



## STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

**ADVERTIMENT.** L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

**ADVERTENCIA.** El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

**WARNING.** Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



UNIVERSITAT  
ROVIRA I VIRGILI

Statistical inference in bipartite networks applied to  
social dilemmas and human microbial systems

Sergio Cobo López

2 December 2019





# Statistical Inference in Bipartite Multilink Networks

---

Sergio Cobo López

Advisors:

Roger Guimerà Manrique

Marta Sales Pardo

Doctoral Thesis

2019



# Abstract

Complex systems are systems comprising many individual elements that interact with each other in highly heterogeneous patterns. Consequently, they display nonlinear dynamics that result in collective behaviors and emergent phenomena that cannot be explained only looking at microscopic interactions. This multiscale behavior can be found in many scientific areas, but complex systems are especially abundant in social sciences and biology. This should not come as a surprise, since both disciplines study many problems with very intricate interactions and large numbers of elements. At the same time, those problems are usually very interesting and relevant. Being able to describe and predict the behavior of biological and social systems is not only very interesting scientifically but also very informative and practical for social or clinical applications.

The goal of this thesis is to make interpretable predictions in complex systems using statistical inference. Interpretable predictions are interesting because it is possible to understand why they are successful or not and because they can reveal the underlying dynamics of the systems under study. This thesis studies the problem of interpretable link prediction in two problems from social sciences and microbiology. These problems, as many others in complex systems can be represented as networks. Networks are simple mathematical artifacts that consist of individual elements called nodes and interactions between them called links. In this regard, they conveniently represent the basic features of most complex systems. Specifically, the problems considered here can be modeled as bipartite networks with different types of links. Bipartite networks are characterized by the existence of two species of nodes. In general, nodes from one species only connect to nodes from the other one. In addition, different types of connections or links allow the study of different forms of interactions.

In order to make predictions in bipartite multilink networks, we implement a family of models called Stochastic Block Models (SBM) that work under the simple assumption that networks have blocks or communities of nodes that define the collective patterns of interactions between nodes. Two particular models from that family are considered here. The first one is a conventional approach in which communities are simply groups of nodes. The second one, in contrast, is a mixed-membership approach that allows nodes to belong to different communities simultaneously. Subsequently, communities become latent or abstract. In both cases the community structure is crucial for formulating predictions, because the probability that two nodes are connected exclusively depends on the communities to which each node belongs. This makes SBM highly tractable, because they reduce the problem to

---

the number of groups identified. Additionally, it makes predictions interpretable, because they ultimately depend on the community structure and probabilities of connections between groups. Finally, SBM are very expressive of the network they represent precisely because they depict it in terms of groups or communities.

In order to test the effectiveness of these methods, we apply them to the problems mentioned above. In the first problem, we study how people behave when they have to make decisions. To that end, we consider a social experiment in which a large group of people make strategic decisions in a game theoretical context. We model this system as a bipartite network in which nodes correspond to players and games and links are represented by the actions performed in each game. Thus, predicting the action taken by a player in a game is equivalent to inferring the existence of a link in the network. We applied the two models mentioned before to this system in order to identify groups of players and groups of games according to the similarities in the decision strategies of players and their perception of the games. We then tested and compared their performance at making predictions in order to select the best model. In both approaches, we found that the classification in groups of players and games is indeed predictive of unobserved actions and informative about the behavior of players. In the case of the conventional approach, we recover 71% of the missing information, and 74% in the mixed approach. We subsequently conclude that the mixed approach is the best model in that it is the most predictive one. Looking at the group structure, we observed that the groups of players reveal consistent strategic phenotypes and that games are perceived by players in a different way than what should be expected from game theoretical criteria.

In the second problem, we study the structure of the human gut microbiome. We have datasets containing microbial concentrations of different species in a large number of human hosts. In a similar fashion as in the first problem, we can also model this system as a bipartite network in which nodes represent by patients and microbes. A patient and a microbe are connected by a link if that microbial species is present in the host. In this case, we only apply the mixed approach to the problem in order to find latent groups of microbes and latent groups of patients. Particularly, we are interested in finding latent microbial profiles that are informative on which groups of microbes are more abundant in patients. We call these microbial profiles latent enterotypes. We test the predictive power of our latent enterotypes by making predictions of unobserved abundances with around 80% accuracy. We also find that taxonomically close microbes tend to be in the same groups identified by our model, which implies that our latent enterotypes are able to capture the biological information of the system. Additionally, we find a well defined ecological order among latent groups of patients and microbes. In particular, we find that there exists an increasing level of specialization in groups of patients and groups of microbes commonly referred to as nestedness.

In both problems, the results show that is possible to find community structures, despite the different nature of the problems. Moreover, these structures are robust in that they are predictive of unobserved events. Finally, they reveal information about the internal dynamics of both systems. This suggests that this inference approach could be extended to other problems in complex systems with similar results.

---

A mis padres.



UNIVERSITAT ROVIRA I VIRGILI

STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

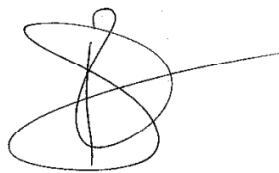
---



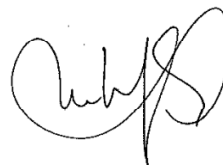
# Declaration

WE STATE that the present study, entitled “Statistical inference in bipartite networks applied to social dilemmas and human microbial systems”, presented by Sergio Cobo López for the award of the degree of Doctor, has been carried out under our supervision at the Department of Chemical Engineering of this university, and that it fulfils all the requirements to be eligible for the International Doctorate Award.

Doctoral Thesis Supervisor/s



Dr. Roger Guimerà Manrique



Dra. Marta Sales Pardo

Tarragona, 2 December 2019

---



# Agradecimientos

Me gustaría comenzar dando las gracias a mis supervisores Marta Sales y Roger Guimerà por todo lo que me han enseñado y porque han contribuido enormemente a que esta tesis haya sido un viaje tan divertido como apasionante. De poca gente he aprendido tantas cosas en tan poco tiempo.

A lo largo de este viaje, he tenido la inmensa suerte no ya de tener a los mejores compañeros de trabajo que pudiera imaginarme, sino de que mis compañeros fueran mis amigos. Muchas gracias a Toñi, a Toni Aguilar, a Pedro, Oscar, Lluís, Lluc y, especialmente, a Oriol, Toni Vallès, Marc e Ignasi.

Muchas gracias también a mis compañeros y amigos del departamento, especialmente a Sandra, Fran, Petra, Noelia, Alberto y Rosa.

Gracias al personal administrativo de la URV y a Núria Mitjana, por felicitar nuestro trabajo a pesar de los pocos medios de que disponen.

A mi amigo Jaeyun y a todas las personas que hicieron de mi estancia en Rochester una experiencia inolvidable. A la gente del Club Natació Tàrraco y a los xinaires, sobre todo a Toni, Gemma, Albert Macarell, Natàlia Bru y Juanvi. A mis amigos de Evohé Juanma y Mar y a mi profesora Ana Corredor y, en general, a toda la gente con la que me he cruzado en Tarragona que ha hecho de esta etapa de mi vida haya sido inolvidable.

Dicen que la parte más complicada de una tesis es el final, pero Ana y yo sabemos que esa es una gran mentira.

A los que siempre están ahí porque solo nos separa la distancia: gracias Mario, gracias Irene.

Y, por encima de todo, gracias a mis padres, por su confianza, su apoyo y su amor incondicionales, sin los cuales habría sido francamente difícil llegar hasta aquí.

---

UNIVERSITAT ROVIRA I VIRGILI

STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
<b>2</b>	<b>Stochastic Block Models</b>	<b>21</b>
2.1	Introduction . . . . .	21
2.2	What are Stochastic Block Models . . . . .	22
2.2.1	Applying SBM to recommender system networks . . . . .	24
2.2.2	Predictions . . . . .	25
2.3	Using bayesian statistics to find the best partitions . . . . .	27
2.4	Beyond regular Stochastic Block Models. Mixed Membership Stochastic Block Models . . . . .	29
2.5	Conclusions . . . . .	32
<b>3</b>	<b>An Application to Social Systems</b>	<b>35</b>
3.1	Introduction . . . . .	35
3.1.1	Data - Game theory . . . . .	36
3.2	Single-strategy model, multiple-strategy model . . . . .	38
3.3	Game metadata and prior modeling . . . . .	39
3.4	Predictive power . . . . .	42
3.4.1	Baseline Model . . . . .	42
3.4.2	Single-strategy models: maximally predictive partitions reveal perception of games by players . . . . .	43
3.4.3	Multiple-strategy models are more predictive and easier to interpret than single-strategy models . . . . .	45
3.5	Discussion and conclusions . . . . .	47
<b>4</b>	<b>An Application to Human Microbiology</b>	<b>51</b>
4.1	Introduction . . . . .	51
4.2	Data . . . . .	52
4.3	Methods . . . . .	53
4.3.1	Enterotypes . . . . .	53
4.3.2	Mixed Membership Stochastic Block Models . . . . .	53
4.3.3	Inference of the most predictive latent enterotypes . . . . .	55

---

4.4	Results . . . . .	55
4.4.1	Cross validation experiments . . . . .	55
4.4.2	Baseline . . . . .	56
4.4.3	Latent enterotype models . . . . .	57
4.4.4	Individual predictability . . . . .	58
4.4.5	Correlation between model parameters and taxonomic information .	59
4.4.6	Nestedness in latent abundance patterns . . . . .	61
4.5	Discussion . . . . .	64
<b>5</b>	<b>Conclusions and perspectives</b>	<b>67</b>
	<b>Appendices</b>	<b>71</b>
5.1	A. Stochastic Block Models . . . . .	71
5.1.1	A1: Multivariate beta function integrals . . . . .	71
5.1.2	A3: Jensen's Inequality . . . . .	72
5.1.3	A2: Simulated annealing . . . . .	73
5.2	B. An Application to Social Systems . . . . .	73
5.2.1	Initial rounds in the empirical data . . . . .	73
5.2.2	Number of groups in the single-strategy and multiple-strategy models	73
5.2.3	Robustness of the results . . . . .	74
5.3	C. An Application to human microbiology . . . . .	74
5.3.1	Robustness of the results . . . . .	74

---

# Chapter 1

## Introduction

### The importance of prediction in science

An important goal common to all scientific disciplines is to make rigorous and accurate predictions. The ability to predict the evolution of a system is not only very informative, but it necessarily implies a good understanding of its underlying dynamics. It is not possible to anticipate unobserved events without a good knowledge of the general principles that rule that system. In science, this knowledge is usually encoded in scientific theories. Indeed, from a scientific point of view, it is precisely the ability to make accurate predictions what validates a hypothesis. When predictions are systematically validated in very different contexts, we talk about a theory.

For example, Charles Darwin was able to predict the evolution of organisms upon the hypothesis that "favorable variations would tend to be preserved, and unfavorable ones to be destroyed" [1]. He formulated this conjecture after discovering fossil bones from large extinct mammals in Argentina and observing the existence of multiple species of finches in the Galápagos islands [2]. He spent the rest of his scientific career making further observations and predictions in different biological systems to validate his hypothesis of natural selection and eventually transform it into his famous theory of the origin of species [3].

Similarly, Newton's law of universal gravitation makes predictions about the trajectories and positions of celestial bodies. However, these predictions arise from a good understanding of the universal principles that govern their interactions. It is precisely the accuracy of these predictions what validated Newton's hypothesis about the motion of celestial bodies.

### Prediction in complex systems

There are many problems in science for which we lack general theories. One particular area in which this happens is complex systems. Complex systems consist of a large number of individual components that interact with each other. However, interactions are highly

---

heterogeneous and, in general, each individual element interacts with the rest of the system in a different way. Take an ecological system, for example. Hundreds or thousands of living organisms coexist in the same physical space, interacting with each other and their environment diversely: insects or herbivores feed on some plants or trees but not on others. Some small predators may prey on rodents or insects, while others, usually bigger, prey on large herbivores. Markets are also a paradigmatic example of complex systems: traders, banks and other financial institutions around the world buy and sell stocks every day with patterns of transactions that change on a daily basis. Complex systems can be found across many different disciplines besides ecology and economics: physics, computer science, and particularly, social sciences and biology. This heterogeneity on the interactions is inherent to all complex systems and implies a lack of symmetries that could simplify their study. It also gives rise to highly nonlinear behaviors. Nonlinearity in complex systems implies that the system cannot be explained as the sum of its components. In our previous example, the extinction of a species can be a good example of a nonlinear behavior: a given species could disappear from the ecosystem because its main source of food also disappeared, a new predator entered the ecosystem or because some environmental changes occurred. In any case, the cause of its extinction would be probably related to different factors concerning the system as a whole and its interactions.

Nonlinear behaviors in complex systems can also give rise to collective or emergent phenomena that again, cannot be explained by looking at the individual elements of the system. Economic crashes or bubbles are classic examples of emergence in market systems. Massive outages in power grids, or outbreaks in contagion networks are other common examples.

Nonlinearity is generally a big hampering for the advancement and development of scientific theories and complex systems science is no exception. Nonetheless, there exist partial theories that can explain particular observed features or families of problems. For instance, societies are very complex systems and no theory can explain in detail the behavior of human beings in social environments, but it is possible to understand the behavior of users in an online social network or costumers in a recommender system. Similarly, we do not know how an aggregate of cells becomes a multicellular living organism, but we have a detailed understanding of how information is transcribed from DNA to proteins, for instance. Besides, there are always advancements towards a more complete knowledge: stochastic processes, dynamical systems, and especially, complex networks are just a few examples.

Complex networks are essentially representations of complex systems. Typically, they consist of two fundamental elements: vertices or nodes and edges or links. Nodes represent the individual components of the system. They could be users in a social network, microbes in a biological network, cities in a transport network and so on. Edges or links represent the interactions between the nodes. In social networks, for instance, a link between two people means that they have interacted at some point or they know each other.

Networks are therefore simple and convenient objects, because they mimic the basic features of almost any complex system, and map the interactions of their individual components and their structure. In many cases, networks have allowed for a much better understanding of complex systems. One classic example is the Canadian cod crisis of the 1990s. During that time there was a dramatic shortage in the cod population in the area of Newfoundland, east Canada [4], [5]. Political authorities and fish industries attributed this shortage to an

---



overpopulation of harp seals, one of the natural predators of cod, and organized hunts of seals. However, these actions did not result in the recovery of the cod population. Meanwhile, ecologists were mapping the food web of the scotian shelf (see. Fig. 1.1) and they eventually found out that cod is just one in the around 150 species that seals prey on. In fact, several of these species turned out to be also cod predators, evidencing the failure of the political solutions attempted. Without a complete picture of the ecologic system of the scotian shelf, it was not possible to understand the complex interactions between harp seals and cod. Besides illustrating the global structure of complex systems, networks also allow

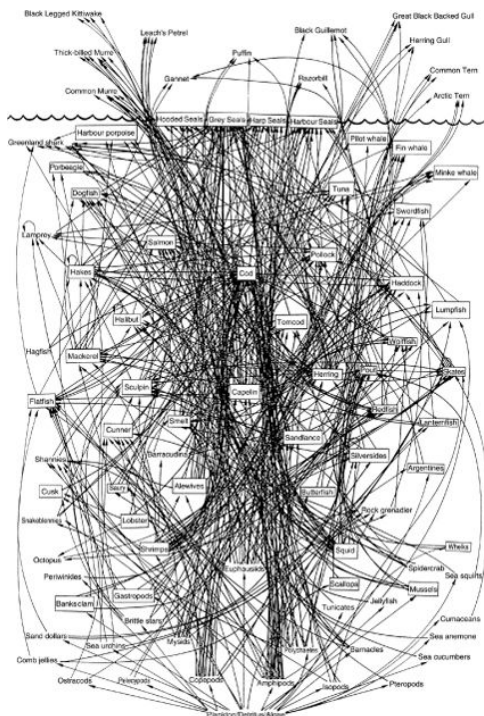


Figure 1.1: **Partial food web of the 'Scotian Shelf' in the north-west Atlantic off eastern Canada. Image reproduced from [4]**

to shift from the macroscopic to the microscopic scales , thus shedding light on the causes of collective and emergent effects. Specifically, networks capture a very important feature of the nature of complex systems: they are not totally regular nor totally random. Precisely because complex systems are not completely random, it is possible to study their regularities and patterns within the network representation. For instance, if we are interested in the statistical properties of a complex system, we can study the degree distribution of its corresponding network, i.e. the number of connections of each node. If we are more interested in topological features, it is possible to analyze other properties such as the node centrality, to see how information flows through and which nodes are more important in that

regard. Alternatively, we may be interested in the structure of a network and its mesoscopic scale and look for communities or groups of nodes that are densely connected in the network. All of these features are informative and descriptive about regularities in networks and about the transition from the microscopic to macroscopic dynamics of complex systems. However, in order to have complete theories about complex systems, we need to elaborate descriptions that are also predictive. Statistical inference is a particularly powerful tool for making predictive descriptions of complex systems. Statistical inference is the process of inferring general properties about a system based on observational data. The idea is to build models that can explain the system in terms of the properties inferred. For example, suppose we have data on a recommender system in which people rate the movies they watch. Probably, we have data on a sample of users and their rating records. Upon these observational data we would like to infer general properties about the system or about the behavior of people in a recommender system context: for instance, are there some popular movies that everybody likes? What is the average number of movies rated? Are there groups of people according to the types of movies they like the most? Using statistical inference approaches, it is possible to test these and other hypotheses and make predictions on the whole system. Besides, these predictions are also easily interpretable.

The network representation is also very helpful when it comes to making predictions in complex systems. Because descriptions of complex systems are done in terms of the network features (nodes and links), the predictive descriptions need to be done in regards to the same network features.

However, statistical inference has not been applied to predictions in networks until recently. The reason is that the formal structure of networks was not well suited for the traditional data types used in statistical inference approaches. This has changed during the last decade. One particular example of prediction in complex network that has become very popular is link prediction [6], [7].

This thesis studies precisely the problem of link prediction with statistical inference modeling. The models used here work on the hypothesis that the nodes in a network can be classified in groups according to their connectivity patterns. In order to test their predictive and descriptive power, we test these models on two novel problems from different disciplines.

## **New data, new opportunities**

Another important advancement in the study of complex systems has come with the recent availability of massive amounts of electronic data. In many cases, these sources of data have opened the door to the study of new problems for which there was simply no data before. In other cases, it has considerably improved the existing knowledge. This thesis considers two of these problems for the application of our statistical inference models. The first problem considered is related to social sciences. Here, we explore the behavior of people when they have to make simple choices in a strategic context. The second problem is a human microbiology one, where we look for similarities in the microbial profiles of healthy patients. We focus on social sciences and biology because they have especially benefitted from this

---

explosion of data. In biology related contexts, for instance, extensive studies on human microbiota [8], [9] or genome sequencing data [10] have boosted the study of these fields and their connection to other levels of systems biology. In social sciences, email networks from companies or organizations [11] or mobile phone data [12], have unveiled patterns of social interactions, leading to important predictions about human behavior.

Although the two problems considered here may seem very different in their nature, they share some structural important similarities. Mainly, they can be both represented as a specific subtype of networks called bipartite multilink networks. These networks are characterized by the existence of two species of nodes and multiple types of links connecting them. The possibility of having different types of links is very convenient because it allows us to consider multiple forms of interactions. For example, it has been observed that two drugs can interact in three different ways: [13] antagonistic (they cancel the effect of each other), synergistic (they enhance the effect of both of them) or additive (they don't interfere with each other). Modeling these interactions requires a network consisting of three types of links to correctly understand the effect of combination of drugs. Social networks are another example of multilink networks, because people tend to be connected in many different ways: they can be family members, friends, work colleagues or simply acquaintances. From a different point of view, social ties can be positive or negative [14]. Each of these connections represents a different interaction with its own characteristics. An accurate representation should take this into account, and treating these networks with a single type of link would hide many important levels of information.

Bipartite networks, on the other hand, are very convenient when one considers systems with two different types of actors or nodes, in which interactions only occur between nodes from different species. Bipartite networks can be found in many ecological systems, such as insect-plant or predator-prey networks. Recommender systems are also a very common example of bipartite networks. These systems, briefly described above, are widely used in e-commerce platforms or online streaming services, where the goal is to make suggestions to costumers based on their records of previous purchases or choices. A recommender system can be modeled as a network in which costumers and items (products or movies) are represented by two types of nodes. A costumer and an item are connected by a link, if the costumer has chosen or bought the item. Overall, multilink bipartite networks are a very convenient representation for the study of many complex systems, because they add some degrees of freedom to the conventional network representation. Particularly, the existence of multiple types of ties introduces a more detailed picture of the problems considered. In addition, the price of the extension of the network modelization is very low both at the formal and computational levels, since the number of parameters required does not increase substantially.

## Machine Learning

Statistical inference models are usually built using Machine Learning (ML) techniques and the models used here are no exception.

---

ML is usually defined as the field of study that gives computers the ability to learn without being explicitly programmed [15]. The performance of ML tools has been increasing over time leading to higher levels of autonomy and computing power.

It is important to note that, contrary to some increasing belief, this does not imply that computers and algorithms are doing science on their own. At the end of the day, their findings are driven by a human scientist that formulates a question, comes up with a plausible hypothesis, tests it on the computer, and interprets the results of the algorithm. If the results are consistent with the hypothesis, the scientist has to build a causal relation or a theory. Otherwise, he or she will have to modify or reformulate the hypothesis and start all over again.

However, it is true that ML and other computer assisted methods can speed up and automate some steps in the scientific method. In particular, we believe that ML tools allow human scientists to formulate and test general hypotheses systematically and very fast, as we will see in this thesis.

## Outline

This thesis is divided in three chapters. The first chapter is devoted to presenting and discussing the statistical inference methods used to study the problems mentioned above. Particularly, we present a family of inference models called Stochastic Block Models (SBM). As already stated, the hypothesis of these models is that nodes in a network can be classified into groups and that these groups are shaped by similarities in the connection patterns and roles of nodes. Two models out of the SBM family are discussed: a conventional one in which nodes are grouped in categories or blocks and a mixed-membership approach in which nodes can belong to different categories simultaneously. We start the chapter discussing the convenience and advantages of SBM. Using a simple example, we next describe the parameters needed to model a network into a conventional SBM and extend the formalism of SBM to model bipartite multilink networks. Next, we explain how SBM can be used to make link predictions in bipartite multilink bipartite networks. Then, we discuss how to find the most predictive partition of nodes into blocks using bayesian statistics analysis. Finally, we discuss the formalism for the mixed-membership approach and the differences with respect to the conventional approach.

In the second chapter, we apply both the conventional and mixed approaches to a human behavior problem. In this problem, we have the results of a social experiment in which real people were interacting in pairs and making strategic decisions within the framework of game theory, i.e. people were playing different types of games. Our goal here is to use both approaches to find groups of people according to behavioral patterns and groups of games according to the perception of players. Using these classifications, we want to make predictions of the actions of players. In addition, we include some existing information about the games into our formalism in the form of bayesian prior distributions. From a methodological point of view, we compare the performance of both models considered. We conclude that the best model is the best predicting one, because it has better captured the internal dynamics of the system. According to this criterion, we find that the best model is

---

the mixed-membership approach.

In the third chapter, we consider a problem of human microbiology. Here, we have different data sets with microbial profiles of a large number of patients. In this case, we only apply the mixed-membership approach. As in the previous case, we use this group structure to make predictions of microbial abundances in individual patients. We also study the predictability of the microbial profiles of individual patients to see if some patients are easier to predict than others. Additionally, we analyze the ecological structure of the groups of patients identified and the correlations of groups of microbes with their taxonomic information.

Finally, we make some conclusions about the results obtained and discuss possible directions for future work. These include, among others, extensions of the problems studied, potential interesting new problems or a detailed analysis of the mechanics of our statistical inference methods.

---

UNIVERSITAT ROVIRA I VIRGILI

STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

---

## Chapter 2

# Stochastic Block Models

### 2.1 Introduction

As stated in the introduction, our goal is to make accurate, reliable, and interpretable predictions of missing links in recommender systems-like networks or bipartite multilink networks. Bipartite networks are characterized by the existence of two types of nodes and by the fact that links only exist between nodes of different species. Multilink networks, on the other side, consist of different types of links. Therefore, our problem translates here into predicting whether two nodes from different species are connected by a link and correctly guessing what type of link connects them. We thus need a method that can efficiently carry out this task of interpretable link prediction.

Link prediction is a widely studied problem in network science, due to its importance in real life applications: successfully predicting an outage in an electric grid, a terrorist attack, an economical crash or simply retrieving missing or corrupted information from an observed system can be of crucial importance. Many methods and approaches have been proposed and studied for link prediction: clustering or neural networks are two prominent examples. However, when it comes to analyzing their results, most of these methods are not easy to interpret. Namely, it is not possible to understand why the method makes successful predictions, neither how it does it.

The problem of interpretability is particularly important because we regard link prediction from a wide perspective: link prediction is not only a tool for recovering information, but also a byproduct of a good understanding of the system represented by the network. That is, we have successfully captured the regularities and patterns arising from the observed data if we can predict missing links accurately. This is something very important from a scientific point of view and underscores the importance of interpretability.

---

## 2.2 What are Stochastic Block Models

In order to find a method that joins predictive accuracy and interpretability, we need a very fundamental approach that emerges from simple assumptions.

One such approach is provided by Stochastic Block Models (SBM). SBM were proposed and developed by sociologists as a tool to identify roles in social data and to find communities [16]. It is precisely this and its simplicity that makes them very convenient for our goal. The main assumption in SBM is actually the idea that there exist blocks or communities inside networks. That is, networks have a non-random internal structure. This structure is shaped by the link patterns among the nodes and the roles played by them. Take the Figure 2.1, for instance. It depicts a very simple network with a clear community structure consisting of three blocks. Note that nodes in the communities display similar roles in the network. For example, all the nodes in the green community are connected to the brown community, but never connect with each other. Nodes in the orange community, on the other hand, are tightly connected to each other, but barely connected to the rest of the network. Finally, the nodes in the brown community tend to be well connected with each other but also to nodes of other communities.

Community detection in networks is not an exclusive feature of SBM, of course. There are many methods to perform this task and indeed, community detection has traditionally been a very active research field. However, there are two important aspects about the SBM that we would like to stand out: first, no assumptions are made about the community structure of the network. That is, there are no constraints on the number or size of the communities. Second, it is very easy (almost straightforward) to analyze the interaction between and within communities. This is due to the fact that, given a community or block, nodes are statistically equivalent for analytical purposes.

To illustrate this idea, let us consider the network in Fig 2.1 . If we look at the orange and brown communities, for instance, we see that they are fully connected: all possible links exist between their nodes. We can thus say that the probability of connections is  $\mathcal{P}_o = \mathcal{P}_b = 1$ <sup>1</sup> (see Fig. 2.1 b). If we now look at the green community, we see exactly the opposite: there is not any connection between their nodes, so that the probability in this case is  $\mathcal{P}_g = 0$ . This should not come as a surprise, since all nodes have the same connections with each other in both cases. Now, let us take a look at the links between the green and brown communities; we see that there are 6 out of 12 possible connections. This means that there is a probability of connection  $\mathcal{P}_{r-b} = 1/2$  between the green and the brown communities. In SBM, we go a step further and state that all nodes are statistically equivalent and any node from the green community is connected with any node from the brown community with  $\mathcal{P}_{r-b} = 1/2$ , despite the fact that nodes have different numbers and patterns of connections.

This has some obvious advantages; the first one is tractability. No matter how many nodes a network has, we only need to know the community structure to estimate the probability of connection of any two nodes. With this in mind, link prediction arises in a very natural and intuitive way. If we want to find a missing or spurious link between two nodes, we only

---

<sup>1</sup>This is actually only true if we have the certainty that that is the real structure of the network. In general, that seldom happens, though: the data could contain observational errors or the network could change over time. However, this simplification allows us to give an intuitive explanation of SBM.



need to know the blocks or block to which each of them belongs. Finally, we see that SBM are very expressive: not only can we identify the community structure of the network but we can also analyze the interplay between communities. All of this, illustrate how SBM are not simply a robust community detection tool, but a very powerful approach for network analysis and link inference.

A key feature to enhance the interpretability of SBM is its formal simplicity. Importantly, two parameters are required to describe any Stochastic Block Model: first, we need to know the membership of nodes to blocks. Second, we need to know the probability of blocks being connected.

We have sketched the second parameter in 2.1b. In reality, it corresponds to a matrix  $P$  of  $K \times K$  dimensions,  $K$  being the number of blocks identified. Each matrix element corresponds to the probability of block  $k_i$  being connected to block  $k_j$ . It is important to keep in mind that the number of nodes in each block is irrelevant, as mentioned before. In our example, the matrix  $P$  can be expressed as:

$$P = \begin{bmatrix} 1 & 1/9 & 0 \\ 1/9 & 1 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix}$$

As for the membership of nodes, this information can be described by an array of membership vectors  $\theta$ . This array consists of  $n$  vectors each with  $K$  dimensions,  $n$  being the number of nodes in the network. It can also be represented as an  $n \times K$  matrix. In our example, this membership array would be:

$$\theta = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

The first, second and third columns correspond to the orange, brown, and green communities shown in the figure, respectively. Note that the only constraint on  $\theta$  is that it can only take 1 or 0 as numerical values, because each node can only belong to a community.

These two parameters  $P$  and  $\theta$  are all that we need to characterize a Stochastic Block Model. While  $\theta$  is basically a record of the membership of every node,  $P$  provides the information about the interaction between blocks. Of course, the numerical values of the matrix elements are also very important in this regard. Having two blocks connected with probability 1 or 0 is clearly more informative than a probability of 0.5.

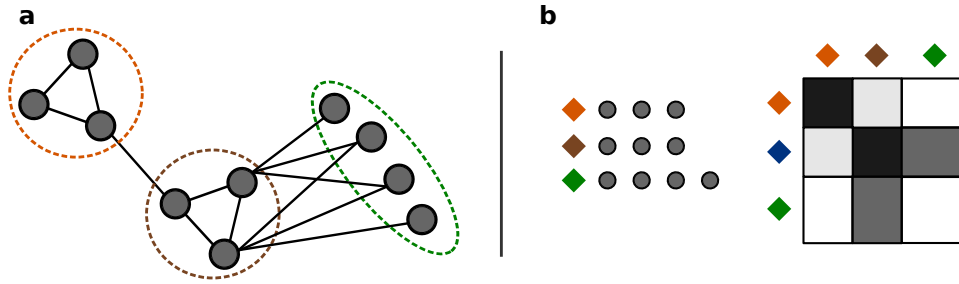


Figure 2.1: **Community and block structure of a simple network.** (a) Network with a community structure. The orange and brown communities have the maximum number of possible links connecting their nodes, while the green community has its nodes completely disconnected. Conversely, the orange and brown communities are only connected by 1 out of the 9 theoretically possible links, the green and brown communities share half of the 12 possible links and the orange and green communities are completely disjoint. (b) A schematic representation of the parameters of the Stochastic Block Models: memberships of nodes into communities and the matrix of interactions among blocks.

### 2.2.1 Applying SBM to recommender system networks

We need to translate this formalism to bipartite multilink networks or recommender systems, since they are our object of interest. Bipartite multilink networks are obviously more sophisticated than the simple networks described above. However, due to the simplicity SBM rather few modifications are required. As mentioned above, bipartite networks consist of two types of nodes. Importantly, nodes from one type only connect to nodes from the other type and blocks are defined by the pattern of connections from one species of nodes to the other one ( see 2.2 below). Consequently, there will also be two types of communities or blocks, which means that we need two membership arrays  $\theta$  and  $\eta$  to account for the membership of nodes to blocks. In the example shown in 2.2  $\theta$  and  $\eta$  would be:

$$\theta = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \eta = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Note that the idea of probabilities of connection of nodes in the same block does not make sense anymore since nodes from the same species never connect in bipartite networks. The question now is the probability of connection between communities of one species and the other one. Furthermore, since we are considering multilink networks, we are actually

interested in the probability of connection with a given type of link. This implies an extension of the matrix  $P$  to an array of  $P$  matrices  $P_1, \dots, P_\Lambda$ , where  $\Lambda$  is the number of species of links considered. In the example described below,  $\Lambda = 2$ , and we therefore have  $2 \times 3 \times 3$   $P$  matrices:

$$P_{red} = \begin{bmatrix} 1/6 & 1/9 & 0 \\ 0 & 1/3 & 0 \\ 0 & 1/12 & 2/3 \end{bmatrix} \quad P_{blue} = \begin{bmatrix} 5/6 & 0 & 0 \\ 0 & 2/3 & 0 \\ 0 & 1/12 & 1/4 \end{bmatrix}$$

With these small extensions of the model, we would be able to successfully model any recommender system. Obviously, there is an increase of complexity, but the formalism remains rather simple and interpretable.

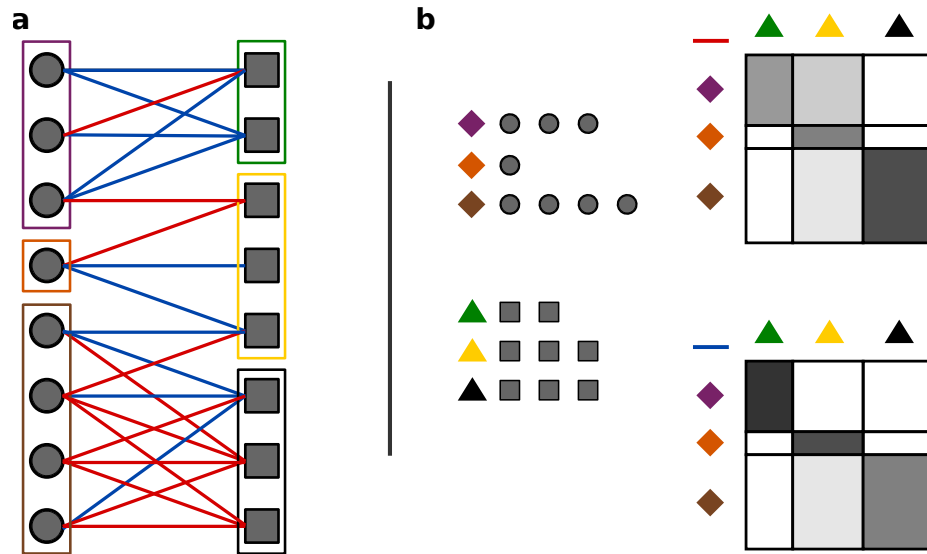


Figure 2.2: **Community and block structure of a bipartite multilink network.** (a) A simple bipartite network with two types of links. Communities are determined by the patterns of connection of both links between nodes of different species. (b) A schematic representation of the parameters of the Stochastic Block Models: memberships of nodes to communities and the matrix of interactions among blocks.

### 2.2.2 Predictions

We have highlighted the potential of SBM for link inference before. However, we haven't explained how they can be used for this task. Suppose we take the example in 2.2b and remove a few links from the original bipartite network as shown in 2.3. Suppose also that

none of those links is so structurally important to modify the original block arrangement in 2.2 . Therefore, the membership matrices  $\theta$  and  $\eta$  will not change in the absence of those links. The P matrices in contrast will do, because they depend on the overall number of links, which is smaller now. The parameters of the Stochastic Block Model of the modified network are like follows:

$$\theta = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \eta = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$P_{red} = \begin{bmatrix} 1/6 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 7/12 \end{bmatrix} \quad P_{blue} = \begin{bmatrix} 2/3 & 0 & 0 \\ 0 & 2/3 & 0 \\ 0 & 1/12 & 1/4 \end{bmatrix}$$

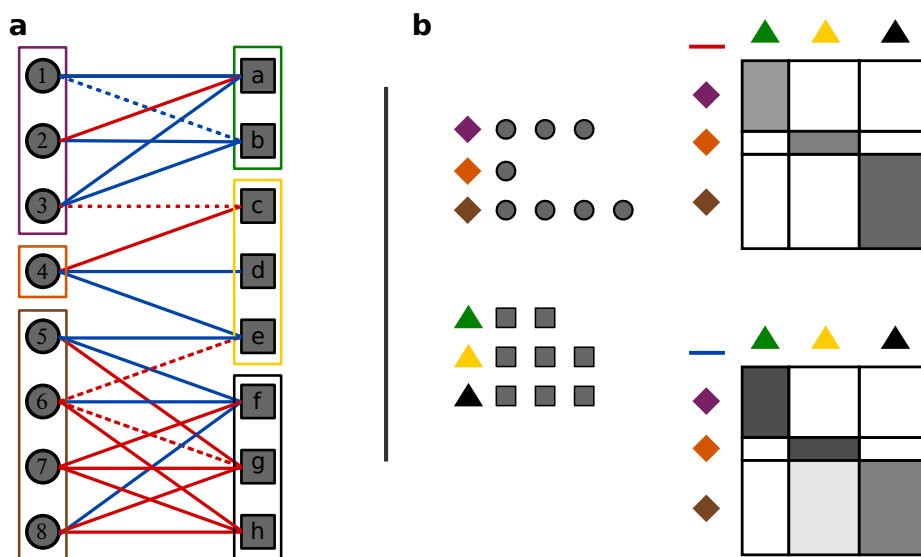


Figure 2.3: **Incomplete observed network.** (a) We remove four links from the original network in a way that does not modify the block structure (b) Only the matrices of probabilities of connection among blocks are modified by the overall reduction in number of links

Once our model is determined, we use it to try predict the existence of the four removed links: 1-b, 3-c, 6-e, and 6-g. The protocol is as follows: first, look at the memberships of both nodes  $i$  and  $j$  to blocks  $m$  and  $n$  in  $\theta_m$  and  $\eta_n$ , respectively. Second, look at the probability of connection of blocks  $m$  and  $n$  with each type of link in  $P_{mn}^{red}$  and  $P_{mn}^{blue}$ . Third, take the highest probability over a certain threshold ( $1/\Lambda = 0.5$ , for example) or choose no connection if both values are below the threshold. Fourth, check whether your prediction matches the actual links to assess the predictive power of your model.

In our example, we can successfully predict half of the missing links (see 4.1). Note that we are able to predict only links between communities that are densely connected both in the original and modified version. This makes sense, since we rely on observations to make predictions and it is generally more difficult to predict rare events.

We also observe that making link prediction in SBM is a very straightforward procedure for which we only need to extract specific information from its parameters.

Table 2.1: Predictions

Node <sub>1</sub>	Block <sub>1</sub>	Node <sub>2</sub>	Block <sub>2</sub>	$P_{red}$	$P_{blue}$	Predicted	Actual	T/F
1	Purple	b	Green	1/6	2/3	Blue	Blue	True
3	Purple	c	Yellow	0	0	None	Red	False
6	Brown	e	Yellow	1/12	0	None	Red	False
6	Brown	g	Black	1/6	2/3	Red	Red	True

## 2.3 Using bayesian statistics to find the best partitions

So far, we have described the main features and formalism of Stochastic Block Models, we have discussed how to adapt them to recommender system networks, and how to exploit their potential to make link prediction. However, we haven't explained how to actually find a partition of a network into blocks or, more specifically, how to find the best partition out of the many possibilities that are theoretically possible. We apply a bayesian approach to solve this problem and find that the best partition is the most plausible one or the one that maximizes the posterior probability of the model parameters. We next discuss how to do this.

Suppose we have a record of observations  $A^o$  from a recommender system network with  $\Lambda$  possible ratings and we want to find the most plausible partition of that network into blocks. Formally speaking, this is equivalent to finding the most plausible membership vectors  $\theta$  and  $\eta$ , i.e. the ones that maximize the posterior probability  $P(\theta, \eta | A^o)$ . That is, the probability of the membership vectors  $\theta, \eta$  given the observations  $A^o$ . Importantly, finding the most plausible partition is in general equivalent to finding the most predictive one [17].

Our first step is to marginalize over the  $\mathbf{p}$ -matrices (what is marginalization?):

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o) &= \int_{\mathbf{p}} d\mathbf{p} P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p} | A^o) \\ &= \int_{\mathbf{p}} d\mathbf{p} \frac{P(A^o | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p})}{P(A^o)} \end{aligned} \quad , \quad (2.1)$$

where we have introduced the Bayes' theorem to express the integrand in terms of the likelihood of the model  $P(A^o | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p})$ , the prior information  $P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p})$ , and the evidence  $P(A^o)$ .  $P(A^o)$  acts as a normalization constant for the whole integrand, and we can actually regard it as a partition function in a similar fashion as it occurs in statistical mechanics:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o) = \frac{\int d\mathbf{p} P(A^o | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p})}{\int d\boldsymbol{\theta}' d\boldsymbol{\eta}' d\mathbf{p}' P(A^o | \boldsymbol{\theta}', \boldsymbol{\eta}', \mathbf{p}') P(\boldsymbol{\theta}', \boldsymbol{\eta}' | \mathbf{p}') P(\mathbf{p}')} = \frac{1}{\mathcal{Z}} \int d\mathbf{p} P(A^o | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p}), \quad (2.2)$$

Next, we have the prior  $P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p})$ . In Bayesian statistics, the prior reflects our previous knowledge about the system. An example of a prior in a recommender system could be some information about a preference towards a certain item from a subset of the users. We will assume  $P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p}) = \text{const}$  in the present discussion.

Finally, the likelihood  $P(A^o | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p})$  represents the probability that the model gives rise to the observed data  $A^o$ , and it can be expressed as:

$$P(A^o | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) = \prod_{k \in K} \prod_{\ell \in L} \prod_{\lambda=1}^{\Lambda} p_{\lambda}(k, \ell)^{n_{k\ell}^{\lambda}}, \quad (2.3)$$

where  $p_{\lambda}(k, \ell)$  is the probability that the group of users  $k$  rates the group of items  $\ell$  with  $\lambda$  and  $n_{k\ell}^{\lambda}$  is the total number of such interactions in the observed network  $A^o$ .

Thus, inserting the likelihood in the previous equation, we get:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o) = \frac{1}{\mathcal{Z}} \int d\mathbf{p} \prod_{k \in K} \prod_{\ell \in L} \prod_{\lambda=1}^{\Lambda} p_{\lambda}(k, \ell)^{n_{k\ell}^{\lambda}}, \quad (2.4)$$

This integral is a multivariable beta function and can be solved analitically (see Appendix A):

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o) = \frac{1}{\mathcal{Z}} \prod_{k\ell} \frac{n_{k\ell}^1! n_{k\ell}^2! \dots n_{k\ell}^{\Lambda}!}{(n_{k\ell}^1 + n_{k\ell}^2 + \dots + n_{k\ell}^{\Lambda} + \Lambda - 1)!}. \quad (2.5)$$

In order to reduce the computational cost of the process, we take the exponential of the natural logarithm of  $P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o)$ :

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o) = \frac{1}{\mathcal{Z}} \exp \sum_{k\ell} \left[ \sum_{\lambda=1}^{\Lambda} \ln (n_{k\ell}^{\lambda})! - \ln (n_{k\ell} + \Lambda - 1)! \right], \quad (2.6)$$

with  $n_{k\ell} = n_{k\ell}^1 + \dots + n_{k\ell}^A$  being the total number of observations from group  $k$  to group  $\ell$ . Then, introducing :

$$\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{k\ell} \left[ \ln (n_{k\ell} + \Lambda - 1)! - \sum_{\lambda=1}^A \ln (n_{k\ell}^\lambda)! \right] \quad (2.7)$$

we can express the previous equation as:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^0) = \frac{1}{\mathcal{Z}} \exp [-\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\eta})] . \quad (2.8)$$

Finally, we obtain the most plausible partitions  $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$  of players and games by minimizing  $\mathcal{H}$ , that is:

$$(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) = \arg \min_{\boldsymbol{\theta}, \boldsymbol{\eta}} \mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\eta}) . \quad (2.9)$$

Because this is a combinatorial optimization problem, we use simulated annealing for the minimization (Appendix A, Section 1). Note that we do not have to fix the number of groups  $k$  and  $\ell$ ; we obtain these groups as a result of the inference process.

## 2.4 Beyond regular Stochastic Block Models. Mixed Membership Stochastic Block Models

In the above discussion we have implicitly assumed that each player and each item in the recommender system belong exclusively to a single block. However, this assumption can be too strong in many contexts. Think of a movie recommender system, for instance. It is very reasonable to think that many movies will not fall under a single category like 'drama', 'comedy' or 'action' and will rather be a mixture of different genres in variable proportions. The same could be possibly applied to users. Certainly, most of them are not "unidimensional" watchers. Very likely, the regular Stochastic Block Models discussed so far, would fail to predict the preferences of users in this system.

This problem can be solved using Mixed Membership Stochastic Block Models (MMSBM) [18]. MMSBM are essentially a generalization of Stochastic Block Models in which nodes belong in general to all blocks simultaneously with different intensities. As a result, blocks are not just groups of nodes anymore. Rather they become latent entities variably represented in nodes. This is comparable to what genres of movies represent in the example discussed above.

From a formal point of view, the MMSBM introduce important changes in the membership vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\eta}$ . Instead of zeros or ones, their elements become now real numbers between 0 and 1 (membership vectors are normalized and  $\sum_{k=1}^K \theta_i^k = 1$  and  $\sum_{\ell=1}^L \eta_j^\ell = 1$  for every user  $i$  and movie  $j$ ). This implies a considerable higher computational cost to find the best partition; it is not possible anymore to exactly marginalize over the p-matrices. Therefore, we assume that the distribution  $P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p} | A^0)$  is very peaked around

$\mathbf{p}^* = \arg \max_{\mathbf{p}} P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p} | A^0)$  and use the approximation [19].

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^0) \approx P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}^* | A^0). \quad (2.10)$$

Because  $\boldsymbol{\theta}$  and  $\boldsymbol{\eta}$  are not discrete as in the single-strategy model, one can use a variational approach to obtain analytic expressions for the optimal membership vectors  $\boldsymbol{\theta}^*$  and  $\boldsymbol{\eta}^*$ , as well as for  $\mathbf{p}^*$ .

As in the regular Stochastic Block Models, our goal is to maximize the posterior probability  $P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p} | A^0)$ . This quantity is unknown, but we can use again the Bayes' theorem and get:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p} | A^0) = \frac{P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p})}{P(A^0)} \quad (2.11)$$

As in the previous case,  $P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p})$  is the likelihood, the prior information is represented by  $P(\boldsymbol{\theta}, \boldsymbol{\eta} | \mathbf{p}) P(\mathbf{p})$ , and  $P(A^0)$  is the evidence. We consider the prior and evidence to be constant, which reduces the equation to:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p} | A^0) = P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) \quad (2.12)$$

and we only need to maximize the likelihood, whose expression is known:

$$P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) = \prod_{u, i \in A^0} \sum_{k, \ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui}), \quad (2.13)$$

where  $r_{ui}$  is the rating given by user  $u$  to item  $i$  in the observed data. Notice that the likelihood in the Mixed Membership approach is different from that in the regular SBM, as it depends indirectly on every couple of user and item. The reason is that, as stated before, blocks are not groups of nodes anymore. This increases the complexity of the model, and it is the main reason we use an Expectation-Maximization algorithm to find the values of the parameters  $\boldsymbol{\theta}$ ,  $\boldsymbol{\eta}$ , and  $\mathbf{p}$  that maximize the posterior.

The first step to develop this algorithm is to take the logarithm of the likelihood (loglikelihood). We do this in order to take advantage of the properties of the logarithm and also to reduce computational costs:

$$\log P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) = \sum_{u, i \in A^0} \log \sum_{k, \ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui}) \quad (2.14)$$

Next, we introduce the auxiliary distribution  $\omega_{ui}(k, \ell)$ :

$$\log P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) = \sum_{u, i \in A^0} \log \left( \sum_{k, \ell} \omega_{ui}(k, \ell) \frac{\theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui})}{\omega_{ui}(k, \ell)} \right), \quad (2.15)$$

$\omega_{ui}(k, \ell)$  being the probability that a rating  $r_{ui}$  is due to  $u$  and  $i$  belonging to groups  $k$  and  $\ell$ , respectively.

Now we apply the Jensen's inequality  $\log \bar{x} \geq \overline{\log x}$  to the right hand of this expression and get (see Appendix A):

$$\log \left( \sum_{k, \ell} \omega_{ui}(k, \ell) \frac{\theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui})}{\omega_{ui}(k, \ell)} \right) \geq \sum_{k, \ell} \omega_{ui}(k, \ell) \log \frac{\theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui})}{\omega_{ui}(k, \ell)}, \quad (2.16)$$



where the lower bound of the inequality holds if:

$$\omega_{ui}(k, l) = \frac{\theta_{uk}\eta_{il}p_{k\ell}(r_{ui})}{\sum_{k', \ell'} \theta_{uk'}\eta_{i\ell'}p_{k'\ell'}(r_{ui})}, \quad (2.17)$$

which gives the expectation step in the algorithm. The loglikelihood thus becomes:

$$\log P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) = \sum_{u, i \in A^0} \sum_{k, \ell} \omega_{ui}(k, \ell) \log \frac{\theta_{uk}\eta_{il}p_{k\ell}(r_{ui})}{\omega_{ui}(k, \ell)}. \quad (2.18)$$

For the maximization step, we need to derive update equations for the parameters  $\boldsymbol{\theta}$ ,  $\boldsymbol{\eta}$  and  $\mathbf{p}$ . We take derivatives of the loglikelihood. We need to include Lagrange multipliers to account for the normalization constraints of the  $\boldsymbol{\theta}$  and  $\boldsymbol{\eta}$ . The function to take derivatives is thus:

$$\mathcal{L} = \log P(A^0 | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) - \gamma \theta_{uk} - \phi \eta_{il} - \chi p_{k\ell}(r) \quad (2.19)$$

with  $\gamma$ ,  $\phi$ , and  $\chi$  being the Lagrange multipliers. That means that if we derive (2.19) with respect to  $\theta_{uk}$ , we would have:

$$\frac{\partial \mathcal{L}}{\partial \theta_{uk}} = 0 \implies \frac{\partial \log \mathcal{P}}{\partial \theta_{uk}} = \gamma \quad (2.20)$$

Similarly for  $\eta_{il}$ :

$$\frac{\partial \mathcal{L}}{\partial \eta_{il}} = 0 \implies \frac{\partial \log \mathcal{P}}{\partial \eta_{il}} = \phi, \quad (2.21)$$

and for  $p_{k\ell}(r)$ :

$$\frac{\partial \mathcal{L}}{\partial p_{k\ell}} = 0 \implies \frac{\partial \log \mathcal{P}}{\partial p_{k\ell}} = \chi. \quad (2.22)$$

With this in mind, we take derivatives with respect to  $\theta_{uk}$ :

$$\begin{aligned} \frac{\partial \log P}{\partial \theta_{uk}} &= \sum_{i \in A^0} \sum_{\ell} \omega_{ui}(k, \ell) \frac{1}{\theta_{uk}} = \gamma \implies \\ &\sum_{i \in A^0} \sum_{\ell} \omega_{ui}(k, \ell) = \gamma \theta_{uk} \end{aligned} \quad (2.23)$$

Now, we take the summation over all  $k$  on both sides:

$$\sum_{i \in A^0} \sum_{kl} \omega_{ui}(k, \ell) = \gamma \sum_k \theta_{uk} \quad (2.24)$$

Because both  $\theta_u$  and  $\omega_{kl}(i, g)$  are normalized, we have:

$$d_u = \gamma, \quad (2.25)$$

where  $d_u$  is the degree of item  $i$ , i.e. the number of users to which it is connected. Combining this result with (2.23), we get:

$$\theta_{uk} = \frac{\sum_{i \in A^0} \sum_{\ell} \omega_{ui}(k, \ell)}{d_u}. \quad (2.26)$$

The derivation for  $\eta_{il}$  is analagous and we get:

$$\eta_{il} = \frac{\sum_{u \in A^o} \sum_k \omega_{ui}(k, \ell)}{d_i}, \quad (2.27)$$

$d_i$ , being the degree of user  $u$  or the number of items that he or she has rated. Finally, for the  $\mathbf{p}$ , we take derivatives respect to  $p_{k\ell}(r)$ :

$$\begin{aligned} \frac{\partial \log P}{\partial p_{k\ell}(r)} &= \sum_{u, i \in A^o | r_{ui}=r} \omega_{ui}(k, \ell) \frac{1}{p_{k\ell}(r)} = \chi \implies \\ &\sum_{u, i \in A^o | r_{ui}=r} \omega_{ui}(k, \ell) = \chi p_{k\ell}(r). \end{aligned} \quad (2.28)$$

Summing over all ratings  $r$  on both sides:

$$\sum_r \sum_{u, i \in A^o | r_{ui}=r} \omega_{ui}(k, \ell) = \chi \sum_r p_{k\ell}(r). \quad (2.29)$$

Because the  $p_{k\ell}(r)$  are normalized, we have:

$$\sum_{u, i \in A^o} \omega_{ui}(k, \ell) = \chi. \quad (2.30)$$

Replacing the value of  $\chi$  in Eq. 2.28, we get:

$$p_{k\ell}(r) = \frac{\sum_{u, i \in A^o | r_{ui}=r} \omega_{ui}(k, \ell)}{\sum_{u, i \in A^o} \omega_{ui}(k, \ell)}. \quad (2.31)$$

## 2.5 Conclusions

We have studied the potential of SBM and MMSBM for community detection and link prediction. Ultimately, we have seen their potential to develop both tasks in a very interpretable way. Specifically, the model parameters allow us to extract valuable information about the internal dynamics of the system in both approaches.

Generally, it has been observed that MMSBM have a higher predictive power than conventional SBM [20]. This is consistent with the idea that in many systems we cannot expect each node to belong to a single category or block but rather to different categories at the same time. Even if we study a system in which nodes have a very unidimensional nature, we must not forget that SBM are a particular case of MMSBM. Finally, although counter-intuitively, interpretability of MMSBM can be easier than in regular SBM in some specific aspects. The reason is that it has been empirically observed that  $\mathbf{P}$  matrices tend to have less informative numerical values in SBM (i.e., closer to 0.5) than in MMSBM, where matrix elements tend to be very close to 0 or 1. As we will see in the next chapters,  $\mathbf{P}$  matrices offer very valuable information in terms of the internal dynamics of the system.

This does not mean that we have to completely discard conventional SBM. First, MMSBM work under the assumption that the posterior distribution is very peaked around a single value of the member parameters. SMBs on the other hand are more robust in that it is possible to marginalize over parameters to find its posterior probability. They also have a lower number of model parameters, making them formally simpler.

All in all, we have a powerful toolset to accomplish our goal of interpretable link prediction in bipartite multilink networks. In the next two chapters we will analyze the results of applying it to two different problems. In particular, the second chapter shows a comparison between the two approaches and a practical case of bayesian prior information. The third chapter is less focused on the methodology and we directly use the Mixed Membership approach to analyze predictability and interpretability.

---

UNIVERSITAT ROVIRA I VIRGILI

STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

---

## Chapter 3

# An Application to Social Systems

### 3.1 Introduction

Many human activities have strong recurrent patterns that make them easy to predict [21, 22]. Other activities involve active decision making and are, therefore, not so obviously predictable. These include relatively simple decisions about, for example, which movie to watch or which book to purchase (as discussed in the introduction [23, 19]), as well as complex strategic decisions in which individuals need to anticipate and take into consideration the decisions of others. For these complex strategic decisions, the question of whether the decision-making process is predictable, and to what extent, has been largely unexplored.

Strategic decision making has been studied in the social sciences, especially in political science and in economics [24, 25]. Experimental approaches, in which individuals play simplified games that pose specific social dilemmas, have been particularly insightful and have demonstrated that individuals often do not act “rationally” to maximize their profits [26, 27, 28]. This makes their behaviors more unpredictable than one may have anticipated. Unfortunately, approaches to analyze data from these experiments have focused mostly on characterizing aggregate behaviors (qualitatively or using regression-based approaches) and on measuring deviations from rationality, on the aggregate and at the level of individuals [29]. In general, however, they have not assessed quantitatively the power of existing theories to predict accurately the actions of each individual.

The lack of such analyses is significant because quantifying predictability provides a rigorous framework to compare theories of decision-making. Indeed, if we formalize theories into models and compare their predictive power, the most plausible theory will be, in general, the one that makes the most accurate predictions [30, 17]. Similarly, given a simple model that we aim to refine and improve, the refinements should increase predictability; otherwise, they ought to be revised or discarded [30]. All in all, studying the predictive power of theories and models opens the door to advance our understanding of human behavior on solid grounds [30].

In this chapter, we aim precisely at narrowing this gap by proposing models for strategic

---

decision making, by developing rigorous model inference approaches, and by showing that individual strategic decisions are, to a large extent, predictable. Specifically, we focus on a recent large-scale study of individuals playing a variety of dyadic games in a controlled setting [31]. We propose the two models and inference approaches discussed in the previous chapter and show that are more predictive than those built upon expectations of individuals' rationality. In this case, Stochastic Block Models are based on the assumption that there are groups of individuals that use similar decision-making strategies (have similar behavioral phenotypes [32, 33, 34, 31]), and groups of games that are perceived similarly by individuals. We are agnostic a priori about which groups of individuals are most appropriate, so that the groups we obtain arise from fitting the models and are the ones that describe the observed behaviors most parsimoniously [17]. Similarly, we do not make strong assumptions about the groups of games that are perceived similarly by individuals and, in particular, we do not assume that games with the same Nash equilibrium [35] are in the same group. However, we do exploit existing information on the similarities between the payoffs of particular pairs of games. Importantly, our approach gives predictive models that are interpretable, which enables us to conclude that: (i) the perception of games by individuals is at odds with their game-theoretical structure; (ii) individuals do not use a single strategy when making decisions but rather a combination of multiple simple strategies, which our approach reveals naturally.

### 3.1.1 Data - Game theory

We consider a dataset [31] consisting of 541 individuals playing a collection of dyadic games in which each player has to choose between two actions: cooperation ( $C$ ) or defection ( $D$ ). Therefore, the games have four possible outcomes and each of them is characterized by a so called payoff matrix (Fig. 3.1a) that displays the results of those outcomes: if both players choose cooperation, they both obtain a "reward"  $R$ ; if both defect, they both get a "punishment"  $P$ ; and if one cooperates and the other defects, the cooperator gets a "sucker's payoff"  $S$  and the defector a "temptation payoff"  $T$ .

For the experiments, the reward and punishment were fixed to  $R = 10$  and  $P = 5$ , whereas the other payoffs took values  $S \in \{0, 1, \dots, 10\}$  and  $T \in \{5, 6, \dots, 15\}$ , summing up to 121 games in total (see Fig. 3.1 b). Depending on the different values of  $S$  and  $T$ , games have optimal strategies or choices. These are called Nash equilibria [35] and can be described as the assumed rational behaviors in which no players have incentives to change their behavior or action. According to these Nash equilibria, the 121 games are divided into four groups (Fig. 3.1 b) [31]: harmony game (HG), snowdrift game (SG) (also known as chicken game or hawk-dove game), stag hunt (SH) game, and prisoner's dilemma (PD). The Nash equilibria of HG and PD prescribe cooperation and defection as pure strategies, respectively. This means that a rational player playing HG would choose cooperation unconditionally. For instance, if she were to play iteratively against the same opponent, she would stick to cooperation every time. SG has a stable equilibrium that corresponds to a mixed strategy (that is, players have a certain probability of defecting and a certain probability of cooperating) [35, 31]. To illustrate the idea of mixed strategy, we can think of Rock-Scissors-Paper. If we always choose Rock, for example, our opponent

---

will eventually guess our action and will play paper consistently. Eventually, we'd be losing every time. Therefore, we would want to choose a mixed-strategy in which we alternate at least two choices with equal probabilities.

Each individual in the dataset played an average of 14 rounds, each one with a randomly selected payoff matrix (that is, a randomly chosen  $(S, T)$  pair) and against a different, unknown, and randomly-selected player (that is, in general there were no repeated interactions between pairs of players). Based on the payoffs they obtained, participants received tickets for a lottery (one ticket for 40 payoff points), in which they could win four coupons of 50 euro redeemable at predefined stores [31].

We observe that the behavior of each player during the first four rounds is erratic, which leads to their behavior being less predictable during those rounds (Fig. B1). After round 4, all rounds are statistically indistinguishable by the metrics we use in what follows. Therefore, we discard the first four rounds of each player and consider all others as indistinguishable.

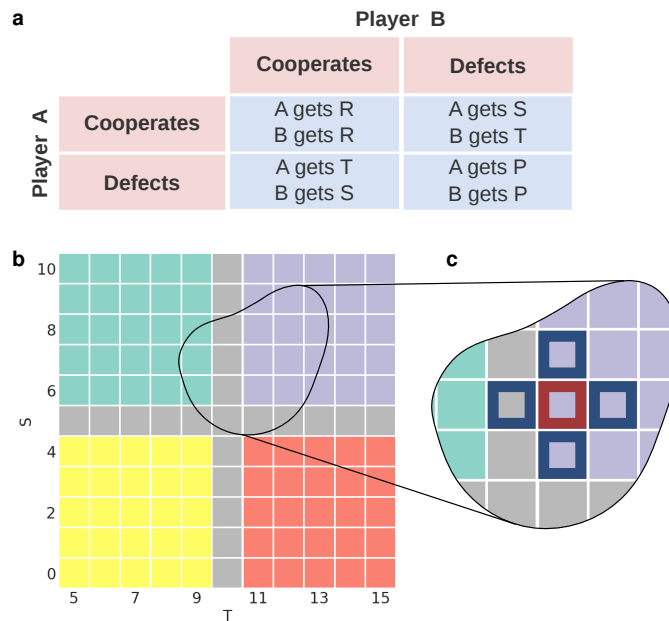


Figure 3.1: **Structure of the dyadic games.** (a) Payoff matrix for the dyadic games played by individuals. (b) The games are typically divided into four groups depending on the different Nash equilibria and on the payoffs  $S$  and  $T$  (given that  $R$  and  $P$  are constant for all games). (c) In our models, we do not impose any group structure for games, but we assume that neighboring games are a priori more likely to belong to the same group. The neighbors of the game marked in red are marked in dark blue.

### 3.2 Single-strategy model, multiple-strategy model

It has been postulated that individuals can be classified into “behavioral phenotypes” depending on how they make decisions when facing social dilemmas [32, 33, 31]. The idea behind the concept of phenotype is that individuals apply general rules when making decisions, regardless of the structure of individual games. For example, players displaying the “envious” phenotype in Ref. [31] cooperate if, and only if, cooperation leads to higher payoffs for them than their opponents, regardless of whether that corresponds to the Nash equilibrium. In practice, among all possible combinations of strategies for single games there are only a few phenotypes that are followed by players. Here, we formalize this idea into group-based generative models of decision making, and go a step further to investigate to what extent these groups can be used to predict unobserved decisions.

We propose two models: a single-strategy model and a multiple strategy model that correspond to the conventional Stochastic Block Models (SBM) [36] and the Mixed Membership Stochastic Block Models (MMSBM) [18, 19, 37] discussed in Chapter 1. In the former, we assume that each player  $i$  belongs to one group of players  $k$ . Similarly, each game  $g$  (defined by the payoff values  $T$  and  $S$ ) belongs to solely one group of games  $\ell$ . We encode this information in the membership vectors  $\theta_i$  and  $\eta_g$  that track the membership of any player and game to a given group. The probability of player  $i$  taking action  $a_{ig} \in \{C, D\}$  on game  $g$  depends exclusively on the groups  $k$  and  $\ell$ , so that the probability of cooperation is

$$\Pr[a_{ig} = C] = p_{k\ell} \quad (3.1)$$

with  $p_{k\ell} \in [0, 1]$ . The element  $p_{k\ell}$  can therefore be interpreted as the strategy of players in group  $k$  for each of the games in group  $\ell$  (see Fig. 3.2). In a slight abuse of language we also call strategy the vector  $\mathbf{p}_k$  of strategies for all game groups; because in the single-strategy model each individual has a single strategy,  $\mathbf{p}_k$  also defines their phenotype. Note that in our approach we do not make any assumption about the strategies for individual games, which can either be pure ( $p_{k\ell} \in \{0, 1\}$ ) or mixed ( $0 < p_{k\ell} < 1$ ) [38].

In the multiple-strategy model, players do not always follow the same strategy but rather adopt different strategies with certain probabilities. We formalize this idea exactly as we did in the discussion of the MMSBM: we allow players to belong to a mixture of groups, with  $\theta_{ik}$  being the probability that player  $i$  belongs to group  $k$  ( $\sum_k \theta_{ik} = 1$ ), that is, that  $i$  adopts strategy  $k$ . Similarly, we assume that games are not always regarded by players as belonging to the same group, and allow games to also belong to a mixture of groups;  $\eta_{g\ell}$  is the probability that game  $g$  is regarded as belonging to group  $\ell$  ( $\sum_\ell \eta_{g\ell} = 1$ ). In this model, the probability that player  $i$  cooperates in game  $g$  is thus

$$\Pr[a_{ig} = C] = \sum_{k,\ell} \theta_{ik} \eta_{g\ell} p_{k\ell}, \quad (3.2)$$

where  $p_{k\ell}$  is the same as in the single-strategy model and the sum is over the  $K$  groups for players and the  $L$  groups for games (Fig. 3.2). In fact, if we restrict the elements of the membership vectors  $\theta$  and  $\eta$  to be either 0 or 1 (that is games/players are restricted to belong to a single group) we recover the single-strategy model, so in what follows we use the multiple-strategy formulation without loss of generality.



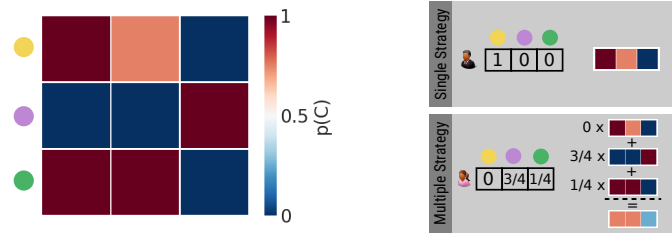


Figure 3.2: **Single-strategy and multiple-strategy models.** The matrix of cooperation probabilities (left) indicates the probability with which players in a player group cooperate in games in game group. In the single-strategy model (top right), players belong to a single group, so that their decisions are given directly by the corresponding row in the matrix of cooperation probabilities. In the multiple-strategy model (bottom right), each player can simultaneously belong to different groups of players with given weights. Their decisions are given by the weighted average of the corresponding rows in the matrix of cooperation probabilities.

### 3.3 Game metadata and prior modeling

A critical aspect in the modeling process (especially if we have limited data) is to specify how the *a priori* information we have about players and games affects the plausibilities of model parameters, namely the strategy matrices  $\mathbf{p}$  and the group membership vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\eta}$ . In our case, we only have auxiliary information (metadata) on the games. Therefore, we make no *a priori* assumptions about which values for  $\mathbf{p}$  and  $\boldsymbol{\theta}$  are more plausible. In contrast, we expect games with similar payoffs ( $S, T$ ) to be regarded by players as similar. This means that games that are neighbors in the  $TS$ -plane are more likely to have similar membership vectors. We model this expectation by choosing a prior distribution that introduces an exponential penalty when membership vectors of neighboring games are dissimilar

$$P(\boldsymbol{\eta}) \propto \exp \left[ -\alpha \sum_{\langle gg' \rangle} (1 - \eta_g \cdot \eta_{g'}) \right]. \quad (3.3)$$

Here, the sum runs over all pairs of games that are nearest-neighbors in the  $TS$ -plane, that is that differ by plus or minus one in  $S$  or  $T$  (but not both; Fig. 3.1c), and  $\alpha \geq 0$  is a parameter that we call the game aggregation factor. Note that  $\alpha$  plays a similar role as the interaction in Ising, Potts and N-spin models [39, 40], so that for  $\alpha = 0$  we recover the uniform prior, whereas increasing values of  $\alpha$  make it more likely that neighboring games have similar mixing vectors. Note that in the single-strategy model, Eq. (3.3) is in reality a prior over partitions of games into groups  $\pi_g$ , since, as for the players, the model considers only the set of membership vectors  $\boldsymbol{\eta}$  that result in disjoint partitions.

These models are reminiscent of existing models, but differ in important aspects. The single-strategy model is the formalization of the idea that there exist “behavioral phenotypes,” and that decisions depend only on those phenotypes [31, 32, 33]. Unlike previous

work, however, we do not assume that the groups of games are known *a priori*. The idea of allowing the existence of more than one strategy is reminiscent of the population strategy models used in evolutionary game theory in which competing strategies can coexist within a population [38]. The difference is that, in our case, each individual is allowed to simultaneously consider more than one strategy to make decisions, rather than having a population in which different strategies are represented. Moreover, as we previously mentioned, in our approach we make no assumptions about which strategy (pure or mixed) players are using in each game; we can determine whether players are using mixed strategies or not as an outcome of the inference process. This possibility is especially interesting in the analysis of real data since the empirical relevance of this concept has been so far hard to prove [41, 42].

Also unlike previous models, we are able to use the available game metadata, which we introduce in the model through the prior distribution of model parameters, an approach that is reminiscent of what has been used in networks [43] (an alternative is to model the metadata together with the data [44]). Modeling game similarity through the scalar product of game membership vectors opens the door to modeling other network-like systems whose nodes, like games, are embedded in a low-dimensional space.

For the single strategy model, we introduce the prior distribution 3.3 in the Eq. 2.2 for the posterior probability:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o) = \frac{1}{Z} \int d\mathbf{p} P(A^o | \boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}) P(\boldsymbol{\eta}). \quad (3.4)$$

Because the prior distribution only depends on  $\boldsymbol{\eta}$  but not on  $\mathbf{p}$  or  $\boldsymbol{\theta}$ , we can take  $P(\boldsymbol{\eta})$  out of the integral. Therefore, the solution to the integral is identical to the case with uniform priors (see Eq. 2.6 in the previous chapter) and the "Hamiltonian" only changes by an additive constant introduced precisely by the prior distribution:

$$\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{k\ell} \left[ \ln(n_{k\ell} + \Lambda - 1)! - \sum_{\lambda=1}^{\Lambda} \ln(n_{k\ell}^{\lambda})! \right] + \alpha F, \quad (3.5)$$

with  $F = \sum_{\langle gg' \rangle} (1 - \eta_g \cdot \eta_{g'})$  being the number of neighboring pairs of games that are not in the same group and  $\alpha$  the game aggregation factor. The effect of the prior distribution is thus to globally penalize partitions where neighbouring games belong to different groups, as mentioned in the previous section.

Furthermore, since only two actions are allowed in conventional games (Cooperation,  $\mathcal{C}$  or Defection,  $\mathcal{D}$ ), we have  $\Lambda = 2$ . Therefore, the Hamiltonian further simplifies to:

$$\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{k\ell} \left[ \ln(n_{k\ell} + 1)! - \ln(n_{k\ell}^{\mathcal{C}})! - \ln(n_{k\ell}^{\mathcal{D}})! \right] + \alpha F. \quad (3.6)$$

This is the function we need to minimize to find the best partition into groups under the game prior condition.

In the case of the mixed strategy model, we started considering the natural logarithm of the likelihood (or posterior, since they were equivalent under the uniform prior condition) in the previous chapter (see Eq. 2.14). Considering now the prior distribution, we take the

logarithm of the posterior probability. Simplifying to two types of links (Cooperation and Defection), we get:

$$\log P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}|A^o) = \sum_{i \in U; g \in \mathcal{C}_i} \log p_{ig} + \sum_{i \in U; g \in \mathcal{D}_i} \log(1 - p_{ig}) - \alpha \left( 1 - \sum_{\langle gq \rangle} \boldsymbol{\eta}_g \cdot \boldsymbol{\eta}_q \right). \quad (3.7)$$

Here,  $U$  represents the total set of players,  $\mathcal{C}_i/\mathcal{D}_i$  is the total set of observed games in which player  $i$  cooperates/defects, and  $p_{ig}$  is a short notation for the probability that player  $i$  cooperates at game  $g$ ,  $p_{ig}(\mathcal{C}) = \sum_{k, \ell} \theta_{ik} \eta_{g\ell} p_{k\ell}$ . Since there are only two available actions and the probabilities are normalized,  $p_{ig}(\mathcal{D}) = 1 - p_{ig}(\mathcal{C})$ .

Introducing the auxiliary distribution  $\omega_{ig}^{\mathcal{C}/\mathcal{D}}(k, \ell)$  and the Jensen's inequality in the posterior with the prior distribution, we have:

$$\begin{aligned} \log P(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbf{p}|A^o) &\geq \sum_{i \in U; g \in \mathcal{C}_i} \omega_{ig}^{\mathcal{C}}(k, \ell) \log \frac{\theta_{ik} \eta_{g\ell} p_{k\ell}}{\omega_{ig}^{\mathcal{C}}(k, \ell)} \\ &+ \sum_{i \in U; g \in \mathcal{D}_i} \omega_{ig}^{\mathcal{D}}(k, \ell) \log \frac{\theta_{ik} \eta_{g\ell} (1 - p_{k\ell})}{\omega_{ig}^{\mathcal{D}}(k, \ell)} - \alpha \left( 1 - \sum_{\langle gq \rangle} \boldsymbol{\eta}_g \cdot \boldsymbol{\eta}_q \right), \end{aligned} \quad (3.8)$$

where the equality holds again for:

$$w_{ig}^{\mathcal{C}}(k, \ell) = \frac{\theta_{ik} \eta_{g\ell} p_{k\ell}}{\sum_{k', \ell'} \theta_{ik'} \eta_{g\ell'} p_{k'\ell'}} \quad (3.9)$$

$$w_{ig}^{\mathcal{D}}(k, \ell) = \frac{\theta_{ik} \eta_{g\ell} (1 - p_{k\ell})}{\sum_{k', \ell'} \theta_{ik'} \eta_{g\ell'} (1 - p_{k'\ell'})}. \quad (3.10)$$

The approach for the maximization step is also the same as in Eq. 2.19, and the presence of the prior information only affects the equation for  $\boldsymbol{\eta}$ . Therefore, we take derivatives of the log posterior under the normalization constraints imposed by the lagrange multipliers:

$$\theta_{ik} = \frac{\sum_{g \in \mathcal{C}_i} \sum_{\ell} w_{ig}^{\mathcal{C}}(k, \ell) + \sum_{g \in \mathcal{D}_i} \sum_{\ell} w_{ig}^{\mathcal{D}}(k, \ell)}{d_i} \quad (3.11)$$

$$\begin{aligned} \eta_{g\ell} &= \frac{\sum_{i \in \mathcal{C}_g} \sum_k w_{ig}^{\mathcal{C}}(k, \ell) + \sum_{i \in \mathcal{D}_g} \sum_k w_{ig}^{\mathcal{D}}(k, \ell)}{d_g + \alpha \sum_{r \in \partial g} \boldsymbol{\eta}_r \cdot \boldsymbol{\eta}_g} \\ &+ \frac{\alpha \sum_{r \in \partial g} \eta_{r\ell} \eta_{g\ell}}{d_g + \alpha \sum_{r \in \partial g} \boldsymbol{\eta}_r \cdot \boldsymbol{\eta}_g}, \end{aligned} \quad (3.12)$$

where  $d_i = \sum_{g \in \mathcal{C}_i} \sum_{k\ell} w_{ig}^{\mathcal{C}}(k, \ell) + \sum_{g \in \mathcal{D}_i} \sum_{k\ell} w_{ig}^{\mathcal{D}}(k, \ell)$  and  $d_g = \sum_{i \in \mathcal{C}_g} \sum_{k\ell} w_{ig}^{\mathcal{C}}(k, \ell) + \sum_{i \in \mathcal{D}_g} \sum_{k\ell} w_{ig}^{\mathcal{D}}(k, \ell)$ , and  $\mathcal{C}_g$  and  $\mathcal{D}_g$  are the set of users that cooperate or defect in game  $g$ , respectively.

Finally, for  $p_{k\ell}$ , we get:

$$p_{k\ell} = \frac{\sum_{(i,g) \in \mathcal{C}} w_{ig}^{\mathcal{C}}(k, \ell)}{\sum_{(i,g) \in \mathcal{C}} w_{ig}^{\mathcal{C}}(k, \ell) + \sum_{(m,n) \in \mathcal{D}} w_{mn}^{\mathcal{D}}(k, \ell)}, \quad (3.13)$$

We have thus shown how to include a prior information about the system into our two inference approaches. In the first part of the section, we took some knowledge we had and modeled it into a prior distribution; we first discussed how that knowledge could affect the dynamics of the system and we then modeled that knowledge in a way that is consistent with our bayesian method for finding the most plausible partitions both in the single and mixed strategy approaches. Importantly, we included a parameter to tune the relative intensity of the prior in a continuous way.

In the second part of the section, we extended the formalism presented in the previous chapter to include a prior distribution. We did this for the membership vectors of games, but this can equally be done for any other model parameters, as long as the prior can be modeled.

It could be argued that there is some sort of arbitrariness in our choice of the prior distribution. However, we can validate it if we test the predictive power of our models and, as we will see in the next section, our assumptions prove to be successful: for non-negligible values of the game aggregation parameter  $\alpha$  we see a significant increase in the predictive power of both approaches. Moreover, since the main effect of the prior distribution is to remove statistical fluctuations, interpretability becomes much easier; again, prediction and interpretability are two sides of the same coin and our results underscore this.

## 3.4 Predictive power

To test the predictive power of our models, we use cross-validation. The idea of cross-validation schemes is to divide a dataset in two splits or sets: one split (usually larger) is used to train the algorithm such that it outputs a predictive model. The other one is used to test the ability of the resulting model at making predictions.

Here, we implement a 5-fold cross-validation scheme: we divide the data in five equally-sized splits, and then use four splits as a training and the remaining split as a test to assess the capacity of the model to predict unobserved data. We repeat this for the five possible train-test combinations.

The idea is to provide enough data to train the model in each iteration, but at the same time run over the whole dataset to prevent statistical fluctuations.

### 3.4.1 Baseline Model

We start by studying the predictive power of the model proposed in Ref. [31], which we will consider as a baseline prediction. In this model, games are divided into four fixed groups (harmony game, snowdrift game, stag hunt game, and prisoner's dilemma). Each user  $i$  is then characterized by a strategy vector  $v_i$  that quantifies their propensity to cooperate in each of the four types of games. In the baseline model, players are grouped according to

the similarity in their strategy vectors using k-means. For each training set, we find the player groups, and estimate  $p_{k\ell}$ , the probability that players in group  $k$  cooperate in games in group  $\ell$ , as the frequency with which players in group  $k$  cooperate in games in group  $\ell$  in the empirical data. Then, we use these frequencies to predict cooperation in the test data, so that if  $p_{k\ell} > 0.5$  then the prediction is that all users in group  $k$  will cooperate in games in group  $\ell$ . We obtain an average predictive accuracy of  $0.683 \pm 0.005$  (Fig. 3.3a).

### 3.4.2 Single-strategy models: maximally predictive partitions reveal perception of games by players

Next, we study the predictive power of the single-strategy model as a function of the game aggregation parameter  $\alpha$ , which controls how strongly neighboring games are pushed into the same group. For the sake of consistency, we use the same 5-fold cross validation scheme as before. For each split, we obtain the optimal partitions of players and groups from Eq. (2.8) and, as before, use the observed cooperation frequencies of groups of players in groups of games to make predictions on the test set.

We find that the predictive power of the model increases with  $\alpha$  and reaches its maximum for  $\alpha = 2$ , at which point it is significantly more predictive than the baseline model with a predictive accuracy of  $0.714 \pm 0.008$  (Fig. 3.3 a).

A close inspection of the optimal partitions for players and games reveals that the game aggregation factor has an effect on the partition of both players and games into groups (Fig. 3.3 b,c–f). With respect to the number of groups of players, we observe that the number of groups decreases as we increase  $\alpha$ , and stabilizes for  $\alpha > 2$  at around 20 groups. Note that many of these player groups are small since 5 or 6 groups typically account over 50% of the players (see bottom row of Fig. 3.3 c–f). With respect to the partitions of games we observe two noteworthy aspects. First, that the absence of a prior for game memberships ( $\alpha = 0$ ) makes game groupings (and as a surrogate player’s groupings) too susceptible to fluctuations, which results in low predictive power. Second, that as  $\alpha$  increases the prior helps disregard statistical fluctuations in favor of a well-defined structure of game groups within the  $TS$ -plane, leading to a higher predictive performance. Despite the fact that the prior acts on the games alone, it also leads to a lower number of player groups because with fewer game groups the number of possible strategy vectors is also smaller. Interestingly, games fall into groups defined by the difference between sucker and temptation payoffs  $\Delta = (T - S)$ , and are qualitatively different from the four regions that follow from game-theoretic considerations (Fig. 3.1b). More specifically, at the optimal aggregation factor  $\alpha = 2$  we observe three regions that are distributed in a roughly diagonal pattern. These regions correspond approximately to: (i)  $S > T$ , where most players cooperate; (ii)  $S < T - P = T - 5$ , where most players do not cooperate (although some do, including a few that always cooperate); (iii) the intermediate region where some cooperate and others do not. These groups are consistent with the observations described in Ref. [31], but in our analysis arise naturally from the rigorous comparison of models, and get incorporated in the models. In this sense, our approach illuminates the way in which individuals are “predictably irrational” [45].

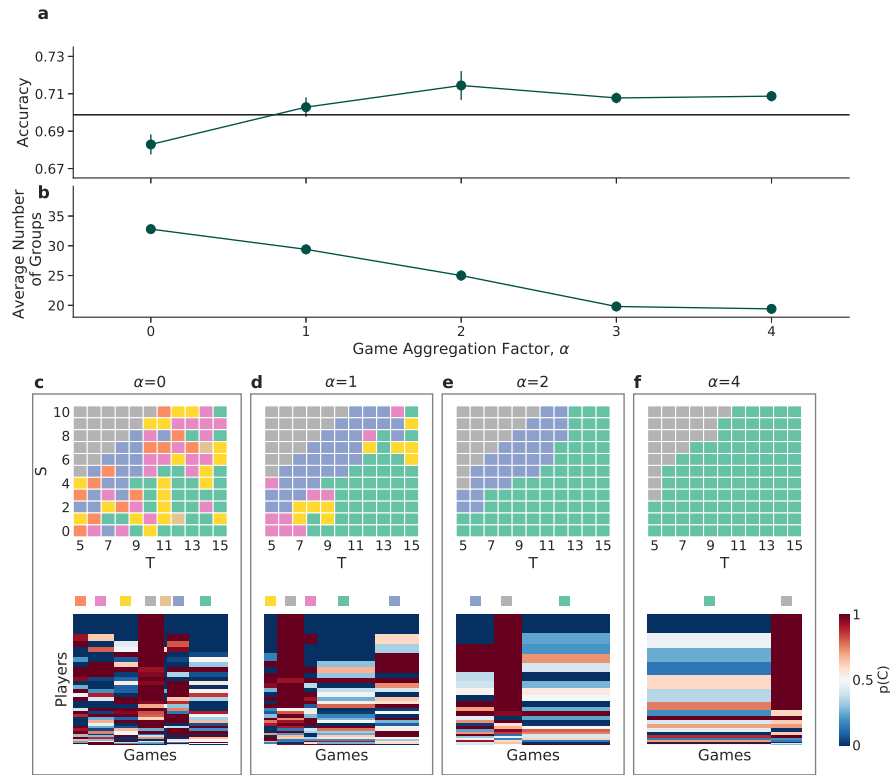


Figure 3.3: **Single-strategy model.** (a) Predictive accuracy of the single-strategy model as a function of the game aggregation factor  $\alpha$ . Each point represents the average of a 5-fold cross-validation; error bars indicate the standard error of the mean. The solid black line represents the accuracy of the baseline model (see text and Ref. [31]). (b) Average number of groups of players  $\langle N_g \rangle$  as a function of the game aggregation factor  $\alpha$ . Each point represents the average of the number of groups for players identified for each of the 5 folds; error bars indicate the standard error of the mean. (c-f) Groups of games in the  $TS$ -plane (top) and cooperation matrix  $\mathbf{p}$  (bottom) for aggregation factors  $\alpha$  equal to: (c) 0, (d) 1, (e) 2, (f) 4. In the game plots, each color indicates a different group of games in the most plausible partition as obtained from Eq. (2.8) for each value of  $\alpha$ . In the cooperation matrices, each row corresponds to a group of players  $k$  (with height proportional to the number of players in the group), and each column to a group of games  $\ell$  as indicated by the colors at the top (with width proportional to the number of games in the group). Each element  $p_{k\ell}$  represents the probability that an individual in group  $k$  cooperates when playing a game in group  $\ell$ , with dark red meaning always cooperate and dark blue always defect. The game groups and the cooperation matrices correspond to one of the folds in the 5-fold cross-validation, but are very consistent across folds (see Fig. 5.3).

To further investigate the collective behavior of players, we focus on the phenotypes associated to the three largest groups of players identified for  $\alpha = 1, 2$  (Fig. 3.4). In both cases we see that the largest group is characterized by two facts: i) players only distinguish between two types of games ( $S \geq T$  and  $S < T$ ); and ii) players display pure strategies in these games: always cooperate in games with  $S \geq T$  and always defect in games with  $S < T$  (Fig. 3.4). Interestingly, this phenotype is precisely the envious phenotype identified in [31], but in our case it arises naturally from our model-selection criteria without having to make any assumptions about the structure of the  $TS$ -plane or about the number of player groups. The two remaining most common phenotypes for  $\alpha = 1, 2$  also show that players fully cooperate in games with  $S \geq T$ . However, the cooperation patterns for  $S < T$  cannot be mapped directly into any behavioral phenotype described previously in the literature, and, despite being strongly correlated with  $\Delta$  and leading to better predictions, they are not as informative as the most common phenotype. In part, this is due to the large number of observed phenotypes (Fig. 3.3c–f); in the following section, we show that the multiple-membership model provides a more parsimonious and straightforward description of this variety of phenotypes.

### 3.4.3 Multiple-strategy models are more predictive and easier to interpret than single-strategy models

Finally, we investigate the predictive power of models in which players are allowed to use mixtures of strategy vectors. As we show in Fig. 3.6, we find that, again, predictive performance grows with the game aggregation  $\alpha$  and saturates after  $\alpha = 3$ . Remarkably, the multiple-strategy model is, in all cases, significantly more predictive than the single-strategy model, with a maximum predictive accuracy of  $0.744 \pm 0.007$ . This is consistent with our remarks about the predictive power of both approaches in the previous chapter, namely that MMSBM are generally more predictive than regular SBM. Likewise, it suggests that our data may be better explained in terms of a mixed membership approach.

Let us consider first the effect of mixed-membership on game grouping. We start looking at how the membership of games is spread across all  $L = 4$  latent groups for different values of  $\alpha = 0, 2, 4, 8$ . To that end, we compute the Shannon Entropy  $H_g$  for every membership vector  $\eta_g$ :

$$H_g = \sum_{\ell=1}^L \eta_{g\ell} \log_L(\eta_{g\ell}), \quad (3.14)$$

where low values of  $H_g$  imply memberships concentrated in one or two groups, whereas values close to 1 would indicate a membership equally distributed among the four groups. The distribution of entropies of membership vectors (Fig. 3.5) shows that most games have a very low Shannon Entropy and that the main effect of  $\alpha$  is actually to further reduce it. If we plot the top membership of each game in the T-S plane (see Fig. 3.6), we see that the membership matrix strongly resembles the game classification for single-strategy models in Fig. 3.3, especially for  $\alpha > 0$ . This confirms that the perception of games by players following the difference  $\Delta = (T - S)$  is robust and that mixed-membership does not play a major role in game grouping, since games end up belonging mostly to a single

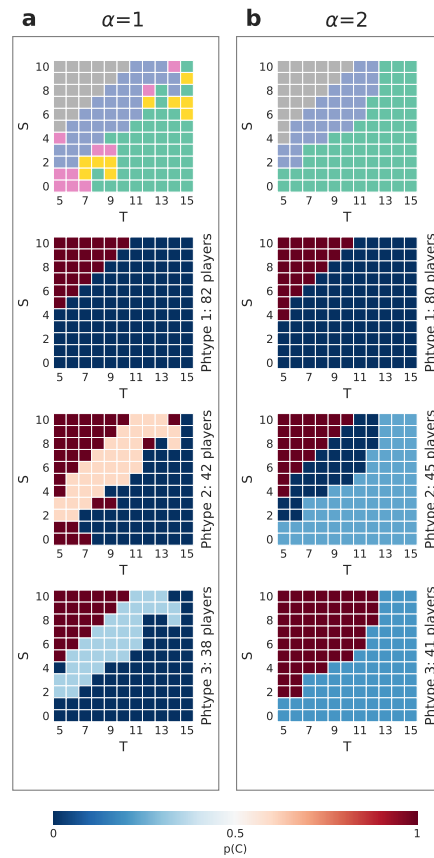


Figure 3.4: **Most common behavioral phenotypes for single-strategy models.** Game groups (top; as in 3.3 , and cooperation probability in each game group for the three largest player groups (or behavioral phenotypes) for: (a)  $\alpha = 1$ , (b)  $\alpha = 2$ . **The groups and cooperation probabilities correspond to one of the folds in the 5-fold cross-validation, but are very consistent across folds.**

well-defined group as in the single-strategy model (Fig. 3.6). Therefore, the increase in predictive performance must be due to the multiple membership of players, that is, to the fact that players are best described as not making decisions following a unique strategy but rather using a combination of strategies. Indeed, we find that the majority of players use a mixture of strategies (center row in Fig. 3.6), and that three global strategies are enough to make the most accurate predictions of players' decisions (bottom row in Fig. 3.6). Note that, unlike the single-strategy model, we need to fix the number of groups; but we find that such a small number of strategies (i.e.,  $K = 3$ ) provides the most accurate predictions



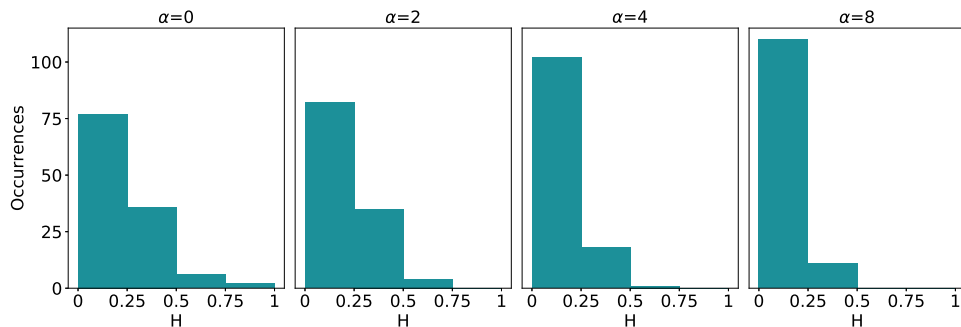


Figure 3.5: **Distribution of Shannon Entropies of the game membership vectors  $\eta$ .** Each bin represents the number of occurrences of games significantly belonging to one, two, three or four memberships.

(Fig. 5.2). Interestingly, these global strategy vectors are for the most part combinations of pure strategies in which players either fully cooperate or fully defect in games. This is an example of a common phenomenon usually seen in MMSBMs and briefly described in the previous chapter. That is, the fact that P matrices (the set of three strategy vectors in our case) tend to have numerical values closer to 1 and 0 than in regular SBMs. In general, this makes interpretability much easier in terms of the interactions between latent groups. In this particular case, we can much better understand the behavioural patterns of groups of players. For example, we can clearly identify the envious strategy (to cooperate only in games with  $S \geq T$ ), which is also present in the single-strategy model and in Ref. [31]. Additionally, aggregating over the  $\theta$  membership vectors, we see that it is slightly more common than the others, also in consistency with the single strategy approach Fig. 3.6 b-e). The other two strategies correspond to: (i) a more rational strategy that leads to cooperation for half of the games (including all harmony games) and to defection for the other half (including all prisoner's dilemma games); (ii) a strategy that accounts for non-rational behaviors, including incomplete cooperation in harmony games and full cooperation for most other games, including prisoner's dilemma games. This last strategy may seem counterintuitive but arises from the need to assign non-zero probability to all behaviors; without this strategy, for example, any observed non-cooperation in games with  $S \geq T$  would lead to zero likelihood. It may seem surprising, however, that this deviant strategy is used in as many as 25% of all the decisions

### 3.5 Discussion and conclusions

In this chapter, we have explored the power of group-based models to predict decisions made by individuals in simple classes of dyadic games that involve strategic thinking. Such decisions are known to deviate from the rationally expected behavior. However, our analysis proves that they still are highly predictable (74% of the decisions can be correctly predicted)

and that group-based models are good models of strategic decision making.

More importantly, proposing interpretable models of human behavior and comparing them in terms of their predictive accuracy sets the bases for advancing the social sciences on solid grounds [30]. In this regard, we have shown that the most explanatory groupings of games reveal the perception of games by players, which differs from game-theoretical expectations. Our approach also gives the most explanatory cooperation strategies followed by players, and suggests that models in which players are allowed to use multiple strategies (rather than sticking to a single strategy) are more predictive than those models in which players are restricted to a single strategy. Multiple-strategy models are also more parsimonious in that they summarize the wide variety of phenotypes suggested by single strategy models as combinations of a small number of simple strategies. In fact, the combination of these two factors (perception of games by players and multiple strategies) accounts well for the rich variety of phenotypes observed in real data.

More broadly, we believe that our approach and models can be used to analyze many other behavioral experiments and datasets. Indeed, whenever humans face distinct (discrete, non-overlapping) situations and take distinct actions, their decisions can be modeled using the exact same approaches we have proposed here.

---

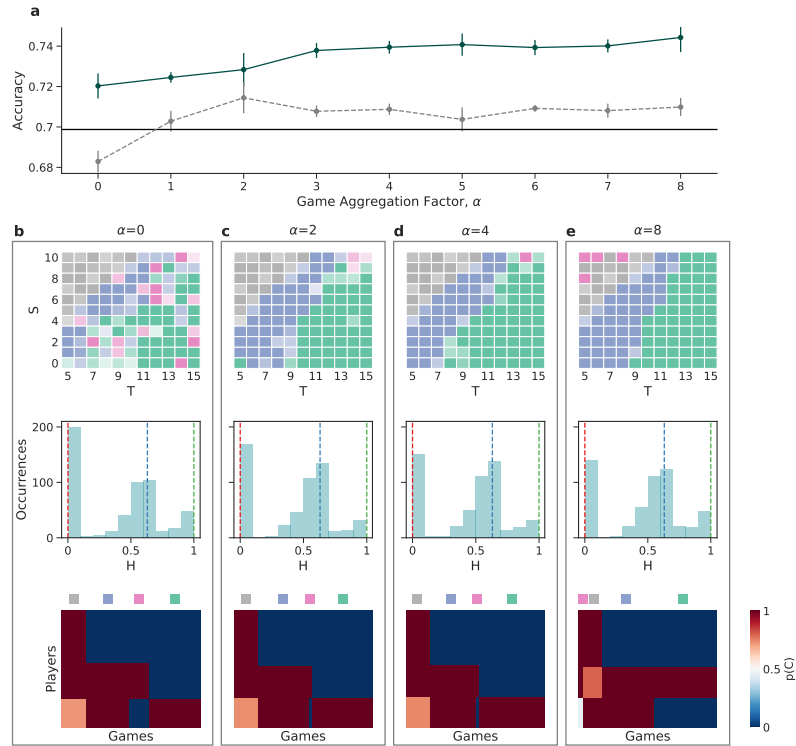


Figure 3.6: **Mixed-strategy model.** (a) Predictive accuracy of the multiple-strategy model (green) as a function of the game aggregation factor  $\alpha$ . We show results for  $K = 3$  latent groups of players and  $L = 4$  latent groups of games; increasing the number of groups did not result in an increase in accuracy (see Fig. 5.2). Each point represents the average of a 5-fold cross-validation; error bars indicate the standard error of the mean. Grey symbols represent the predictive accuracy of the single-strategy model as in Fig. 3.3 a, and the solid black line represents the accuracy of the baseline model. (b-e) Top group memberships for each game (top), distribution of the entropies of player membership vectors (middle), and cooperation matrices  $\mathbf{p}$  (bottom) for  $\alpha = 0, 2, 4, 8$ . In the game membership plots, each color indicates a different group of games in the most plausible model as obtained from Eqs. (3.11)-(3.13). The saturation of the color indicates how distributed a game is among groups, so that games with multiple memberships are paler (Fig. 3.5). In the distribution of the player entropies, players with a single strategy vector have entropy  $H = 0$  (red dashed line); players that mix two strategy vectors with equal weights have  $H = \log_3 2 = 0.63$  (blue dashed line); and players that mix three strategy vectors with equal weights have  $H = 1$  (green dashed line). In the cooperation matrices  $\mathbf{p}$ , each row corresponds to a group of players  $k$ , and each column to a group of games  $\ell$  as indicated by the colors at the top. The height of player group  $k$  is proportional to the effective number of players in that group  $\sum_i \theta_{ik}$ . The width of game group  $\ell$  is proportional to the effective number of games in that group  $\sum_g \eta_{g\ell}$ . Each element  $p_{k\ell}$  represents the probability that an individual in group  $k$  cooperates when playing a game in group  $\ell$ , with dark red meaning always cooperate and dark blue always defect. The groups, entropies, and cooperation matrices correspond to one of the folds in the 5-fold cross-validation, but are very consistent across folds (see Fig. 5.4).

UNIVERSITAT ROVIRA I VIRGILI

STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

---

## Chapter 4

# An Application to Human Microbiology

### 4.1 Introduction

The study of the human microbiome has become very popular during the last decade. This is partially due to the recent availability of data provided by large scale experiments such as the Human Microbiome Project (HMP) [8], [9]. These experiments have shed light on the importance of microbiome in multiple aspects. For instance, many pieces of evidence suggest that the microbial system is tightly related to several immune and digestive functions and that it may be related to certain health disorders such as Inflammatory Bowel Disease (IBD) and other autoimmune diseases [46], [47], [48]. However, the microbiome is a very complex system that interacts at different scales with human biology, hence making it difficult to understand its functioning.

One interesting question in this regard is whether there are universal traits in the human microbiome. In particular, it has been proposed that there might be universal types of human gut microbial profiles according to their compositions. These classifications are usually referred to as enterotypes. However, results on the existence of enterotypes have been controversial [49], mainly because it is difficult to find regularities in microbial data. Here, we redefine this idea by proposing latent enterotypes, a more general and abstract version of the conventional ones. The idea is that hosts have microbial profiles that are a combination of latent enterotypes in different proportions. We believe that latent enterotypes could be more consistent with the extreme variability of microbial data. Taking real datasets from open repositories, we use MMSBM to find groups of microbes and patients upon which latent enterotypes are defined.

We validate our results by testing the ability of our model to predict abundances of missing microbial species in individual patients. We find that our approach is significantly more predictive than previously used ones. We also find a correlation between the predictability of individual patients and microbes and the degree of mixing among different groups.

---

Importantly, we also observe a significative correlation between membership vectors of microbes and their taxonomic classifications. Finally, we observe a well defined ecological order among latent enterotypes called nestedness[50]. The nestedness corresponds to an increasing degree of specialization on the enterotypes, in the sense that some enterotypes contain fewer groups of microbes than others.

Altogether, our results show that it is possible to find universal patterns of individual microbial profiles, that these patterns are similar to those found in some ecological networks and that it is possible to make predictions upon them, both at an individual and a global scale. We also see that patients and microbes have varying degrees of predictability and that our model parameters can partially capture the taxonomic proximity of microbial species.

## 4.2 Data

We have datasets for the five human gut microbiome studies shown in Table 4.1 . Each dataset consists of an Operational Taxonomic Unit (OTU) matrix  $M$ , where columns represent human samples or patients and rows correspond to microbial species. Each element  $M_{ij}$  corresponds to the relative abundance of the microbial species  $i$  relative to the host sample  $j$ . For practical purposes, we have considered abundances below  $10^{-4}$  as negligible. Also, we have removed microbes that were present in less than 5 % of all patients. In addition to OTU matrices, we have taxonomic information on the microbes provided by the same sources. For each microbre we have the Phylum, Class, Order, Family, and Genus level.

Table 4.1: Datasets

Dataset	Study Name	Microbes	Patients
S-8	Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis [51]	128	107
V-10	Alterations of the human gut microbiome in liver cirrhosis [52]	137	92
V-22	Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity [53]	134	467
V-23-24	Structure, function and diversity of the healthy human microbiome & Strains, functions and dynamics in the expanded Human Microbiome Project [54], [55]	118	222
V-25	Personalized Nutrition by Prediction of Glycemic Responses [56]	144	883

We model OTU matrices as bipartite multilink networks. Here, nodes correspond to patients and microbes, and edge values correspond to relative abundances in each host. Since relative abundances are continuous and our models require a discrete number of edge

values or categories, we need to discretize relative abundances. We choose to discretize them according to orders of magnitude. In all five datasets, abundances range from  $10^{-4}$  to  $10^{-1}$ , so we consider three types of links or discrete abundances corresponding to: null or negligible (0) ( $M_{ij} < 10^{-4}$ ), low ( $\mu$ ) ( $10^{-4} \leq M_{ij} \leq 10^{-3}$ ), and high ( $\nu$ ) abundance ( $M_{ij} > 10^{-3}$ ). This discretization is more informative from a clinical point of view and it allows us to keep a connection with the real values of abundances. This connection would not have existed with other discretization schemes such as quantiles. On the other hand, it could be argued that  $9.99 \cdot 10^{-3}$  and  $1 \cdot 10^{-4}$  are too close to belong to different categories, but this problem will arise in any other discrete categorization we could think about.

## 4.3 Methods

### 4.3.1 Enterotypes

A common question in the study of the human gut microbiome is whether people can be classified into a reduced number of groups according to similarities in their gut microbial ecosystems. Each one of these groups is characterized by an enterotype: a specific profile of microbe concentrations in the gut [57], [58], [59]; for instance, a set of clusters of microbes with different average abundances each. The idea of enterotypes has been often questioned; the lack of a standardized procedure to define them [49], and the extreme variability of microbial composition related to diet [60],[61], [62], age [63], health state [64], or simply time [65], challenges the idea of universal rigid patterns in gut microbial composition. However, it is certainly plausible that some regularities exist across humans, since it appears that the human gut microbiome performs critical biological functions [66], [67].

A possible explanation that would reconcile the existence of enterotypes with the observed variability is that patients are mixtures of enterotypes in different proportions. Therefore, changes according to diet, age, time or even health state could be explained as transitions between enterotypes. In that regard we propose the idea of latent enterotypes. Latent enterotypes are universal microbial profiles, but they do not generally correspond to a group or population of patients. Instead, each patient belongs to several latent enterotypes with different weights (see Fig. 4.1). The same logic applies to microbes. Each microbial species can belong to many groups in different proportions. In reality, our model identifies latent groups of patients and latent groups of microbes. Thus, we define a latent enterotype (hereafter, enterotype) as an abundance pattern. Each pattern associated to a specific latent group of patients is characterized by the abundances of the latent groups of microbes.

### 4.3.2 Mixed Membership Stochastic Block Models

The idea of latent enterotypes is formally analogous to the idea of the multiple strategy approach discussed in the previous chapter. Players and games are thus replaced by patients and microbes. Therefore, given  $K$  groups of patients and  $L$  groups of microbes, we have

---

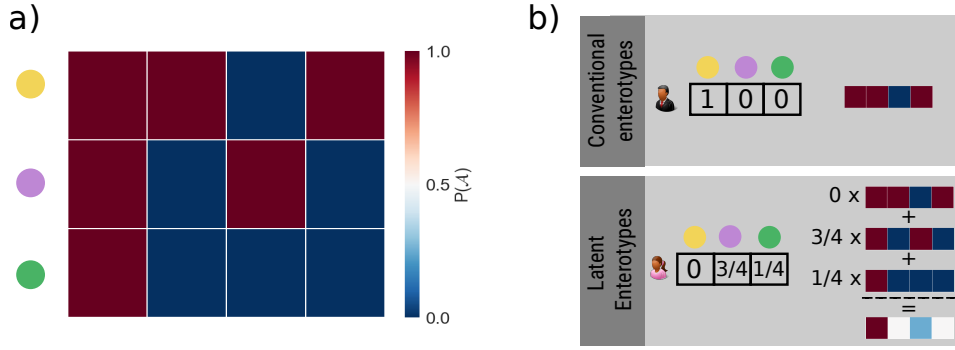


Figure 4.1: **Conventional enterotypes and latent enterotypes.** a) Matrix of probabilities. Rows correspond to groups of patients and columns represent groups of microbes. Colors represent the probability that a group of patients hosts a group of microbes. Each entire row represents an enterotype or microbial profile. b) Example of a conventional enterotypes and latent enterotypes. In a conventional enterotype, each patient corresponds to a microbial profile or enterotype. When considering latent enterotypes, a patient can have an enterotype that is a linear combination of latent enterotypes.

mixed-membership vectors for each patient  $p$  and microbe  $m$ :

$$\theta_p = [\theta_p^1, \theta_p^2, \dots, \theta_p^K] \quad \eta_m = [\eta_m^1, \eta_m^2, \dots, \eta_m^L] \quad (4.1)$$

where  $\theta_p$  and  $\eta_m$  are normalized, so that:

$$\sum_{k=1}^K \theta_k^p = 1 \quad \sum_{l=1}^L \eta_l^m = 1 \quad (4.2)$$

As for the interaction between groups of patients and microbes, instead of latent strategies, we have latent enterotypes. Latent enterotypes in our model correspond to the vector of probabilities of connection of a group of patients to all  $L$  groups of microbes.

In order to validate the existence of latent enterotypes, we test the ability of our method to predict unobserved microbial abundances in patients. That is, given a patient  $p$  and a microbe  $m$ , we ask ourselves what is the probability that microbe  $m$  is present in patient  $p$  with an abundance  $\mathcal{A}$ .

This probability is given by:

$$\Pr[a_{pm} = \mathcal{A}] = \sum_{k,\ell} \theta_{pk} \eta_{m\ell} p_{k\ell}^{\mathcal{A}} \quad (4.3)$$

with  $p_{k\ell}^{\mathcal{A}}$  being the probability that the group of patients  $k$  contains the group of microbes  $\ell$  with an abundance  $\mathcal{A}$ <sup>1</sup>. Our estimation of the abundance of microbe  $m$  in patient  $p$ , is

<sup>1</sup> Note that  $p^0$ ,  $p^\mu$ , and  $p^\nu$  are matrices of  $K \times L$  dimensions. Rows in these matrices represent exactly the latent enterotypes



given by the maximum numerical value of  $\Pr[a_{pm} = 0]$ ,  $\Pr[a_{pm} = \mu]$ , and  $\Pr[a_{pm} = \nu]$ , i.e. :

$$\text{Prediction} = \max_x \{\Pr[a_{pm} = x]\} \quad (4.4)$$

We thus look for the membership vectors  $\theta$  and  $\eta$  that yield the highest predictive power of unobserved abundances.

### 4.3.3 Inference of the most predictive latent enterotypes

Given an OTU matrix  $R^o$  of observed microbial abundances, we want to find the most plausible membership vectors  $\theta$ ,  $\eta$ , as well as the most plausible  $\mathbf{p}$  matrices.

Since we do not have any metadata that we could use as a prior to model the system, the formalism is just the particular case of the Mixed Membership Stochastic Block Model (MMSBM) with three types of links or ratings (see Chapter 1).

Therefore, the equations for the model parameters are:

$$\theta_{pk} = \frac{\sum_{m \in 0_p} \sum_{\ell} w_{pm}^0(k, \ell) + \sum_{m \in \mu_p} \sum_{\ell} w_{pm}^{\mu}(k, \ell) + \sum_{m \in \nu_p} \sum_{\ell} w_{pm}^{\nu}(k, \ell)}{d_p}, \quad (4.5)$$

$$\eta_{m\ell} = \frac{\sum_{p \in 0} \sum_k w_{pm}^0(k, \ell) + \sum_{p \in \mu} \sum_k w_{pm}^{\mu}(k, \ell) + \sum_{p \in \nu} \sum_k w_{pm}^{\nu}(k, \ell)}{d_m}, \quad (4.6)$$

$$p_{k\ell}^{\mathcal{A}} = \frac{\sum_{[(p,m) \in R^o | a_{pm} = \mathcal{A}]} \omega_{pm}(k, \ell)}{\sum_{(p,m) \in R^o} \omega_{pm}(k, \ell)}. \quad (4.7)$$

Here,  $0$ ,  $\mu$ , and  $\nu$  are the sets of low, medium, and high abundances observed in the OTU matrix  $R^o$ .  $0_p$ ,  $\mu_p$ , and  $\nu_p$  are the subsets of observed low, medium, and high abundances corresponding to the patient  $p$ . Correspondingly,  $0_m$ ,  $\mu_m$ , and  $\nu_m$ , are the subsets of patients that contain microbe  $m$  with low, medium, and high abundance.  $d_p$  and  $d_m$  are the total degrees of  $p$  and  $m$ , that is, the number of microbes in patient  $p$ , and the total number of patients in which microbe  $m$  is present with any type of abundance (low abundance included).

Finally,  $w_{pm}^{\mathcal{A}}(k, \ell)$  is a distribution representing the probability that patient  $p$  having the microbe  $m$  with relative abundance  $\mathcal{A}$  is due to  $p$  and  $m$  belonging to latent groups  $k$  and  $\ell$ :

$$\omega_{pm}^{\mathcal{A}}(k, \ell) = \frac{\theta_{pk} \eta_{m\ell} p_{k\ell}^{\mathcal{A}}}{\sum_{k'\ell'} \theta_{pk'} \eta_{m\ell'} p_{k'\ell'}^{\mathcal{A}}}. \quad (4.8)$$

## 4.4 Results

### 4.4.1 Cross validation experiments

In order to assess the prediction accuracy, we perform cross validation experiments with OTU matrices corresponding to the datasets described before. That is, we remove a given

number of relative abundances in the OTU matrix and use the rest of the matrix as an input to the model. Because we want to make predictions on individual patients, we always remove abundances from a single patient. Particularly, we do a 5-fold cross validation for each patient: we divide its microbial abundances in five equally sized groups at random, and remove one of those groups of abundances to test the performance of our model. We train our model with the remaining data in the OTU matrix (i.e. the four remaining groups of patients plus the whole information on all other patients). We repeat this process for all five groups in order to prevent statistical fluctuations.

#### 4.4.2 Baseline

To assess the advantages of the modeling approach we propose, we start by selecting a reference model with which we will obtain a baseline. We will use this baseline to compare our results. Specifically, we consider the pipeline described in [49] to obtain enterotypes from OTU matrices. In this work, enterotypes are found using multiple clustering approaches that combine 2 different methods, 5 distance metrics, and 9 different possible numbers of clusters (ranging from 2 to 10), yielding 90 clusterings in total. From these 90 possibilities, the best clustering is then selected measuring the performance on three different metrics: Silhouette width (SI), Prediction Strength (PS), and Jaccard threshold. There is a threshold for each of these metrics below which no clustering is generated. Note that this work was not intended to make predictions of unobserved abundances, but to find (conventional) enterotypes in a rigorous way.

We test the predictive power of the enterotypes produced by this pipeline with three different metrics: accuracy, recall, and precision. Because recall and precision are binary magnitudes, we reduce the problem to predicting whether a microbial species is present or not in a patient. To that end, we merge the low ( $\mu$ ) and high ( $\nu$ ) abundances into a single category. For simplicity, we will denote both categories as 0 and 1. Our goal is thus to correctly predict 0 and 1 abundances under the three aforementioned metrics. Accuracy is simply the fraction of correctly predicted data points (true positives or 1s and true negatives or 0s) out of the total (see Fig. 4.2). It is a particularly useful metric if the number of 1s is comparable to the number of 0s. However, that is not often the case, and thus accuracy can be a misleading metric. For instance, in very sparse datasets (OTU matrices with a large proportion of entries equal to 0), even the simplest baseline would have very high accuracy rates. To prevent this, we use the recall and precision. The recall measures the fraction of true positives out of the total existing ones in the test dataset. However, the recall on its own is not completely informative; we could come up with a method that simply overclassifies data points as 1s or positives. Therefore, the recall would be very high, but many predictions would be wrong due to a biased tendency to misclassify 0s as 1s. To circumvent this problem, we compute the precision, which measures the fraction of true positives out of all predicted positives. These three metrics are complementary and provide an overall picture of the performance of a method in a prediction task. In Fig. 4.3, we show the prediction performance together with the results of our model. For each patient, we measure her aggregated accuracy, precision, and recall over the five folds. We then take the average over all patients in the data set.

---

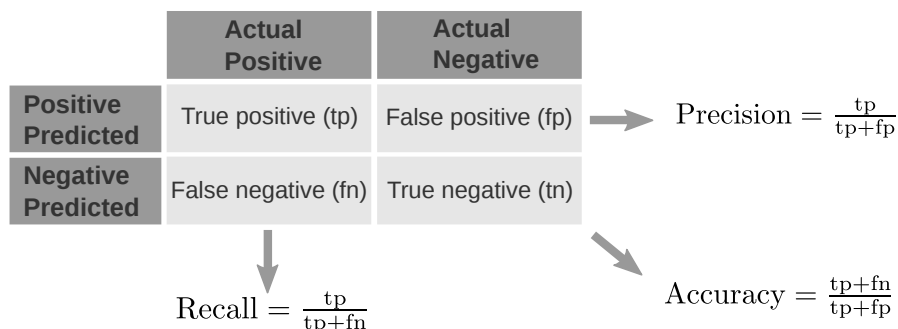


Figure 4.2: **Contingency matrix of a classification method** Elements in the diagonal are correctly predicted data points (true positives and true negatives). Off-diagonal elements represent misclassified data points either as false positives or false negatives. The precision measures the ratio of true positives out of all predicted positives, recall measures the fraction of true positives out of the total positives, and accuracy measures the ratio of correctly predicted observations.

### 4.4.3 Latent enterotype models

We now look at the results for our MMSBM latent enterotype model. For the sake of consistency, we apply the same cross validation scheme as before. First of all, we observe that the best predictive combination of parameters corresponds to considering 10 latent groups of patients ( $K = 10$ ) and 20 latent groups of microbes ( $L = 20$ ). Both in our model and the baseline, we observe that the performance is highest for the accuracy, slightly lower for the precision, and significantly lower for the recall (see Fig. 4.3). We speculate that this is due to the larger proportion of 0s or null abundances, which drives both models to correctly classify most 0s and probably to misclassify a considerable number of 1s abundances (especially those corresponding to former low abundances  $\mu$ , at least in our model). However, we observe that our model outperforms the baseline in all datasets and metrics, particularly for the precision and recall. The precision is slightly lower, but still very consistent across all five datasets. Finally, the recall is significantly lower than accuracy and precision, but over 0.7 in all datasets except V-23-24.

An interesting question is whether these predictive metrics vary across patients and microbes. That is, are there individuals or microbes where abundances are easier to predict? From a clinical point of view this is a relevant question, since patients harder to predict fall out of the common patterns.

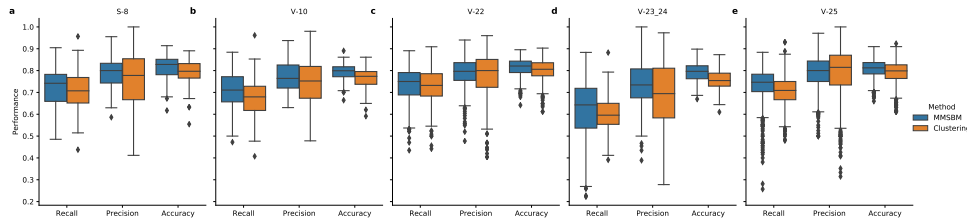


Figure 4.3: **Predictive performance [a-e]** Patient wise accuracy, precision, and recall for the latent enterotype model and the baseline for all five datasets considered

#### 4.4.4 Individual predictability

To answer this question, we explore the predictability of individual patients and microbes according to how their memberships are spread across latent groups. We measure the spreading of memberships by computing the Shannon entropy  $H_p/H_m$  of each patient  $p$ /microbe  $m$  membership vector  $\theta_p/\eta_m$ , respectively. Low values of Shannon entropy correspond to a membership being concentrated in one or two groups, while high values indicate a membership distributed across many groups. Fig. 4.4 [a, f], shows the distribution of Shannon entropies for patients and microbes, respectively. For patients, we observe that the majority of membership vectors have entropies that lie between  $H_p = 0.4$  and  $H_p = 0.6$ , with relatively few membership vectors having very low Shannon entropies and the highest entropy being  $H_p = 0.81$ . For microbes, Shannon entropies are more equally distributed. There are some membership vectors with very low entropies and  $H_m = 0.701$  is the highest value.

We then measure the relation of the Shannon entropies  $H_p$  and  $H_m$  with individual precision, recall, accuracy, and held-out loglikelihood of patients and microbes (see Fig. 4.4). Here precision, recall, and accuracy have the same meaning as in the previous section and they are also aggregated. The held-out loglikelihood represents the loglikelihood assigned to the abundances in the test set. The numerical values of the held-out loglikelihood range between 0 and 1. Low values indicate that the abundances in the test set are very likely according to the model. High values, on the contrary, indicate that these values are not expected by the model. Therefore, very low values are a good indicator of the performance of the model in the sense that it has captured the important patterns of the data.

In regards to the relation of these predictive measures and Shannon entropy, we see a general trend of  $H_p$  and  $H_m$  being negatively correlated with prediction metrics. That is, the lower the Shannon entropy, the higher the predictability of the patient/microbe. In the case of microbes, the correlation is stronger and more statistically significant, the only possible exception being the held-out loglikelihood (p-value=0.0118). In the case of patients, the correlation between  $H_p$  and precision is practically inexistent.

Such negative correlations are consistent across the four remaining datasets, although statistical significance varies and it is lowest in the datasets S-8 and V-10. Here, the p-values for the correlation between  $H_p$  and patient's precision are 0.952635 and 0.528857, respectively. In the case of  $H_m$  and the held-out loglikelihood, the p-values are 0.960426 and 0.493331,

respectively (see Supplementary Materials). Our results suggest that microbes and patients with lower Shannon entropies have a more regular pattern and are therefore easier to predict.

#### 4.4.5 Correlation between model parameters and taxonomic information

We have validated our modeling approach by testing the predictive power of the latent enterotype model. To further underpin the biological meaning of our results, we now check whether taxonomically close microbes are also close in latent space, i.e. if they have similar membership vectors. This would mean that our model is capable of capturing a biological information that is not directly provided. It would also suggest that taxonomic proximity is relevant to define the enterotypes. In order to test this, we consider all taxonomic levels: phylum, class, order, family, and genus. For each taxonomic level, we group microbes according to all existing taxa in that level and compare the similarity of membership vectors from the same and different taxa, i.e. the intra and inter-taxa similarities.

Suppose we have a taxonomic level  $T$  with  $t_1, \dots, t_n$  taxa. For each taxa  $t_k$ , we take all possible pairs of microbes' membership vectors  $(\eta_k^i, \eta_k^j)$ , such that  $i, j \in t_k$  and  $i \neq j$  and measure their similarity  $d_{ij}(t_k)$  calculating the euclidean distance between them. We repeat this process over all taxa and then compute the mean intra-taxa distance as:

$$\langle d_{intra} \rangle = \frac{1}{N_p^{intra}} \sum_{k=1}^n \sum_{i \neq j} d_{ij}(t_k), \quad (4.9)$$

where  $N_p^{intra}$  is the number of pairs of microbes from the same taxa. The next step is to measure the mean inter-taxa distance. To that end, we take all possible pairs of membership vectors from different taxa  $(\eta_k^i, \eta_{k'}^j)$  such that  $k \neq k'$  and measure the euclidean distances. We repeat this process for all pairs of different taxa and calculate the mean:

$$\langle d_{inter} \rangle = \frac{1}{N_p^{inter}} \sum_{k \neq k'}^n \sum_{i, j} d_{ij}(t_k, t_{k'}). \quad (4.10)$$

Finally, we take the log of the ratio intra-inter distances  $\log(\langle d_{intra} \rangle / \langle d_{inter} \rangle)$  for all five taxonomical levels (see Fig. 4.5).

We observe that the average intra-category distance is smaller than the average inter-category distance in all cases, except for the Phylum, Class, and Order of the dataset V-10 (the smallest one). That is, on average, microbes of the same taxa are closer than those from different clades. This means that our model partially captures the taxonomic similarities of the microbes. Also, we see that the ratio becomes smaller for the lowest taxonomic levels: family and genus. This is probably due to the fact that the lower the taxonomic level, the more similar the microbes in the same taxa are.

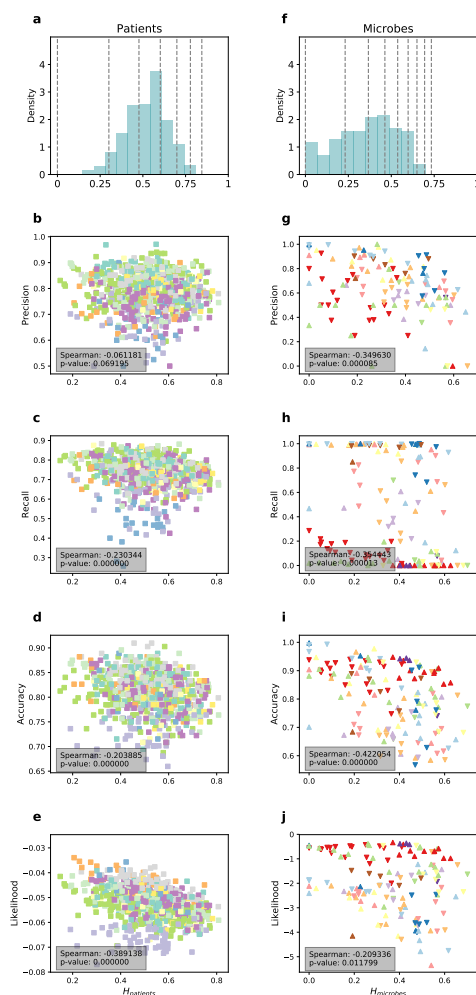


Figure 4.4: **Individual predictability.** [a,f] Distribution of Shannon entropies of memberships. Gray dashed lines represent numerical values of the Shannon Entropy such that the membership is equally spread among  $n$  groups. [b,g] Precision of individual patients and microbial species vs Shannon Entropies. Each colored square/triangle represents an individual patient/microbe. Color and shape (if applies) represent its top membership. In both cases there is a negative correlation between individual precisions and Shannon Entropies, although it is only statistically significant and larger for microbes. [c,h] Recall for individual patients and microbial species vs Shannon Entropies. Again, there is a negative correlation between individual recalls and Shannon Entropies, that only shows statistical significance for the microbes. [d,i] Individual accuracies for patients and microbes vs Shannon Entropies. Negative correlations and p-values are similar in both cases. [e,j] Individual held-out-loglikelihoods vs Shannon Entropies. The correlation is only significant for the patients.

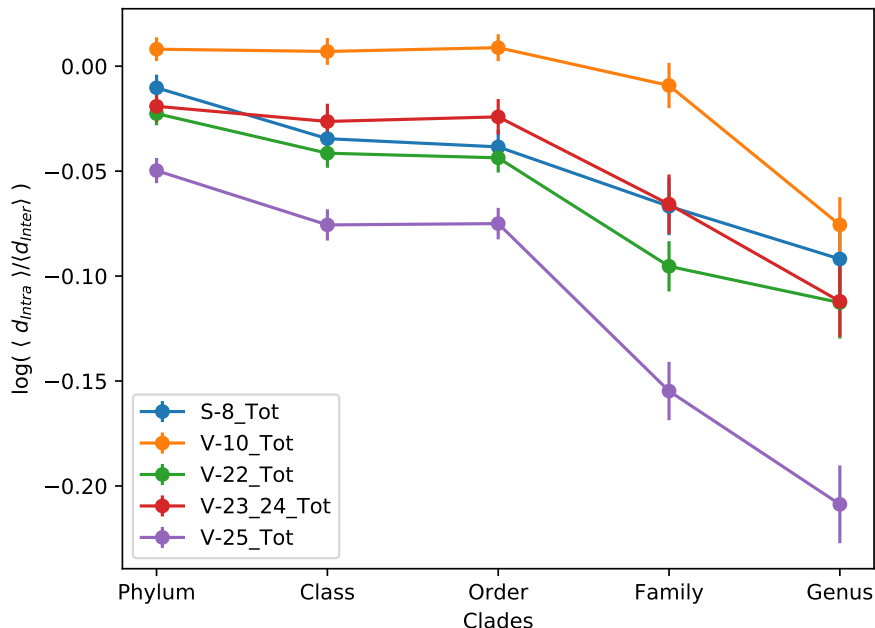


Figure 4.5: **Log-ratio of the average intra/inter clade distances of microbes** For each dataset, we take the 20-dimensional membership vectors of all microbes and group them in the different existent taxonomic groups (phyla, classes, orders, families, and genera). Then, we measure the euclidean distances of all possible pairs of intraclade and interclade vectors. Finally, we compute the main intraclade and interclade distance and take the log-ratio. We observe a generalized trend of decreasing log-ratios, especially at the Family and Genus level. Also, we note that most log-ratios are negative, indicating a smaller intra-clade than inter-clade distance. This suggests that our model successfully captures the taxonomic proximity of different microbes.

#### 4.4.6 Nestedness in latent abundance patterns

Nestedness is a well known property of some ecological systems, particularly so called mutualistic networks [68], [50], [69], which are characterized by mutually beneficial trophic or reproductive relations. A common example are insect-plant networks, where both insects and plants benefit from their mutual interactions: insects obtain their food, while plants are pollinized by insects or have their seeds dispersed around. In many cases, there is a nested pattern in these systems: some insects (usually called specialists) feed on a very specific subset of plants, while others (generalists) feed on a majority of plants in the ecosystem. The same is also true for plants, with some being specialized in a few species of insects an

others being prevalent among all insects. Additionally, the insects that feed from a specialist plant are a subset of those insects that feed on generalist plants (see Fig. 4.6). Nestedness can actually be quantified for real mutualistic networks. Usually, it ranges from  $n = 0$  to  $n = 1$ , with larger values indicating more nested networks.

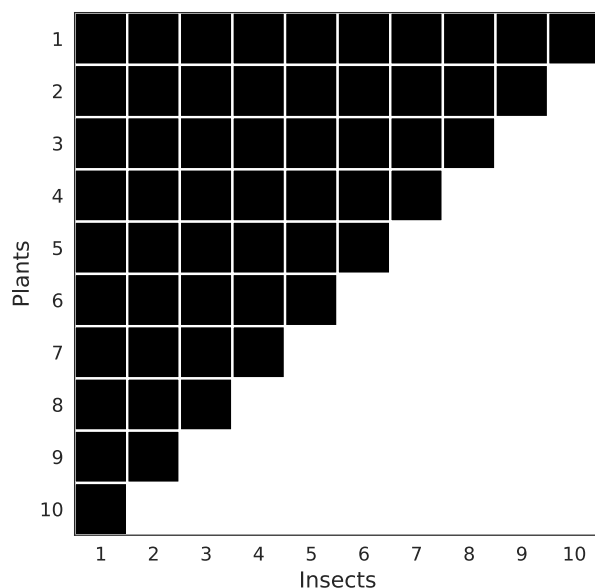


Figure 4.6: **Perfectly nested mutualistic network** Specialist animals or insects (rightmost columns) feed on plants that are a subset of the plants from which generalist animals or insects feed (leftmost columns). (Figure based in [50])

In healthy individuals, it also occurs that both the host and its gut microbial system benefit from mutual interactions. For example, dietary fibers are indigestible for humans, but the gut microbiome is involved in generating digestible byproducts, while at the same time obtaining energy for the microbes involved in this task [70]. We therefore explore the possibility of a nested ecological order in these systems. In particular, we analyse it at the level of latent enterotypes. As mentioned before, latent enterotypes are microbial profiles or abundance patterns of a set of groups of microbes. For simplicity, we limit ourselves to consider whether a group of microbes exists in an enterotype or not. That is, we consider again only negligible (0) or non-negligible (1) abundances. Also, we approximate the probability  $\Pr[a_{pm} = 1]$  that a group of microbes  $m$  is present in a group of patients  $p$  to 0 or 1: if this probability is larger than 0.5, we assign it the value 1 and a 0 otherwise. This choice is justified because most of the probabilities are indeed equal to 0 or 1 or very



close to these values. Because our optimal choice of parameters (i.e., the one leading to the best predictive power) was  $K = 10$  groups of patients and  $L = 20$  groups of microbes, we have 10 latent enterotypes containing 20 groups of microbes each.

In order to measure the nestedness of the latent enterotypes, we use the metric presented in [71] and given by:

$$n = \frac{\sum_{i < j} q_{ij}^{(p)} + \sum_{i < j} q_{ij}^{(m)}}{\sum_{i < j} \min(q_i^{(p)}, q_j^{(p)}) + \sum_{i < j} \min(q_i^{(m)}, q_j^{(m)})}, \quad (4.11)$$

with  $q_{ij}$  representing the number of shared interactions between latent groups  $i$  and  $j$  and  $q_i$  being the number of interactions of group  $i$ .  $m$  and  $p$  represent latent groups of microbes and patients, respectively. As mentioned above, the numerical value of  $n$  ranges from 0 to 1.

Fig. 4.7 shows the enterotypes for all five datasets. In all plots we observe a consistent pattern of nestedness across the enterotypes when we order rows and columns according to the number of non-negligible matrix values: on the bottom rows, there are specialist enterotypes that contain just a small fraction of groups of microbes, whereas enterotypes in the top rows are generalists in that they contain a majority of groups of microbes. On top of that, the groups of microbes contained in a specialist enterotype are a subset of the groups of microbes contained in generalist enterotypes, confirming our initial expectation. The nestedness can be observed in all five datasets with high statistical significance.

It is important to note that nestedness here has a slightly different meaning than that described in insect-plant mutualistic networks, because patients are isolated ecosystems not interacting with each other. Here, the nestedness implies that some groups of patients are specialists in that they interact with a reduced number of groups of microbes, and others are generalists because they interact with a majority of groups of microbes. From the point of view of the groups of microbes, it implies that some microbial species are prevalent in almost all patients while others only exist in a small fraction of them.

One may object that this nested pattern only happens in the latent space and it is not representative of the actual system. However, making the same two approximations as before, it is possible to observe a nested pattern in the actual OTU matrices as well (see Fig 4.8). This makes sense, since patients and microbes belong to all groups with different weights, and are therefore linear combinations of all latent groups. Unsurprisingly, the nestedness signal is weaker in the actual OTU matrices, because it is precisely the latent enterotypes that capture the collective patterns and regularities of the system.

The fact that there is a well ordered ecological structure in human microbiome systems is something remarkable. It suggests the possibility that health is not necessarily related to microbial diversity.

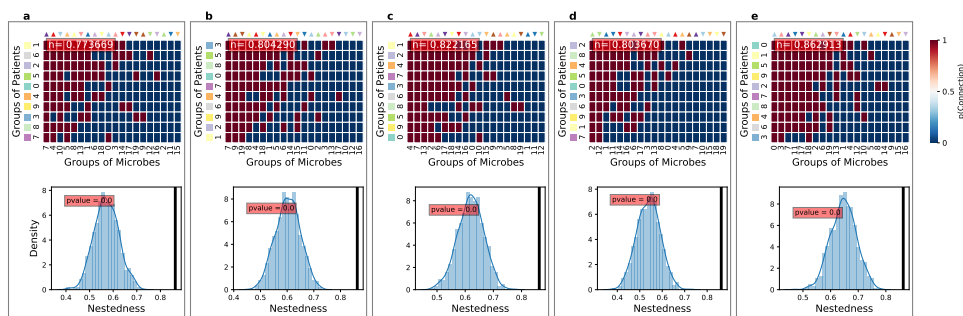


Figure 4.7: **Nestedness of the Latent Enterotypes [a-e]** Nestedness of the latent enterotypes and significance plots for all five datasets considered. In the heatmaps (top), rows represent enterotypes, and columns correspond to latent groups of microbes in a similar fashion as the example shown in Fig. 4.1. Red/blue colored squares represent the presence/absence of a group of microbes in an enterotype. The colored squares and triangles label enterotypes and groups of microbes. The color and shape (if it applies) code is the same as in Fig. 4.4. Red and blue squares in the heatmap indicate that the corresponding latent group of microbes is not connected to the latent enterotype. Rows and columns have been ordered to highlight the nestedness of the system. The numerical value of the nestedness is shown in the inset of the plots. The bottom plot shows the distribution of nestedness values of 1000 randomizations of the latent enterotypes. We randomize the latent enterotypes by preserving the average number of connections between latent groups of patients and microbes. The black solid line corresponds to the observed value of the nestedness.

## 4.5 Discussion

Our work proposes an approach in which enterotypes are not necessarily rigid microbial profiles. Subsequently, individuals' microbiomes are combinations of different enterotypes in variable degrees. We believe that this flexibility is more consistent with the variability of biological data, while still capturing universal patterns, hence increasing the predictive power. On top of that, it enables to uncover features such as the nestedness, suggesting an analogy with mutualistic systems. To our knowledge, this is the first time nestedness is reported in human microbial systems. The fact that there is an ecological order in the human gut microbiome across individuals raises some questions about their collective behavior: for instance, we wonder whether nestedness is a property of healthy populations and whether it could be affected by certain diseases or health disorders.

We validate our results showing that our latent enterotype model can make successful predictions of unobserved microbial abundances both on a global scale and on individual patients. Additionally, we observe that our model can partially capture the taxonomic proximity of microbial species. We believe that this could have potential clinical applications in individ-

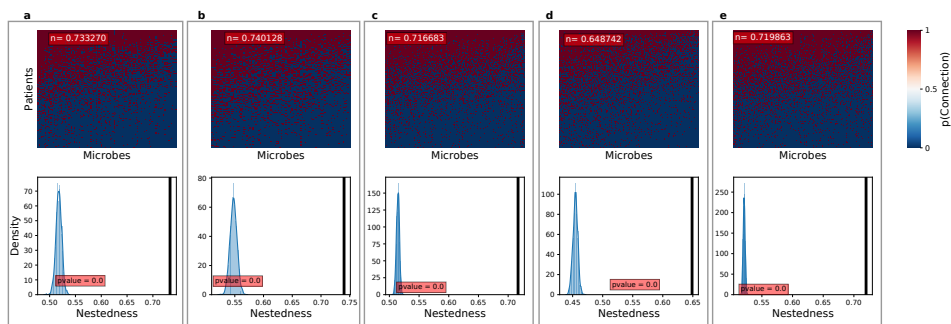


Figure 4.8: **Nestedness of the OTU matrices [a-e]** Nestedness of the OTU matrices and significance plots for all five datasets. In the upper plots, rows represent patients, and columns represent microbial species. Red and blue squares in the heatmap mean that the corresponding groups of patients and microbes are or are not connected. Rows and columns have been ordered to highlight the nestedness of the system. The numerical value of the nestedness is shown in the inset of the plots. The bottom plot shows the nestedness of 1000 randomizations of the latent enterotypes with the black solid line representing the nestedness value of the original OTU matrix. Randomizations are made preserving the average number of connections between patients and microbes.

ualized medicine. We leave for future work the possibility of making predictions in time evolving systems such as transitions to a healthy/diseased state. Another interesting question is if we could increase the predictive power of our model extending our approach to other levels of systems biology such as metabolomics, epigenetics etc. Indeed, this could improve our understanding of the human gut microbiome.

UNIVERSITAT ROVIRA I VIRGILI

STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

---

## Chapter 5

# Conclusions and perspectives

The main goal of this thesis was to build statistical inference models for interpretable link prediction in multilink bipartite networks. The interest in bipartite networks relied on their ubiquity in many natural and social systems. Likewise, we were interested in a multilink approach because we wanted to consider different types of interactions in order to make more sophisticated analyses. Formally, this required an additional level of complexity, but we believed it is well justified due to the potential to study many interesting and relevant problems.

Link prediction is a very important problem in that it has many obvious practical applications, ranging from information retrieval to prediction of future events in different contexts. But more importantly, we believe that the ability of making predictions about the behavior of a system is a natural byproduct of the understanding of its underlying dynamics, which is ultimately the goal of science and research.

In order to fulfill this goal, we started this thesis presenting two models of the family of Stochastic Block Models (SBM). SBM are generative models that allow us to infer unknown properties and features of a system given data corresponding to that system. The choice of SBM was motivated by three important features: their formal simplicity, their tractability and their expressiveness. As it turns out, interpretable link prediction arises very naturally from these three aspects. The formal simplicity of Stochastic Block Models can be partly explained because the only assumption made is that there exist groups of nodes in networks. Given a group structure in a network, it is possible to infer the existence of a missing link between two nodes based exclusively on the groups to which those nodes belong. This greatly simplifies the problem of link prediction, because it reduces it to the number of groups identified, hence the tractability of SBM. Finally, it is very easy to analyze the interaction of groups or communities, which allows for expressive explanations of the dynamics of the group structure (and the system, implicitly) and a straightforward interpretation of the predictions.

The first model presented is a conventional approach that classifies nodes in different groups or communities according to their similarities in terms of the roles displayed in the network and their connectivity patterns. The second model is a mixed-membership approach

---

in which nodes can belong to several communities simultaneously. Therefore, communities become latent and they do not generally correspond to groups of nodes any longer. Although the idea of latent communities adds an important degree of abstraction to the model, it turns out that predictions are more interpretable in the mixed approach than in the conventional one.

After discussing the Stochastic Block Models, we applied these models to two different problems. In the first problem we considered a social experiment in which a large sample of people were playing dyadic games iteratively. The experiment consisted of 541 participants or players and 121 games. Each player was confronted with an average of 14 rounds of randomly selected games against randomly selected players. Based on the records of all players and implementing our models, we showed that people can be grouped in behavioural phenotypes according to the similarities in their decision-making strategies. Similarly, we found classifications of games according to the perception of people, which turned out to be different from game-theoretical criteria. Additionally, we incorporated existing information (metadata) about the games into the formalism. This resulted in a more explanatory group structure and in an increase of the predicting performance. Upon these classifications, we could predict as much as 71% of people decisions in the conventional approach and 74% in the mixed-membership approach. Our results thus suggest that people are very consistent when confronted to simple decisions, although further experiments might be helpful to provide more evidence. On the other hand, these results arose from a rigorous comparison between three models: a baseline model, the conventional and the mixed-strategy approach: in all three cases, we treated the data in the same way and concluded that the best model was the best predicting one. We suggest that this could pave the way for the design of model comparison protocols in future works.

Our second problem was a human microbiology one. Here, we had data about the gut microbiome of a large number of patients. In particular, we had five studies each on a different cohort of healthy people. In all cases, we had the microbial profiles of each patient, consisting on the concentrations of the microbial species detected, together with their taxonomic information. This system can also be regarded as a bipartite network in which patients and microbes are represented the two kinds of nodes and links correspond to different intervals of microbial concentrations or abundances. In this case, we directly implemented the mixed-membership approach, given our evidence of its higher predictive strength and the large size of the datasets considered here (usually it is computationally cheaper to implement the mixed-strategy model). In a similar fashion as in the first problem, we found latent groups of patients and latent groups of microbes with which we built an inference model that can predict whether a microbe is present in a given patient with 80% accuracy. On top of that, we found that there is a consistent nestedness pattern among latent groups of patients and microbes. This pattern can also be found to a lesser extent in the original datasets.

Overall, we have shown that a single family of statistical inference models allows us to make predictions of unobserved events in two very different problems. Moreover, we could actually learn general aspects of their dynamics despite the different nature of both systems. Perhaps counterintuitively, the fact that both problems are so different from each other is not a drawback, but it actually underpins the versatility and robustness of our models. We believe that these models have a huge unexplored potential to solve many other problems in

---

which interpretable prediction can be of fundamental importance. Additionally, their ability to provide information about the internal dynamics of the systems under study could unveil many important details of their nature and advance theories that explain them. We would like to explore some of these problems and also see how our models would adapt to different types of networks or sources of prior information. For instance, if we had had information about patients (gender, age, diet) in our second problem, how could we have introduced it as a prior distribution in the model?

However, several questions remain yet unanswered about the problems studied here. In the case of the social system, subjects in the experiment were playing an average of 14 rounds, as mentioned before. This is certainly not a very large number. We wonder if our predictions would have been more accurate with larger records of actions or they would have remained stable because the strategies of players were mostly fixed at that number of rounds. Besides, we have considered the simplest possible scenario to model human interactions and decisions, with games involving only two people with two available choices each. Despite the large number of variations for this scenario accounted for (121 games), decisions faced by human beings in their daily life are often far more complicated, with many possibilities available and more than two people involved. We wonder if it would be possible to come up with formal approaches that would allow us to come closer to real life decision making scenarios and how predictions would be affected.

In regards to our human microbiology problem, even more questions remain open; for a start, we wonder whether nestedness could be sensitive to the health state of the population considered. Would nestedness change or disappear if we considered cohorts of diseased patients? This question would be especially relevant in autoimmune or microbiome-related diseases, such as inflammatory bowel diseases. We also wonder if we could study the temporal evolution of the microbiome within our approach. Taking snapshots of the microbiome of patients at different times, we could establish connections between microbiome and dietary habits, age or potentially related diseases. From a clinical point of view, we could try to predict the success of different therapies such as antibiotics or fecal microbial transplant (FMT). In the same spirit, we wonder if we could add more elements to the microbiome picture; systems biology is a multiscale problem in which microbiome is just one layer deeply connected to several others, such as genetics, proteomics, metabolomics or even the environmental conditions to which we are exposed. By bringing in some of these layers to our analysis, we could improve our understanding of microbiome dynamics and human biology. Needless to say, this might also shed some light on the nature of the diseases mentioned before. We could probably model this idea very intuitively using multilayer networks.

Finally, we observe that despite the high accuracy of our predictions in both problems, there is a 5% difference between them. This significant disparity poses some questions. First of all, are some systems more predictable than others? Surely data plays a role: some datasets can be noisier than others, for instance. And generally, a larger amount of data will yield better predictions because the signal is better represented. But, even in large datasets, is there a way we could tell the signal from the noise? Actually, is there a limit to predictability that we could identify? On the other hand, is the predictive power of our models limited? If so, to which extend and how could we improve them? Moreover, do our models constrain the way in which we see the data? For example, we have said before that the  $p$  matrices of the

---

MMSBM tend to have very informative values (i.e., close to 0 or 1), but we don't know with certainty why this happens. It is also left for future work the task of better understanding the inner mechanics of our models.

---



# Appendices

## 5.1 A. Stochastic Block Models

### 5.1.1 A1: Multivariate beta function integrals

In this section we show the solution to the Eq 2.4 . Let us start decoupling the integrals for all  $\mathbf{p}$ :

$$P(\boldsymbol{\theta}, \boldsymbol{\eta} | A^o) = \frac{1}{Z} \prod_{k \in K} \prod_{\ell \in L} \int_0^1 dp_1(k, \ell) p_1(k, \ell)^{n_{k\ell}^1} \int_0^{1-p_1} dp_2(k, \ell) p_2(k, \ell)^{n_{k\ell}^2} \dots \int_0^{1-p_1-\dots-p_{\Lambda-2}} dp_{\Lambda-1}(k, \ell) p_{\Lambda-1}(k, \ell)^{n_{k\ell}^{\Lambda-1}} (1 - p_1(k, \ell) - \dots - p_{\Lambda-1}(k, \ell))^{n_{k\ell}^{\Lambda}}. \quad (5.1)$$

Note that we impose the constraint that all probabilities have to sum up to 1, i.e.:

$$\sum_{\lambda=1}^{\Lambda} p_{\lambda}(k, \ell) = 1, \quad (5.2)$$

hence the upper limits of the integrals. Now, let us start solving the innermost integral. For the sake of simplicity, we'll reduce the notation by taking  $p_{\lambda} = p_{\lambda}(k, \ell)$  and  $n^{\lambda} = n_{k\ell}^{\lambda}$ :

$$\int_0^{1-p_1-\dots-p_{\Lambda-2}} dp_{\Lambda-1} p_{\Lambda-1}^{n_{\Lambda-1}^{\Lambda-1}} (1 - p_1 - \dots - p_{\Lambda-1})^{n^{\Lambda}}. \quad (5.3)$$

The first step is to take  $v = 1 - p_1 - \dots - p_{\Lambda-2}$ , which gives:

$$\int_0^v dp_{\Lambda-1} p_{\Lambda-1}^{n_{\Lambda-1}^{\Lambda-1}} (v - p_{\Lambda-1})^{n^{\Lambda}}. \quad (5.4)$$

Now we perform the change of variable  $p_{\Lambda-1} = vt$ , which gives  $dp_{\Lambda-1} = vdt$  and changes the integral limits to 1 and 0:

$$\int_0^1 dt v^{n^{\Lambda-1} + n^{\Lambda} + 1} t^{n^{\Lambda-1}} (1 - t)^{n^{\Lambda}} = v^{n^{\Lambda} + n^{\Lambda-1} + 1} \int_0^1 dt t^{n^{\Lambda-1}} (1 - t)^{n^{\Lambda}}. \quad (5.5)$$

The right-hand side integral is a beta function, whose solution is known:

$$v^{n^{\Lambda-1}+n^{\Lambda}+1} B(n^{\Lambda-1} + 1, n^{\Lambda} + 1) = (1 - p_1 - \dots - p_{\Lambda-2})^{n^{\Lambda}+n^{\Lambda-1}+1} \frac{n^{\Lambda}!n^{\Lambda-1}!}{(n^{\Lambda} + n^{\Lambda-1} + 1)!}. \quad (5.6)$$

Inserting this in Eq. 5.1, we have:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta}|A^o) = \frac{1}{\mathcal{Z}} \prod_{k \in K} \prod_{\ell \in L} \int_0^1 dp_1(k, \ell) p_1(k, \ell)^{n_{k\ell}^1} \int_0^{1-p_1} dp_2(k, \ell) p_2(k, \ell)^{n_{k\ell}^2} \dots \int_0^{1-p_1-\dots-p_{\Lambda-3}} dp_{\Lambda-2}(k, \ell) p_{\Lambda-2}(k, \ell)^{n_{k\ell}^{\Lambda-2}} (1 - p_1(k, \ell) - \dots - p_{\Lambda-2}(k, \ell))^{n_{k\ell}^{\Lambda}+n_{k\ell}^{\Lambda-1}+1} \frac{n_{k\ell}^{\Lambda}!n_{k\ell}^{\Lambda-1}!}{(n_{k\ell}^{\Lambda} + n_{k\ell}^{\Lambda-1} + 1)!}. \quad (5.7)$$

We can now solve the next innermost integral using the same procedure as before:

$$\int_0^{1-p_1-\dots-p_{\Lambda-3}} dp_{\Lambda-2}(k, \ell) p_{\Lambda-2}(k, \ell)^{n_{k\ell}^{\Lambda-2}} (1 - p_1(k, \ell) - \dots - p_{\Lambda-2}(k, \ell))^{n_{k\ell}^{\Lambda}+n_{k\ell}^{\Lambda-1}+1} = (1 - p_1(k, \ell) - \dots - p_{\Lambda-3}(k, \ell))^{n_{k\ell}^{\Lambda}+n_{k\ell}^{\Lambda-1}+n_{k\ell}^{\Lambda-2}+2} \frac{n_{k\ell}^{\Lambda-2}!(n_{k\ell}^{\Lambda} + n_{k\ell}^{\Lambda-1} + 1)!}{(n_{k\ell}^{\Lambda} + n_{k\ell}^{\Lambda-1} + n_{k\ell}^{\Lambda-2} + 2)!}. \quad (5.8)$$

Introducing this again in Eq. 5.3, we get:

$$P(\boldsymbol{\theta}, \boldsymbol{\eta}|A^o) = \frac{1}{\mathcal{Z}} \prod_{k \in K} \prod_{\ell \in L} \int_0^1 dp_1(k, \ell) p_1(k, \ell)^{n_{k\ell}^1} \int_0^{1-p_1} dp_2(k, \ell) p_2(k, \ell)^{n_{k\ell}^2} \dots \int_0^{1-p_1-\dots-p_{\Lambda-4}} dp_{\Lambda-3}(k, \ell) p_{\Lambda-3}(k, \ell)^{n_{k\ell}^{\Lambda-3}} (1 - p_1(k, \ell) - \dots - p_{\Lambda-3}(k, \ell))^{n_{k\ell}^{\Lambda}+n_{k\ell}^{\Lambda-1}+n_{k\ell}^{\Lambda-2}+2} \frac{n_{k\ell}^{\Lambda-2}!(n_{k\ell}^{\Lambda} + n_{k\ell}^{\Lambda-1} + 1)!}{(n_{k\ell}^{\Lambda} + n_{k\ell}^{\Lambda-1} + n_{k\ell}^{\Lambda-2} + 2)!} \frac{n_{k\ell}^{\Lambda}!n_{k\ell}^{\Lambda-1}!}{(n_{k\ell}^{\Lambda} + n_{k\ell}^{\Lambda-1} + 1)!}. \quad (5.9)$$

Doing the remaining integrals and repeating the same protocol in each of them, we get the result in Eq. 2.6.

### 5.1.2 A3: Jensen's Inequality

In our particular case, the complete form of Jensen's inequality is:

$$\log \sum_x p(x)x \geq \sum_x p(x)\log(x) \quad (5.10)$$

replacing  $x$  and  $p(x)$ , by the auxiliary distribution  $\omega_{ui}(k, \ell)$  and the quotient  $\frac{\theta_{uk}\eta_{i\ell}p_{k\ell}(\tau_{ui})}{\omega_{ui}(k, \ell)}$  respectively, we recover Eq. 2.16.

### 5.1.3 A2: Simulated annealing

As discussed in the text, we use simulated annealing to find the group assignments  $\theta$  and  $\eta$  that maximize the expression for the posterior in Eq. 2.8 or alternatively minimize the energy  $\mathcal{H}_{min}$  in 2.7 of Chapter 2. However, solving this problem analytically is computationally infeasible, given the vast number of possible partitions (group assignments) of the system. Therefore, we implement a simulated annealing algorithm to perform this task [72]. The idea of this method is the following: starting from a given partition of players and games whose energy  $\mathcal{H}_0$  is known, new partitions are proposed by randomly moving players and games to different groups and energies are computed. The new partitions are automatically accepted if  $\mathcal{H}_{new} < \mathcal{H}_0$ . Otherwise, they are accepted with probability  $e^{-\Delta\mathcal{H}/T}$ , where  $T$  represents the temperature. In this case, the temperature basically controls the tolerance of the system to switching towards partitions with higher energies. The key point of the simulated annealing is that the temperature gradually decreases with the number of iterations. That way, the system can initially explore the whole landscape and escape from a local minima, for instance. For each value of the temperature, we allow  $N_{players}^2 + N_{games}^2$  movements of players and games. Then, we cool the system by a factor  $\lambda = 0.99$ . That is,  $T_{new} = \lambda T_{old}$ . Finally, if the system doesn't change its energy after 10 temperature changes, the algorithm automatically stops its execution.

## 5.2 B. An Application to Social Systems

### 5.2.1 Initial rounds in the empirical data

We observe that the behavior of each player during the first four rounds is erratic, which leads to their behavior being less predictable during those rounds (Fig. 5.1). After round 4, all rounds are statistically indistinguishable by the metrics discussed in the main text. Therefore, we discard the first four rounds of each player and consider all others as indistinguishable.

### 5.2.2 Number of groups in the single-strategy and multiple-strategy models

In the single-strategy model, the number of groups is determined automatically by the simulated annealing optimization. Since group plausibilities are calculated by marginalizing exactly over the  $\mathbf{p}$  matrices (Eq. 1 above), Eq. 5 above already penalizes complex models, and the optimization will naturally choose the optimal number of groups. The optimal model consists of around 20 groups (depending on the cross-validation fold), although 5 or 6 of them alone typically account for more than 50% of the players.

For the multiple-strategy model, the number of groups needs to be fixed manually. As shown in Fig. B2, we find that the optimal predictions are obtained for  $K=3$  groups of players and  $L=4$  groups of games, although performance is not very sensitive to these values. In fact, for larger values of  $K$  and  $L$ , the performance is similar but some groups are, in practice, left empty.

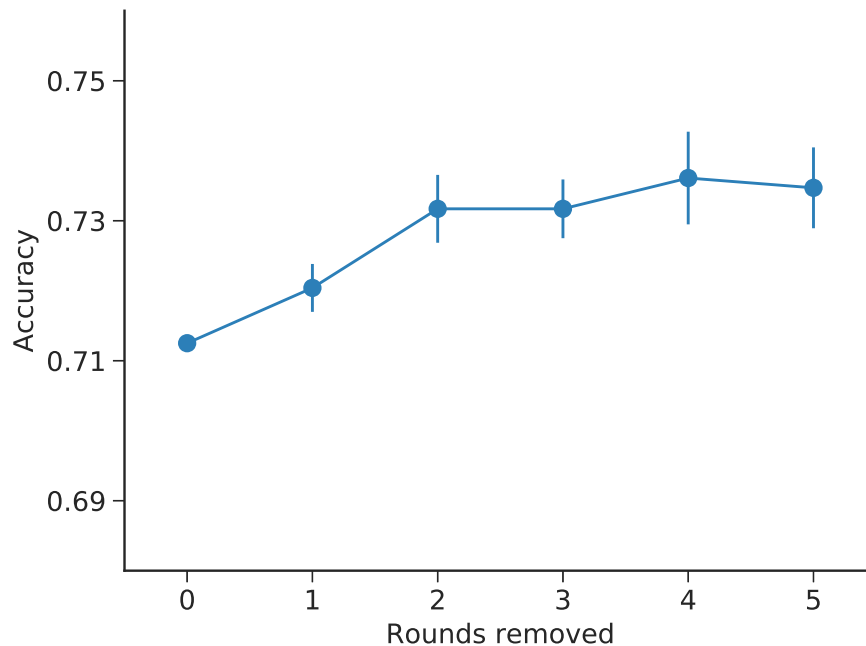


Figure 5.1: **Predictive accuracy of the single-strategy model after removing the first rounds of the players' histories.** The error bars show the standard error of the mean of the results for the 5 folds.

### 5.2.3 Robustness of the results

In Figs. 3.3 and 3.4 of the main text, we show results for a single fold of the 5-fold cross-validation (except panel **a** in each of them, that shows the average over the five folds). Figures 5.3 and 5.4 show equivalent results for a different fold, and are very similar to those in the main text, thus indicating that they are robust.

## 5.3 C. An Application to human microbiology

### 5.3.1 Robustness of the results

We show the correlations between  $H_p$  and  $H_m$  and predictive metrics for the four remaining datasets not shown in the main text: S-8, V-10, V-22, and V-23-24.

---

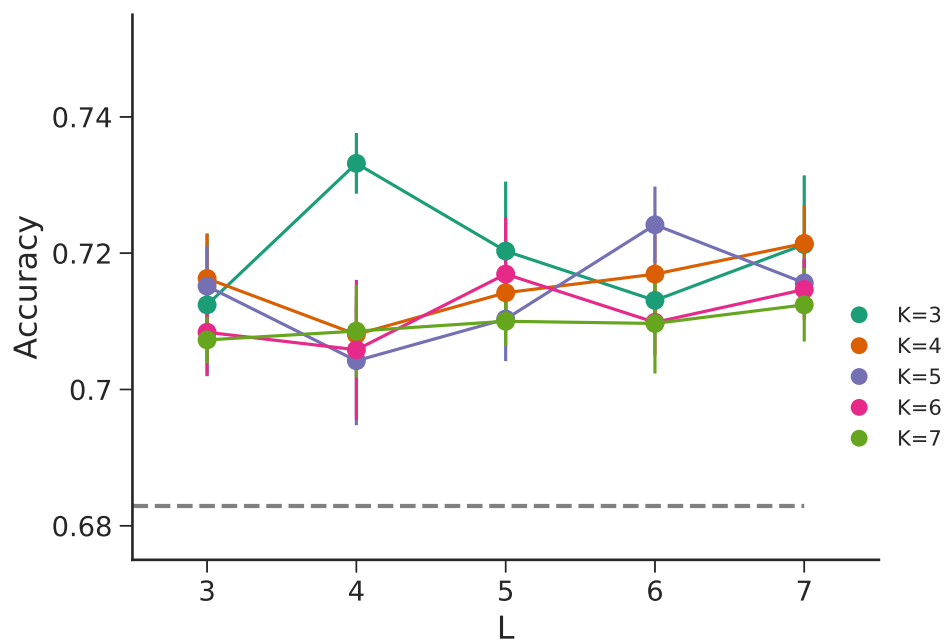


Figure 5.2: **Dependence of the predictive accuracy of the multiple-strategy model for different values of  $K$  and  $L$ .** We show the predictive accuracy for different combinations of  $K$  and  $L$  (the number of latent groups of players and games) for  $\alpha = 0$ . Each point represents the average of a 5-fold cross-validation; error bars indicate the standard error of the mean. Note that for any choice of parameter values the accuracy of the multiple-strategy model is above the accuracy of the single-strategy model.

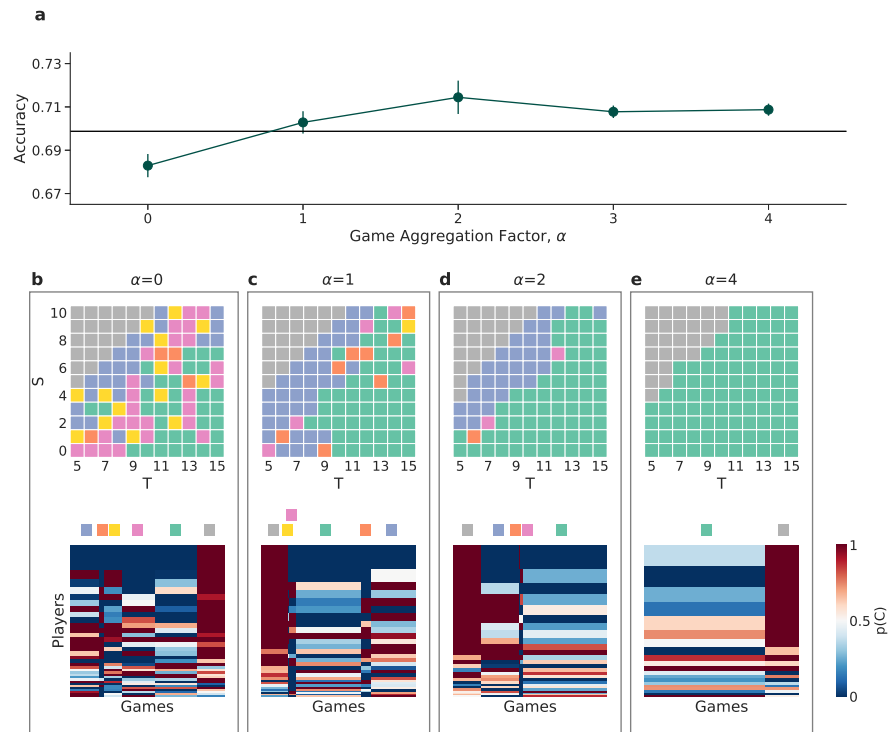


Figure 5.3: Same as Fig. 3.3 in the main text for a different split of the 5-fold cross validation. Note how the results we obtain for partitions of games and users are very similar to the ones shown in Fig. 2 of the main text, thus suggesting that the results are robust.

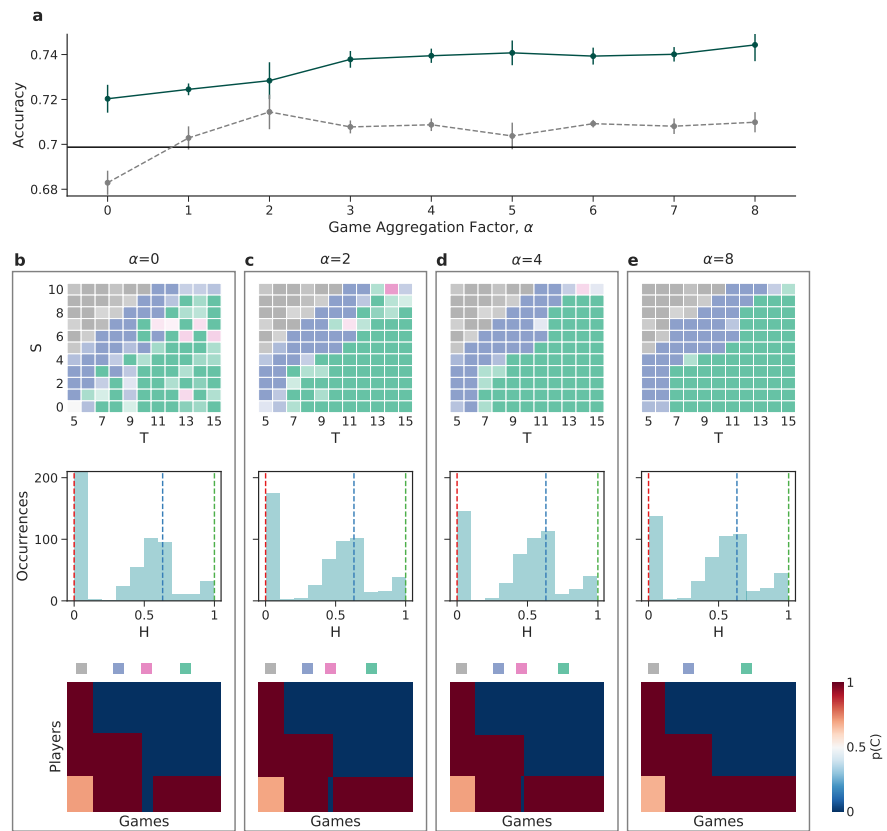


Figure 5.4: Same as Fig. 3.5 in the main text for a different split of the 5-fold cross validation (same split as in Fig. 5.3). Note how rows b), c) and d) are very similar to those shown in 3.5 of the main text, thus suggesting that the results are robust.

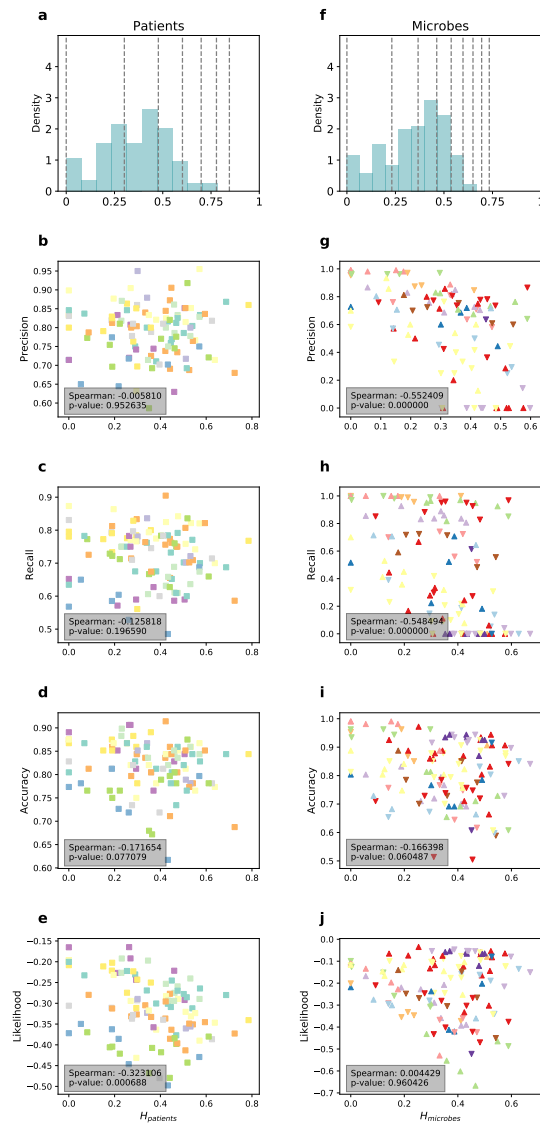


Figure 5.5: Same as Fig. 4.4 in the main text for data set S-8.



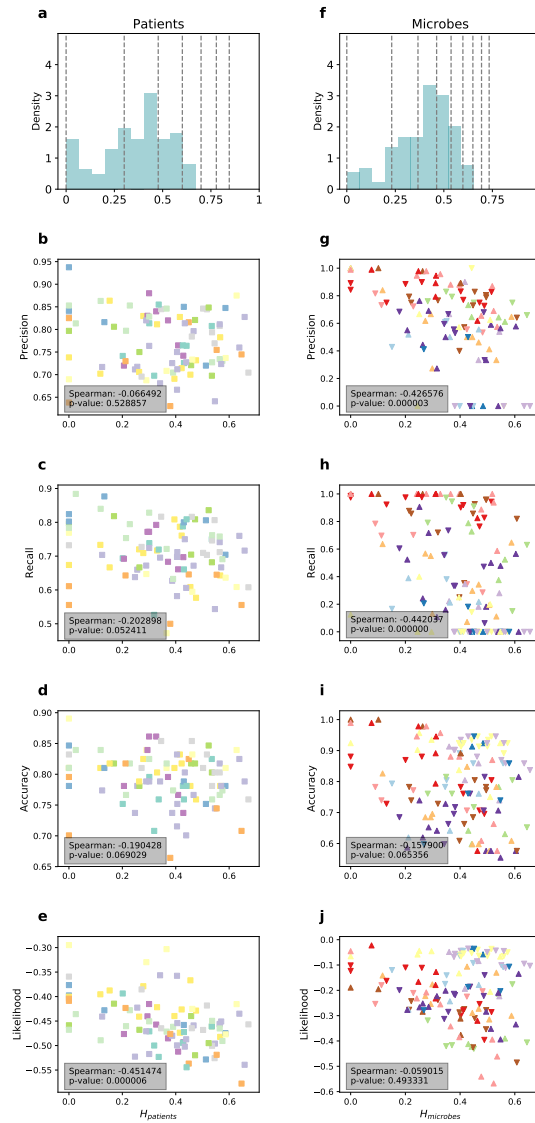


Figure 5.6: Same as Fig. 4.4 in the main text for data set V-10.

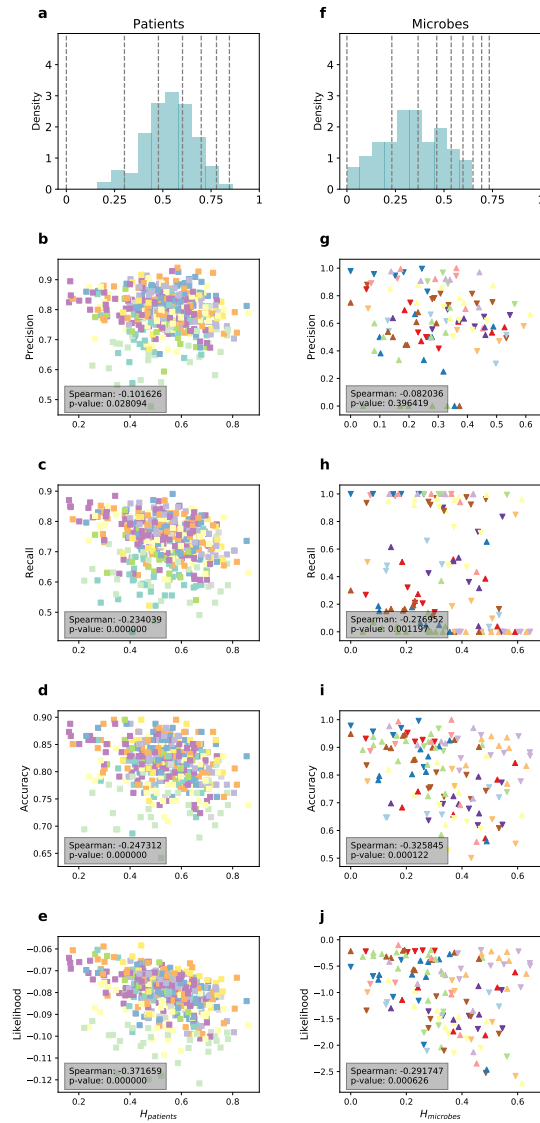


Figure 5.7: Same as Fig. 4.4 in the main text for data set V-22.

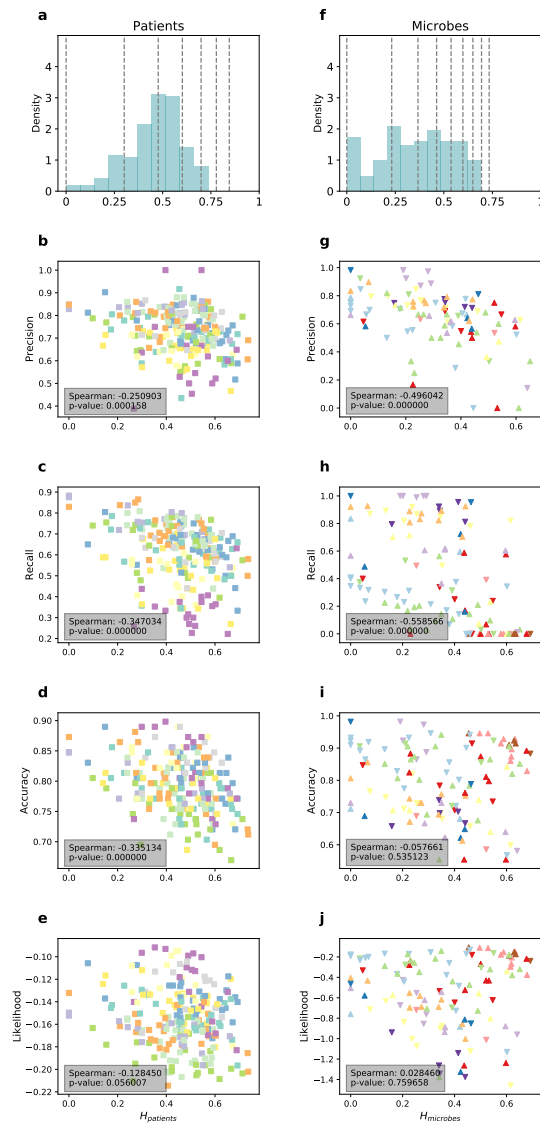


Figure 5.8: Same as Fig. 4.4 in the main text for data set V-23-24.

UNIVERSITAT ROVIRA I VIRGILI

STATISTICAL INFERENCE IN BIPARTITE NETWORKS APPLIED TO SOCIAL DILEMMAS AND HUMAN MICROBIAL SYSTEMS

Sergio Cobo López

---

# Bibliography

- [1] Nora Barlow. *The autobiography of Charles Darwin 1809-1882*. Collins, London, 1958.
  - [2] Francisco J. Ayala. Darwin and the scientific method. *Proceedings of the National Academy of Sciences*, 106(Supplement 1):10033–10039, 2009.
  - [3] C Darwin. *On the Origin of Species by Means of Natural Selection, Or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859.
  - [4] Mark Hindell and Roger Kirkwood. *Marine mammals: fisheries, tourism and management issues*. Csiro Publishing, 2003.
  - [5] Guido Caldarelli and Michele Catanzaro. *Networks: A Very Short Introduction*. Oxford University Press, Oxford, 2012.
  - [6] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.
  - [7] R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. U. S. A.*, 106(52):22073–22078, 2009.
  - [8] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, et al.
  - [9] Barbara A. Methé, Karen E. Nelson, Mihai Pop, Heather H. Creasy, Michelle G. Giglio, et al. A framework for human microbiome research. *Nature*, 486(7402):215–221, 2012.
  - [10] C. Harger, G. Chen, A. Farmer, W. Huang, J. Inman, et al. The Genome Sequence DataBase. *Nucleic Acids Research*, 28(1):31–32, 01 2000.
  - [11] A Godoy-Lorite, R Guimerà, and M Sales-Pardo. Long-term evolution of email networks: Statistical regularities, predictability and stability of social behaviors. *PLoS ONE*, 11(1):e0146113, 2016.
-

- 
- [12] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [13] R. Guimerà and M. Sales-Pardo. A network inference method for large-scale unsupervised identification of novel drug-drug interactions. *PLoS Comput. Biol.*, 9(12):e1003374, January 2013.
- [14] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.
- [15] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44(1.2):206–226, Jan 2000.
- [16] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Soc. Networks*, 5:109–137, 1983.
- [17] T Vallès-Català, T.P. Peixoto, R. Guimerà, and M. Sales-Pardo. On the consistency between model selection and link prediction in networks. arXiv:1705.07967 [stat.ML].
- [18] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9(2008):1981–2014, 2008.
- [19] Antonia Godoy-Lorite, Roger Guimerà, Cristopher Moore, and Marta Sales-Pardo. Accurate and scalable social recommendation using mixed-membership stochastic block models. *Proc. Natl. Acad. Sci. U.S.A.*, 113:14207 — 14212, dec 2016.
- [20] Sergio Cobo-López, Antonia Godoy-Lorite, Jordi Duch, Marta Sales-Pardo, and Roger Guimerà. Optimal prediction of decisions and model selection in social dilemmas using block models. *EPJ Data Science*, 7(1):48, 2018.
- [21] R. Dean Malmgren, Daniel B. Stouffer, Andriana S. L. O. Campanharo, and Luis A. N. Amaral. On universality in human correspondence activity. *Science*, 325(5948):1696–1700, September 2009.
- [22] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [23] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Syst.*, 46:109–132, 2013.
- [24] Donald P Green and Ian Shapiro. *Pathologies of rational choice theory: A critique of applications in political science*. Yale University Press, New Haven, CT, US, 1994.
- [25] Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Russell Sage Foundation., NY,NY, 2003.
- [26] Robert J Aumann. Rationality and bounded rationality. *Games Econ. Behav.*, 21(1):2–14, 1997.
-

- [27] J. O. Ledyard. *The Handbook of Experimental Economics*, chapter Public goods: A survey of experimental research, pages 111—194. Princeton Univ. Press, Princeton, NJ, 1997.
- [28] Vincent P. Crawford, Miguel A. Costa-Gomes, and Nagore Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *J. Econ. Lit.*, 51(1):5–62, March 2013.
- [29] Alan P. Kirman. Whom or what does the representative individual represent? *J. Econ. Perspect.*, 6(2):117–136, 1992.
- [30] Jake M. Hofman, Amit Sharma, and Duncan J. Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017.
- [31] Julia Poncela-Casasnovas, Mario Gutiérrez-Roig, Carlos Gracia-Lázaro, Julian Vicens, Jesús Gómez-Gardeñes, et al. Humans display a reduced set of consistent behavioral phenotypes in dyadic games. *Sci. Adv.*, 2(8):e1600451, 2016.
- [32] Martin G. Kocher, Todd Cherry, Stephan Kroll, Robert J. Netzer, and Matthias Sutter. Conditional cooperation on three continents. *Econ. Lett.*, 101(3):175–178, 2008.
- [33] Alexander Peysakhovich, Martin A. Nowak, and David G. Rand. Humans display a ‘cooperative phenotype’ that is domain general and temporally stable. *Nat. Comm.*, 5:4939, sep 2014.
- [34] M Gutiérrez-Roig, C Segura, J Duch, and J Perelló. Market imitation and win-stay lose-shift strategies emerge as unintended patterns in market direction guesses. *PLoS ONE*, 11:e0159078, 2016.
- [35] F. Vega-Redondo. *Economics and the Theory of Games*. Cambridge University Press, Cambridge, MA, US, 2003.
- [36] R. Guimerà and M. Sales-Pardo. Justice blocks and predictability of U.S. Supreme Court votes. *PLoS ONE*, 6(11):e27188, 2011.
- [37] Tiago P. Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Phys. Rev. X*, 5(1):011033, mar 2015.
- [38] Martin A. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Belknap Press of Harvard University Press, September 2006.
- [39] F. Y. Wu. The Potts model. *Rev. Mod. Phys.*, 54:235–268, Jan 1982.
- [40] H. E. Stanley. Dependence of critical properties on dimensionality of spins. *Phys. Rev. Lett.*, 20:589–592, Mar 1968.
- [41] Ido Erev and Alvin E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.*, 88(4):848–881, 1998.
-

- [42] P.-A. Chiappori, S. Levitt, and T. Groseclose. Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *Am. Econ. Rev.*, 92(4):1138–1151, September 2002.
- [43] M. E. J. Newman and Aaron Clauset. Structure and inference in annotated networks. *Nat. Comm.*, 7:11863, 06 2016.
- [44] Darko Hric, Tiago P. Peixoto, and Santo Fortunato. Network structure, metadata, and the prediction of missing nodes and annotations. *Phys. Rev. X*, 6:031038, Sep 2016.
- [45] D. Ariely. *Predictably Irrational*. HarperCollins, New York, NY, 2008.
- [46] M. Hasan Mohajeri, Robert J. M. Brummer, Robert A. Rastall, Rinse K. Weersma, Hermie J. M. Harmsen, et al. The role of the microbiome for human health: from basic science to clinical applications. *European Journal of Nutrition*, 57(1):1–14, May 2018.
- [47] Tao Zuo and Siew C. Ng. The gut microbiota in the pathogenesis and therapeutics of inflammatory bowel disease. *Frontiers in Microbiology*, 9:2247, 2018.
- [48] Christoph Becker, Markus F. Neurath, and Stefan Wirtz. The Intestinal Microbiota in Inflammatory Bowel Disease. *ILAR Journal*, 56(2):192–204, 08 2015.
- [49] Beatriz García-Jiménez and Mark D. Wilkinson. Robust and automatic definition of microbiome states. *PeerJ*, 7:e6657, March 2019.
- [50] Jordi Bascompte, Pedro Jordano, Carlos J Melián, and Jens M Olesen. The nested assembly of plant-animal mutualistic networks. *Proc. Natl. Acad. Sci. USA*, 100(16):9383–9387, Aug 2003.
- [51] Wenjun Liu, Jiachao Zhang, Chunyan Wu, Shunfeng Cai, Weiqiang Huang, et al. Unique features of ethnic mongolian gut microbiome revealed by metagenomic analysis. *Scientific Reports*, 6:34826 EP –, Oct 2016. Article.
- [52] Nan Qin, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513:59 EP –, Jul 2014. Article.
- [53] Melanie Schirmer, Sanne P. Smekens, Hera Vlamakis, Martin Jaeger, Marije Oosting, et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell*, 167(4):1125 – 1136.e8, 2016.
- [54] Curtis Huttenhower, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.
- [55] Jason Lloyd-Price, Anup Mahurkar, Gholamali Rahnavaard, Jonathan Crabtree, Joshua Orvis, et al. Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 550:61 EP –, Sep 2017. Article.
-



- 
- [56] David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, et al. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079 – 1094, 2015.
- [57] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.
- [58] Dan Knights, Tonya L. Ward, Christopher E. McKinlay, Hannah Miller, Antonio Gonzalez, et al. Rethinking “enterotypes”. *Cell Host & Microbe*, 16(4):433 – 437, 2014.
- [59] Omry Koren, Dan Knights, Antonio Gonzalez, Levi Waldron, Nicola Segata, et al. A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLOS Computational Biology*, 9(1):1–16, 01 2013.
- [60] Jun Wang, Miriam Linnenbrink, Sven Künzel, Ricardo Fernandes, Marie-Josée Nadeau, et al. Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. *Proceedings of the National Academy of Sciences*, 111(26):E2703–E2710, 2014.
- [61] Gary D. Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.
- [62] Lars Christensen, Henrik M Roager, Arne Astrup, and Mads F Hjorth. Microbial enterotypes in personalized nutrition and obesity management. *The American Journal of Clinical Nutrition*, 108(4):645–651, 09 2018.
- [63] Huanzi Zhong, John Penders, Zhun Shi, Huahui Ren, Kaiye Cai, et al. Impact of early events and lifestyle on the gut microbiota and metabolic phenotypes in young school-age children. *Microbiome*, 7(1):2, 2019.
- [64] Falk Hildebrand, Thi Loan Anh Nguyen, Brigitta Brinkman, Roberto Garcia Yunta, Benedicte Cauwe, et al. Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biology*, 14(1):R4, 2013.
- [65] J. Gregory Caporaso, Christian L. Lauber, Elizabeth K. Costello, Donna Berg-Lyons, Antonio Gonzalez, et al. Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50, 2011.
- [66] Maria Gloria Dominguez-Bello, Filipa Godoy-Vitorino, Rob Knight, and Martin J Blaser. Role of the microbiome in human development. *Gut*, 68(6):1108–1114, 2019.
- [67] Yasmine Belkaid and Timothy W Hand. Role of the microbiota in immunity and inflammation. *Cell*, 157(1):121–141, Mar 2014.
-

- [68] Judith L. Bronstein. *Mutualism*. Oxford University Press, Oxford, 2015.
- [69] J Bascompte and P Jordano. Plant-animal mutualistic networks: the architecture of biodiversity. *Annu. Rev. Ecol. Evol. Syst.*, 38:567–593, 2007.
- [70] Seraj Zohurul Haque and Mainul Haque. The ecological community of commensal, symbiotic, and pathogenic gastrointestinal microorganisms - an appraisal. *Clinical and experimental gastroenterology*, 10:91–103, May 2017. 28503071[pmid].
- [71] Ugo Bastolla, Miguel A Fortuna, Alberto Pascual-García, Antonio Ferrera, Bartolo Luque, et al. The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, 458(7241):1018–1020, Apr 2009.
- [72] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
-



UNIVERSITAT  
ROVIRA i VIRGILI

