Networks, Epidemics and Collective Behavior:

from Physics to Data Science

Alberto Aleta

Universidad de Zaragoza
2019

Abstract

In the final quarter of the XX century the classical reductionist approach that had been driving the development of physics was questioned. Instead, it was proposed that systems were arranged in hierarchies so that the upper level had to convey to the rules of the lower level, but at the same time it could also exhibit its own laws that could not be inferred from the ones of its fundamental constituents. This observation led to the creation of a new field known as complex systems. This novel view was, however, not restricted to purely physical systems. It was soon noticed that very different systems covering a huge array of fields, from ecology to sociology or economics, could also be analyzed as complex systems. Furthermore, it allowed physicists to contribute with their knowledge and methods in the development of research in those areas.

In this thesis we tackle problems covering three areas of complex systems: networks, which are one of the main mathematical tools used to study complex systems; epidemic spreading, which is one of the fields in which the application of a complex systems perspective has been more successful; and the study of collective behavior, which has attracted a lot of attention since data from human behavior in huge amounts has been made available thanks to social networks. In fact, data is also the main driver of our discussion of the other two areas. In particular, we use novel sources of data to challenge some of the classical assumptions that have been made in the study of networks as well as in the development of models of epidemic spreading.

In the case of networks, the problem of null models is addressed using tools coming from statistical physics. We show that anomalies in networks can be just a consequence of model oversimplification. Then, we extend the framework to generate contact networks for the spreading of diseases in populations in which both the contact structure and the age distribution of the population are important.

Next, we follow the historical development of mathematical epidemiology and revisit the assumptions that were made when there was no data about the real behavior of this kind of systems. We show that one of the most important quantities used in this kind of studies, the basic reproduction number, is not properly defined for real systems. Similarly, we extend the theoretical framework of epidemic spreading on directed networks to multilayer systems. Furthermore, we show that the challenge of incorporating data to models is not only restricted to the problem of obtaining it, but that it is also really important to be aware of its characteristics to do it properly.

Lastly, we conclude the thesis studying two examples of collective behavior using data extracted from online systems. We do so using techniques that were originally developed for other purposes, such as earthquake prediction. Yet, we demonstrate that they can also be used to study this new type of systems. Furthermore, we show that, despite their unique characteristics, they possess properties similar to the ones that have been observed in the offline world. This not only means that modern societies are intertwined with the online world, but it also signals that if we aim to understand socio-technical systems a holistic approach, as the one proposed by complex systems, is indispensable.

Contents

List of Figures

1 Introduction

Caminante, son tus huellas
el camino y nada más;
Caminante, no hay camino,
se hace camino al andar.

Al andar se hace el camino,
y al volver la vista atrás
se ve la senda que nunca
se ha de volver a pisar.

Caminante no hay camino
sino estelas en la mar.

Wanderer, it is your footprints
winding down, and nothing more;
wanderer, no roads lie waiting,
roads you make as you explore.

Step by step your road is charted,
and behind your turning head
lies the path that you have trodden,
not again for you to tread.

Wanderer, there are no roadways,
only wakes upon the sea.

("Proverbios y cantares", Antonio Machado)

"But although, as a matter of history, statistical mechanics owes its origin to investigations in thermodynamics, it seems eminently worthy of an independent development, both on account of the elegance and simplicity of its principles, and because it yields new results and places old truths in a new light in departments quite outside of thermodynamics" wrote Josiah W. Gibbs in the preface of his seminal book published in 1902, *Elementary principles in statistical mechanics* [1]. Yet, starting from that point might fool the reader into thinking that science progresses in *eureka* steps, isolated sparks of inspiration only attainable by geniuses. In reality, however, its progress is much more continuous than discrete. As Picasso said, *inspiration exists, but it has to find you working.*

XIX The birth of statistical mechanics

We shall begin this thesis with the work published by Carnot in 1824 [2]. The industrial revolution had brought steam engines all around Europe, completely reshaping the fabric of society. Yet, in Carnot's words, "their theory is very little understood, and the attempts to improve them are still directed almost by chance"[1]. Although clear efforts had been done in the pursuit of understanding the science behind what was yet to be named as thermodynamics, Carnot's work is usually regarded as the starting point of modern thermodynamics. The book was, however, slightly overlooked until 1834 when it laid on the hands of Clapeyron, who found its ideas "fertile and incontestable" [4]. In fact, it was Clapeyron the one that used the pressure-volume diagram (developed, in turn, by Watt in the late XVIII century) to represent the Carnot cycle, an image that is nowadays etched into the memory of every student of thermodynamics.

Subsequently, the ideas behind thermodynamics were developed mainly by Clausius and Kelvin, with the indispensable insights provided by experiments such as the ones carried out by Joule. Actually, it was Clausius the one that coined the term *entropy* in

[1] Quote extracted from the English translation published by Thurston in 1897 [3]

1865 [5] although, interestingly, he had obtained that same quantity a few years before, in 1854, but did not realise its full potential [6]. Albeit in a different shape, this concept will be one of the cornerstones of chapter 2.

To continue discussing the path followed by Clausius we need to bring two more theories to the table: the kinetic theory of gases and the theory of probability. The former theory stated that a gas was composed by many tiny particles or *atoms*, whose movement was responsible for the observed pressure and temperature of the gas. This theory had been proposed a century ago, in 1738, by Daniel Bernoulli, although it did not attract much attention at the time [7]. The theory of probability, on the other hand, had received several contributions along the years, being the one by Daniel's uncle, Jacob Bernoulli, one of the most well known (for instance, Bernoulli trials or the Bernoulli distribution are named after him). Probability had been regarded for some time as something mainly related to gambling[2], but step by step it started to loose that negative connotation when scientists in the XVIII century introduced it into error theory for data analysis. The process was then firmly established with the works by Gauss in the early 1800s and by the middle of the century it was already common in physics textbooks[3].

This leads us to the year 1859. Clausius had published a work about molecular movement for the kinetic theory of gases that, at first glance, implied that molecules could move freely in space. This was criticized by some scientists as, if it were true, they wandered why clouds of tobacco smoke extended slowly rather than quickly filling up the whole room. Clausius regarded that objection legitimate and further developed his theory to account for "how far on an average can the molecule move, before its centre of gravity comes into the sphere of action of another molecule" [10]. In other words, he calculated what we know today as the mean free path. Furthermore, he introduced the concept of average speeds and random impacts. As we will see, this work was fundamental for the development of statistical physics.

St John's College, in Cambridge, started a scientific competition in 1855 whose solution was to be delivered by the end of 1857. The problem was to explain the composition of Saturn's rings, something that had puzzled scientists for over 200 years. Laplace had already shown that a solid ring would be unstable, but nevertheless the examiners proposed three hypothesis, the rings could be: solid, liquid or composed of many separate pieces of matter. The only contestant that submitted a proposal was James C. Maxwell, who showed that the only stable solution was the last one, granting him the prize [11]. Interestingly, he claimed in his solution that collisions between those pieces were possible, but that he was unable to calculate them [12]. But then, just three years later, in 1860, he derived an equation that today is regarded as the origin of statistical mechanics, the Maxwell distribution of velocities, obtained precisely by calculating collisions between particles [13]. In that three years period only one thing had change, the paper by Clausius in 1859, which he actually cites at the beginning of his paper. Notwithstanding his great achievement, it seems clear that the spark was not isolated, but came from a burning wick.

At this point we need to add a new scientist to the group, Ludwig E. Boltzmann. The scientific career of Boltzmann started in 1866, when he tried to give an analytical proof of the second law of thermodynamics. His actual accomplishment was quite modest. Yet, two years later, in 1868, he changed his approach and started a series of lengthy memoirs where he extended the results from Maxwell resulting into the full development of statistical mechanics [14]. The fundamental work by Boltzmann was, in turn, expanded by Gibbs in the classical treatise of 1902 that opened this chapter [1] (although some of his ideas had been already proposed by Boltzmann, they had been slightly overlooked by his colleges,

[2] In Maxwell's own words: "This branch of Math., which is generally thought to favour gambling, dicing and wagering and therefore highly immoral [...]" [8]

[3] See [9] for an overview of the introduction of probability into physics.

see [15] for a discussion on why this might have happened). The development of statistical mechanics would culminate in 1905 with the work by Einstein on the Brownian motion, regarded by Born as the final proof for physicists "of the reality of atoms and molecules, of the kinetic theory of heat, and of the fundamental part of probability in the natural laws" [16].

It should be noted, however, that one of the key elements of the theory introduced by the aforementioned scientists was the concept of *ensemble*, that we will further describe in chapter 2. In spite of its importance, their use raised some mathematical problems that they could not solve, "but like the good physicists they were, they assumed that everything was or could be made all right mathematically and went on with the physics" [17]. Some years later this lead to the subject of *ergodic theory* which we will not address in this thesis.

XIX.1 Meanwhile, in the social sciences

The XIX century can be also acknowledged as the century when social science was born. There were multiple factors leading to such enterprise, of which we might highlight: the social changes induced by the industrial revolution, the standardization of statistical approaches beyond physics and the publication of Darwin's *On the origin of Species*. Some of the ideas developed in these areas, perhaps surprisingly, echoed in the own development of physics in the XX century.

To give a brief overview, we shall start with Quetelet's view of statistics. Back in the beginning of the XIX century, statistics was mostly restricted to the calculation of errors in astronomy. In 1823 Quetelet traveled to Paris to study astronomical activities, and he became impassioned by the subject of probability. Since then, he went on to put it to practical use in the study of the human body, in an attempt to find the *average man*. This led to the creation of the Body Mass Index, which we still use today [18]. In subsequent years he developed his ideas further and applied statistics not only to the human body but also to states and even to the *social body*, i.e. the aggregation of the whole human race. We find particularly interesting that, as one of the precursors of social sciences, he believed that it was possible to find laws for the social body "as fixed as those which govern the heavenly bodies: [like in] physics, where the freewill of man is entirely effaced, so that the work of the Creator may predominate without hindrance. The collection of these laws, which exist independently of time and of the caprices of man, form a separate science, which I have considered myself entitled to name *social physics*"[4].

Another important (for our interests) branch of science that started in this century is the mathematical study of demography. In the beginning of the XIX century it was claimed that surnames were being lost (particularly among the nobility). Francis Galton, a famous statistician (and cousin of Darwin), thought that this could be addresses mathematically and put it as an open problem for the readers of Educational times in 1873. The proposed solutions did not please him, so he joined another mathematician, Henry W. Watson, and together developed the theory that later came to be known as the Galton-Watson process [21]. Their theory, based on generating functions, was a novel way of tackling the study of demography, specially for not being deterministic. This seemingly ingenuous problem is credited as the origin of the theory of branching processes [22], which in turn was very important in the development of graph theory, epidemiology and the theory of point processes, as we shall see in chapters 2, 3 and 4 respectively. Furthermore, this

[4] August Comte, the father of sociology, held a similar view. He proposed that sciences could be arranged in order of generality of their theories and complexity: astronomy, physics, chemistry and physiology. However, there was one type of phenomena yet to be addressed, the "most individual, the most complicated, the most dependent on all others, and therefore [...] the latest", social phenomena. Oddly enough, he coined the term *social physics* to refer to this new branch of science that had to be affected in part by physiology and, at the same time, by the influence of individuals over each other [19]. However, once he discovered that Quetelet had used the same term, he changed it to *sociology* [20].

problem, together with the great impact that the book of his cousin had on him, lead him to the study of the infamous "cultivation of race" or *eugenics* [23]. Lastly, in 1906 he went back to the roots of statistics and performed the first experiment on collective intelligence to which we shall further return in chapter 4.

XX The century of Big Science

The enormous expenditures for research and development during World War II brought a revolution in the physical sciences [24]. For instance, branching processes and Monte Carlo methods, which we will use throughout this thesis, were developed at that time. Branching processes (term coined by Kolmogorov [25]), following the path started by Galton in the previous century, were used to study neutron chain reactions and cosmic radiation [26]. In turn, Monte Carlo methods were a tool used to study several stochastic processes. In particular, we will use these methods in chapters 3 and 4 (see [27] for a nice review of their history and the origin of the name).

Both branching processes and Monte Carlo methods are intimately related to percolation processes, proposed by Broadbent and Hammersley in 1956. Initially, the idea was to study the diffusion of a fluid in a medium but focusing on the medium, in contrast to common diffusion processes that used to focus on the fluid, in order to design better gas masks for coal miners (Broadbent received support from the British Coal Utilisation Research Association during his PhD[5]). Interestingly, though, they gave examples of which problems could be tackled with this formalism and included the spreading of a disease in an orchad [29]. In recent years these processes have then been applied to the study of disease spreading and network percolation, as we will see in chapter 3.

But there was also space for fundamental research. Using the information theory established by Shannon [30], Jaynes proposed in 1957 that statistical physics could be derived from an information point of view [31]. Rather than deriving the theory from dynamical arguments, he argued that the objective of statistical physics was to infer which probability distribution was consistent with data while having the least possible bias respect to all other degrees of freedom of the system. In this sense, the entropy would be a measure of the information about the system, so that maximizing the entropy would be equivalent to maximizing the ignorance subject to the data that is known to be true. This view was not unanimously embraced, as some scientists believed that this definition depends on the observer, which goes against the fact that entropy is a definite physical quantity that can be measured in the laboratory [32].

This debate is still open today. For instance, in the book by Callen that is widely used to teach thermodynamics in physics courses, this view of statistical mechanics is deeemed to be "a subjective science of prediction". Instead, he proposes the more common view of entropy as a measure of "objective" disorder [33]. Yet, if one goes to the original work by Bridgman, published in 1943, where he introduces the notion of disorder, he claims that the definition is "anthropomorphic" and not absolute [34]. Others, like Ben-Naim, claim that the problem lays in the history of the development of statistical mechanics. As we saw in XIX, this discipline started from thermodynamics. Hence, entropy was defined as a quantity of heat divided by temperature, yielding units of energy over temperature. If entropy, instead, refers to information, it has to be a dimensionless quantity. The problem, he argues, derives from the fact that the concept of temperature was developed in a pre-atomistic era. Once Maxwell identified the temperature with the kinetic energy of atoms, the own definition of temperature could have been changed into the units of energy. In such case, heat over temperature would result in a dimensionless quantity, making it much easier to be accepted as a measure of information. Furthermore, he claims that most

[5]For a historical review of branching processes in general and their relation to percolation processes see [28]

people have misinterpreted the own concept of information in this context, see [35]. In spite of these controversies, in chapter 2 we will follow Jaynes' definition to be able to apply the formalism of statistical physics to graphs.

Jaynes' proposal is one of the first hints about the usefulness of statistical physics outside the classical realm of physics. In particular, the framework of statistical physics has shown to be quite useful in a new branch of science that started to develop during the 1960s, complex systems.

XX.1 More is Different

This section is named after the famous paper by P. W. Anderson, where he claimed that the reductionist approach followed by physicisits up to that moment had to be revisited [36]. He argued that obtaining information about the fundamental components of a system did not mean that you could then understand the behavior of the whole. Instead, he proposed that systems were arranged in hierarchies so that each upper level had to convey to the rules of the lower level, but it could also exhibit its own laws that could not be inferred from the ones of the fundamental constituents. His book *More and Different* offers some insights about the reasons that led him to write that paper[6], as well as a glimpse of how condensed matter physics was born from the ashes of the Second Wold War [38].

This view of systems as a multiscale arrangement grew during the 1960s and 1970s lead by several observations that would end up composing what we know as complex systems. For instance, in condensed matter physics the intereset in disordered systems (such as spin glasses or polymer networks) started to increase. The expertise obtained with these models, allowed for the study of collective behavior of completely different systems but whose components were also heterogeneous, such as in biological systems [39].

Another example of results that composed the early theory of complex systems is chaos theory. Despite the pioneering work by Poincarè in the late XIX century, chaos theory was mainly developed in the middle of the XX century [40]. On of its main starting points was the study by Lorenz in 1963 on hydrodynamics [41]. In that study, he showed that a simple nonlinear differential system of equations, meant to reproduce weather dynamics, exhibited wildly different evolution with very similar starting conditions. In other words, he had discovered chaos. His study had a huge impact in the community because not only it opened the door to a whole new area of research, but also in the particular case of weather prediction signaled that maybe long term predictions were not attainable, as perfect knowledge of the initial conditions was not achievable. He summarized that statement saying "if the theory were correct, one flap of a sea gull's wings would be enough to alter the course of the weather forever" (the sea gull was turned into a butterfly latter on for aesthetic reasons) [42].

From that point forward, the study of nonlinear systems explode. During the 1970s mathematicians, physicists, biologists, chemists, physiologists, ecologists... found a way through disorder. One of those scientists was R. May, a theoretical physicist who initially focused his research on superconductivity. However, he was suddenly trapped by the ideas behind nonlinear equations and their application to population dynamics [43]. In fact he is considered the founder of theoretical ecology in the 1970s, and as we shall see in chapter 3, his contributions to theoretical epidemiology were also outstanding.

In the following decades many more concepts where added to the body of complex systems: critical phenomena, self-similarity, fractals... It was found that several systems

[6] Although we need to look for another source for the origin of the own sentece *more is different*. According to Pietronero[37], Anderson confessed that the paper originated from a sort of resentment that physicists in the field of condensed matter developed with respect to a certain arrogance of the field of elementary particles, who thought that their research was the only true intellectual challenge. Back in those dates the British environmental movement had various slogans such as "small is beautifull" and "more is worse", from which he drew inspiration.

from very different fields exhibited similar properties, such as scale free distributions [44]. Unfortunately, their greatest strength is also their main weakness. Complex systems are everywhere, but there is not yet a universal law that can be applied to a wide array of them. For some time, it was proposed that self-sustained criticality (critical phenomena arising independently of the initial condition) could be that holy grial [45] but it turned out not to be the case [46]. Having scientists from so many different backgrounds tackling problems set in such a wide array of fields (from molecular biology to economy, urban studies or disordered systems) without common laws tying them up together, renders even the own definition of a complex system a daunting task. Finding that common law, or framework, common to all such systems is still today one of the greatest challenges in the field.

This problem, unifying complex systems, has been addressed from many diverse perspectives. For instance, by the end of the century an international group of ecologists, economists, socials scientists and mathematicians collaborated in the "Resilience project" with the objective of deepen the understanding of linked socio-ecological systems. From that 5-year project they developed the concept of *panarchy*, in an attempt to remove the rigid top-down nature that is associated with hierarchies [47]. Interestingly, they claimed to be inspired by the work by Simon, but do not mention the work published by Anderson in 1972 that is essentially the same concept, but in the context of physics. In turn, Anderson did not cite Simon in his article, even though he had been discussing on the problem of hierarchies since the 1960s [48]. This is a great example of the fragmentation that ballasts the development of complex systems.

The view of Anderson in this matter is particularly interesting. He argued that physics in the XX century solved problems that had clear hierarchical levels such as atomic theory, electroweak theory or classical hydrodynamics. Consequently, the XXI century should be devoted to building "generalizations that jump and jumble the hierarchies". Furthermore, he claimed that, by embracing complexity, the "theorist" will no longer be confined by a modifier specifying "physics", "biology" or "economics" [38].

An arguably more pesimystic view is presented by Newman in his great resource letter for complex systems. He claims that since there is not a general theory of complex systems, and it might never arrive, maybe it should be better to talk about "general theories" as complex systems is not a monolithic body of knowledge. He summarizes this view saying that *"complex systems theory it not a novel, but a series of short stories"*. [49]. In this thesis we will revise some of those stories.

XX.2 From lattices to networks

Towards the end of the century, however, the initial hype on nonlinear dynamical systems started to decline. It was time to add more details into the models, and consequently more data, although it was a more challenging venture than it might seem. For instance, the common approach of making theoretical predictions and comparing them with experiments is not that straightforward in complex systems, as the own definition of prediction can have very different meanings in chaos or stochastic systems due to the extreme sensibility to the initial conditions. Thus, the comparison with data in these systems had to focus initially on extracting universal patterns rather than going to specific details [50]. A great leap forward in this direction was the introduction of the classical graph theory developed in the middle of the XX century in the shape of *networks*.

A holistic view of a system implies that its components are no longer isolated and their interactions have to be properly taken into account. Networks represent a particularly useful tool for such endeavor. In order to gauge their importance in complex systems, suffice to say that in the Conference of Complex Systems of 2018 out of over 400 contributions, 60% of them explicitly mention the term "network" in their abstract. However, networks were not originated along complex systems, but way before.

The origin of networks as a tool to study other systems, rather than as a mathematical object of their own, began in 1934. The psychologist Jacob Moreno proposed that the evolution of society might be based on some laws, which he wanted to uncover in order to develop better therapies to treat his patients. In order to do so, he proposed to study communities up to their "social atoms". He then studied the relations between those social atoms, which in his first work were babies and children in a school. This represented a shift from classical sociological and psychological studies, where the attributes of the actors (a generic term used to refer to the element under study in sociology) used to be more important. Furthermore, he represented the actors in his studies with circles and connected them with lines if they had some relation. These diagrams, which he denominated *sociograms*, where the first examples of networks [51]. This procedure was mostly forgotten until the 1960s, when it was picked up by sociologists who further developed the theory of networks, with new tools and frameworks[7].

A particularly interesting example of social research on networks is the well-known study performed by Milgram in 1967 [53]. In his experiment, Milgram sent a package to random individuals with the instructions that, if they wanted to participate, they should send the package to someone they knew that might, in their opinion, new a person that Milgram had chosen, or at least get closer to her. The purpose of the experiment was to determine the number of steps that the package had to take to navigate the social network of the country. Interestingly, he found that the *average path length* was close to six. These results led to the notion of *six degrees of separation* and *small world* that have been part of popular folklore ever since.

Yet, the use of networks remained constrained to the fields of sociology and some areas of mathematics until the end of the century, when research in networks exploded in several fields at the same time. This, however, meant that lot of advances that had been done during decades were not widely known, leading to multiple rediscoveries of the same concepts. For instance, the multiplex networks that we will see in chapter 2 were introduced around 2010, although the term *multiplex* had been coined in 1955 by the anthropologist Max Gluckman during his studies of judicial processes among the Barotse [54]. Similarly, Park and Newman introduced in 2004 [55] the exponential random graph model that had been already developed in 1981 by Holland and Leinhardt [56]. Nonetheless, it should be noted that they acknowledge the work by Holland and Leinhardt in their paper and present a different formulation of the model. In fact, chapter 2 will mostly be devoted to the formulation by Newman and Park which, in turn, is based in the statistical physics framework proposed by Jaynes that we discussed in earlier.

The first paper of what we might call the "modern" view of networks, was the work by Watts and Strogatz on small-world networks, published in 1998 [57]. In their work, they took three very different networks (the neural network of a worm, the power grid of the western United States and the collaboration network of film actors) and measured their average path length. Surprisingly, they found that the three of them exhibited the same small-world behavior as observed by Milgram 30 years before. Besides, they created a model to explain those networks that interpolated between the well-known lattices and random networks. One may wonder, then, why that paper was so successful if most results were one way or another already known. And the answer, we believe, is data and universality.

Indeed, in sociology most networks analyzed were fairly small, as they were collected manually. The spreading of the internet in the 1990s, however, allowed scientists to share information in an unprecedented way. Even more, larger sets of data could be analyzed and stored. The fact that they showed that three large systems of completely different nature had the same properties was determinant in its success. In fact, those networks were not extremely interesting on their own, they selected them "largely because they were

[7] See [52] for the history of the development of social network analysis.

available in a format that could easily be put into a computer as the adjacency matrix of a graph" [58]. But it was clearly a good choice. Thanks to that variety, researchers from many different areas saw small-world networks as a Rorschach test, in which every scientist saw different problems depending on their disciplines [59].

We can summarize this point of view using Stanley et al. words in the seminal paper that started the area of econophysics, "if what we discovered is so straightforward, why was it not done before? [Becuase] a truly gargantuan amount of data analysis was required" [60].

XXI The Information Age

Undoubtedly, we live in the information age. To put it into perspective, while in the small-world paper previously mentioned 3 networks were used, in figure 2.4 we will compare 1,326 networks that were collected with just a couple of clicks.

Obtaining meaningful information from high-dimensional and noisy data is not an easy task. To achieve this, the limits of the theoretical framework of statistical physics will have to be extended [61]. A large amount of data also means data from very different sources. Hence, we need to combine temporal and spatial scales and nonlinear effects in the context of out of equilibrium systems. Furthermore, it is not only important to extract the information and build appropriate models to increase our knowledge of a given system, but also to develop quantities that might be useful to describe multiple sociotechnical systems at the same time [62].

For instance, in epidemics data from very different scales, from flight data to physical contact patterns can be combined, together with economic and social analysis to produce much more informative spreading models. But mathematical models able to capitalize such data stream are not available yet [63]. The question on whether the information shared in the internet can be used to track epidemic evolution is also open. Google Flu claimed that could predict the evolution of flu using search statistics, but it was shown that it was better at predicting winter than diseases [64].

Interestingly though, one of the areas which might seem would benefit more for having huge amounts of data about human behavior and communications patterns, sociology, has not embraced it yet [65]. This is even more striking given that the precursors of sociology, Quetelet and Comte, as previously discussed, believed in the possibility of addressing social systems in a similar fashion as other experimental sciences, i.e., with data.

There is currently a huge debate in sociology about the impact that this amount of data can have in the own field. For instance, McFarland et al. talk about sociology being subverted to computer science. Their fear is that data might be used only to seek for solutions without explaining why. Moreover, they argue that the scientific culture of both disciplines are completely different. While computer science is characterized by large collaborations, fast review periods and quick development, sociology is slower, with larger review periods, more theory and a more "monastic" science [66]. But they also observe that it is a new opportunity, as data from behaviors that could not be analyzed before is being collected now. In fact, data about new behaviors, which deserve scientific analysis is also being collected, as we will see in chapter 4. Their proposal is to move towards a *forensic social science* in which applied and theory-driven perspectives are merged.

A similar approach is proposed by Halford and Savage, who fear that big data might corner sociology into a defensive position [67]. Instead, they propose to forget about inductive theory and woven it with data, in what they call *symphonic social science*. Even more, they believe that the limitations of the data, under proper guidance, can be leveraged. For instance, it is known that most classical psychological experiments are done on *WEIRD* population (western, educated from industrialized, rich, democracies) [68]. On the other hand, Twitter has a disproportionate number of young, male black and Hispanic users

compared to the national population. Thus, it might offer some insights into groups that are underrepresented in some traditional scenarios.

There are clear signs that the interconectedness of society is bringing changes into our sociotechnical systems, even if they are not yet understood. For instance, it has been observed that since the appearance of Google Scholar the citation patterns among scholars have changed [69]. Nowadays, older articles are being cited more commonly than before and, at the same time, non top journals are getting more attention [70]. Still, many sociologists remain unconvinced that the sources of data and methods present something new or claim that instead of studying society, the use of data will lead us to study technology instead. But maybe society and techonology cannot be disentangled anymore, and they have to be addressed together [71]. In terms of Castells, we live in the culture of real virtuality and society is no longer structured over individual actors but around networks [72].

In any case, it is clear that in the XXI century the world will not longer be controlled by those who merely posses the information, but by those who are able to understand it. In Edward O. Wilson words, *we are drowning in information, while starving for wisdom* [73].

2 Statistical mechanics of networks

In this chapter, we will introduce some of the most basic concepts in network science. We will begin by giving a brief overview of the mathematical framework used to study networks in section 2.1. This section is not meant to be a thorough introduction into the field, for which several good books are available [74, 75, 76], but rather as a way of establishing the terminology that will be used for the rest of this thesis. For this reason, we will only focus on those properties of networks that are indispensable for its correct comprehension. Thus, key concepts in network science such as clustering or community structure will not be addressed here, as they are not used in the works composing this thesis.

Special attention will be paid to multilayer networks in section 2.1.3. These networks are a particular generalization of classical graphs that, as we shall see, will play a key role in several sections of this thesis. The overview of multilayer networks will be based on one of the works developed during this thesis:

- A. Aleta and Y. Moreno, Multilayer Networks in a Nutshell, *Annu. Rev. Condens. Matter Phys.*, vol. 10, pp. 45–62, Mar 2019.

Next, in section 2.2, we will introduce the problem for which this chapter is devoted, generating appropriate null models of real networks. Our methodology will be based on the exponential random graph (ERG) model, which will be introduced in section 2.3. In section 2.4, we will show how this model can be used as a null model of real networks. As we shall see, the mathematical framework of ERGs has clear similarities with statistical mechanics, hence the name of this chapter.

Lastly, we will apply this framework to study two particular problems. In the first one, section 2.5, we will analyze several transportation networks in an attempt to determine if there are anomalies in them. Anomalies in this context refers to properties that would differ significantly from what is found in a null model of the network, and we will show that the assumptions implicit in those models might be one of the main causes of the anomalies. This section will be based on the work

- L. G. A. Alves, A. Aleta, F. A. Rodrigues, Y. Moreno, and L. A. Nunes Amaral, Centrality anomalies in complex networks as a result of model over-simplification, *Sent for publication*, 2019.

of which I am first co-author.

The second problem, presented in section 2.6, will focus instead on generating synthetic networks from real data. In particular, we will use socio-demographic data in order to build multilayer networks in which the mixing patterns of the population are properly accounted for. In section 3.4, we will measure the impact that incorporating this kind of data can have on disease dynamics. Both sections will be based on the work

- A. Aleta, G. Ferraz de Arruda, and Y. Moreno, Generating data-driven age contact networks, *In preparation*, 2019.

2.1 Brief introduction to graph theory

Graph theory is the mathematical framework that allows us to encode the properties of real networks into a mathematically tractable object. In fact, the term *network* is often used when talking about a real system while *graph* is usually used to discuss its mathematical representation. Yet, this distinction is seldom made and the terms graph and network are nowadays synonyms of each other [76]. Hence, during this thesis both terms might be used interchangeably.

Following [74], we define a network as a set of entities, which we will refer to as *nodes*, that are related to each other somehow. Said relationship will be encoded in *links* connecting the nodes. This very general definition allows us to describe systems of very diverse nature. For instance, in spin glasses two nodes (spins) will have a link if they can interact with each other. In other systems, such as transportation networks, two nodes (e.g. cities) cannot interact per se in the strict sense, but if travelers can go from one destination to the other somehow, we will establish a link between them to encode the relationship.

In general, it is possible to have more than one link between two nodes. For example, in order to increase the resilience of a power grid, there could be two independent transmission lines between two cities. In these cases, we refer to those links as *multilinks*. Similarly, it is possible to find nodes connected to themselves through *self-loops*. However, in this thesis, we will mostly work with *simple graphs*. That is, graphs with neither self-loops nor multiedges.

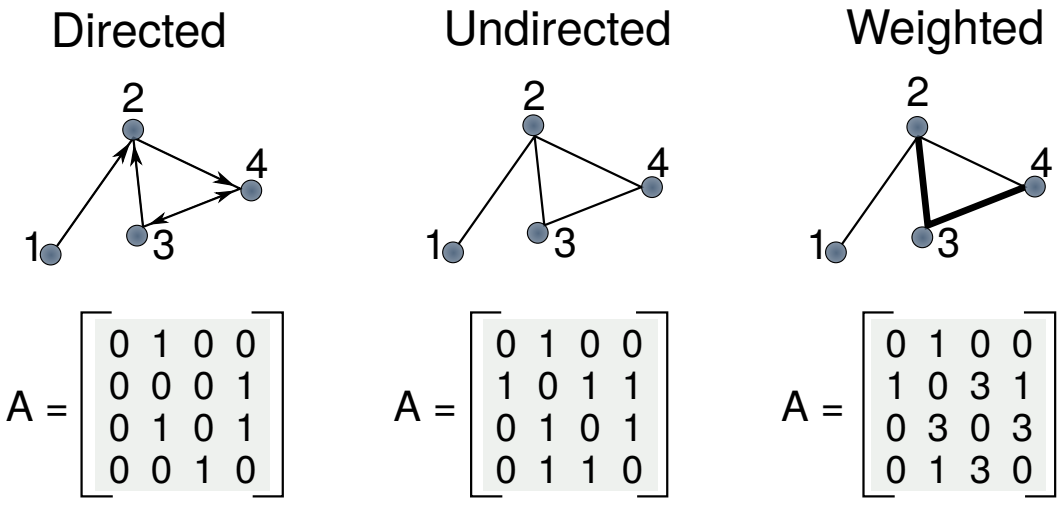

Figure 2.1: Schematic representation of directed, undirected and weighted networks. In directed graphs a relationship from i to j does not imply that the reciprocal is true, hence the adjacency matrix is not symmetric. In undirected graphs all relationships are reciprocal, making the matrix symmetric. In weighted graphs links have associated weights which can represent various quantities such as the strength or duration of the interaction.

The fundamental mathematical representation of a network is the *adjacency matrix*, A. Given a graph composed by N nodes, its adjacency matrix is a square matrix of size $N \times N$ in which $a_{ij} = 0$ if there is no relationship between nodes i and j and $a_{ij} \neq 0$ otherwise. Depending on the values allowed for a_{ij} we can define several types of graphs. The most common one is the undirected binary graph, undirected graph or, simply, graph, in which the matrix is symmetric and a_{ij} can only take binary values. If we relax the symmetry condition we obtain a *directed graph*, in which a relationship from node i to node j does

not imply the existence of a relationship from j to i. Even more, if we allow a_{ij} to take any positive value, we obtain a *weighted graph*. In these networks links have an associated weight that represents the relative strength of the interaction. A schematic representation of these three types of networks is shown in figure 2.1. During this thesis we will explore the three of them, although the undirected graph is admittedly the most common one.

Although the adjacency matrix completely describes a network, it is rather difficult for humans to comprehend the structure of the network by just looking at its adjacency matrix. For this reason, it is common practice to define mathematical measures that capture interesting features of network structure quantitatively. In the following section we will describe the ones that will be used all over the thesis.

2.1.1 Topological properties of networks

The most basic microscopic property is the *degree* of the nodes, which represents the number of links it has. For a network of N nodes the degree can be written in terms of the adjacency matrix as

$$k_i = \sum_{j=1}^{N} a_{ij} \,. \tag{2.1}$$

This quantity plays a key role in many networks. In fact, networks can be classified according to their *degree distribution*, $P(k)$, which provides the probability that a randomly selected node in the network has degree k. It has been observed that the precise functional form of $P(k)$ determines many network phenomena in a wide array of systems. Besides, with this distribution, it is possible to obtain other important quantities such as the *average degree* of a network

$$\langle k \rangle = \sum_{k=0}^{\infty} kP(k) \tag{2.2}$$

In directed networks (2.1) is slightly different as the network is not symmetric. Thus, we have to define the *in-degree* as the number of incoming links to a node and the *out-degree* as the number of outgoing links,

$$k_i^{\text{in}} = \sum_{j=1}^{N} a_{ij} \,, \qquad k_i^{\text{out}} = \sum_{j=1}^{N} a_{ji} \tag{2.3}$$

In the case of weighted networks, it is common to denote with the binary quantity a_{ij} the existence of a link, while the weight is encoded in the variable w_{ij}. This way, it is possible to use the same definition for the degree (2.1) and, similarly, define the *strength* of a node as

$$s_i = \sum_{j=1}^{N} w_{ij} \,. \tag{2.4}$$

In a similar fashion, one can define as many observables as desired, some of which will be more important than others depending on the system under consideration. In particular, a lot of effort has been devoted to the concept of *centrality* in networks, which tries to answer the question of which are the most important nodes in a network. There are multiple ways of defining the centrality of a node and the best one usually depends on the specific system that is being analyzed. For instance, the degree of a node by itself can be considered as a measure of its importance. However, as we shall see, our interest in this thesis lays on the role nodes play in transportation networks. Hence, for us, it is more interesting to know which nodes are indispensable for the correct operation of the network rather than which are the most popular. For instance, one can think of two large parts of a city divided by a river with one bridge. It is clear than in said situation the bridge holds

a key position in the system, as if there was a problem there, both parts of the city would become disconnected. A quantity that can capture this information is the *betweenness*.

To understand betweenness we first need to talk about *paths*. A path is a route that runs along the links of a network without stepping twice over the same link. We define the *shortest path* between nodes i and j as the path between them with the fewest number of links. Note that if we had a weighted network representing some spatial system, with the real spatial distance acting as weight, this is equivalent to simply look for the route between two nodes that minimizes the total distance traveled.

Denoting by σ_{rs}^i the number of shortest paths between nodes r and s that pass through i, the betweenness of node i can be defined as

$$b_i = \frac{2}{(N-1)(N-2)} \sum_{r \neq i} \sum_{s \neq i} \frac{\sigma_{rs}^i}{\sigma_{rs}} , \tag{2.5}$$

that is, the fraction of all shortest paths in the system that include node i without starting or ending in it. The leading factor is just a normalization constant so that the betweenness of networks of different sizes can be compared.

2.1.2 Important degree distributions

As previously mentioned, the degree distribution of a network can determine many of its properties. Although any probability distribution can be used as a degree distribution, there are two prototypical distributions to which this section will be devoted: the Poisson distribution and the power-law distribution.

The Poisson distribution attains its importance from being the distribution that arises naturally in *the random graph* model. In general, a random graph is a model network in which the values of certain properties are fixed, but it is in other respects random. One can think of many properties that could be settled, but undoubtedly the simplest choice is to establish the number of nodes N and the number of links L. Such model is known as *Erdős-Rényi model* (honoring the contributions of Paul Erdős and Alfréd Rényi in the study of the model [80]), *ER graph*, *Poisson random graph* or, simply, *"the" random graph*. Henceforth, we will refer to this model as ER.

In the ER model the only two elements that are fixed are the number of nodes, N, and the number of links, L. Hence, this model does not define a single network but rather a whole collection of networks, or *ensemble*, that are compatible with those constraints. Indeed, as there are $\binom{N}{2}$ ways of selecting pairs of nodes, there are $\binom{\binom{N}{2}}{L}$ ways of placing the L links, or different graphs. The probability of selecting any of those graphs will be given by

$$P(G) = \frac{1}{\binom{\binom{N}{2}}{L}} . \tag{2.6}$$

Although this is the original formulation of the model, nowadays it is more common to use a definition that is completely equivalent for large N. Specifically, one defines p as the probability of including any possible link, independently from the rest of links. Therefore, for instance, the average degree in both formulations would be

$$\langle k \rangle = \frac{2L}{N} = p(N-1) . \tag{2.7}$$

To obtain the degree distribution of this model we need to consider that a given node in the network will be connected with probability p to each of the $N-1$ other nodes. Thus the probability of being connected to k nodes and not to the rest of them is $p^k(1-p)^{N-1-k}$.

As there are $\binom{N-1}{k}$ ways to choose k nodes, the probability of being connected to k nodes is

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}\,, \tag{2.8}$$

which is a binomial distribution. Yet, in many cases we are interested in the properties of large networks, so that N can be assumed to be large. In the large-N limit equation (2.8) tends to a Poisson distribution,

$$P(k) = \frac{(pN)^k e^{-pN}}{k!} \approx \frac{\langle k\rangle^k e^{-\langle k\rangle}}{k!}\,. \tag{2.9}$$

Due to its simplicity, this model is often used as the basic benchmark to determine if a real network has non trivial topological structures. Yet, most real networks do not resemble an ER graph at all. Actually, the most common distribution of real networks is a *power law distribution*. A power law distribution can be expressed as

$$P(k) = Ck^{-\gamma} \tag{2.10}$$

where C is the normalization constant that ensures that the probability is correctly defined. Networks with a degree sequence that follows a power law distribution are known as *scale-free* (SF) networks.

The term scale-free comes from the fact that the moments of the power-law distribution,

$$\langle k^n\rangle = \int_{k_{\min}}^{k_{\max}} k^n P(k)\mathrm{d}k = C\frac{k_{\max}^{n-\gamma+1} - k_{\min}^{n-\gamma+1}}{n-\gamma+1}\,, \tag{2.11}$$

diverge under certain conditions. In particular, if $2 < \gamma < 3$ and $k_{\max} \sim \sqrt{N} \to \infty$ the first moment is finite, but the second moment diverges. Interestingly, most networks following a power law distribution have an exponent between 2 and 3. Hence, the fluctuations around $\langle k\rangle$ are so large that a given node could have a very tiny degree or an arbitrarily large one. In other words, there is no "scale" in the network. In contrast, if the degree distribution is Poisson, a node is expected to have degree in the range $k = \langle k\rangle \pm \langle k\rangle^{1/2}$. Thus, in those cases, $\langle k\rangle$ serves as the scale of the network.

This property of both degree distributions gives rise also to another terminology. Sometimes ER networks are said to be *homogeneous* while SF networks are called *heterogeneous*. As we shall see, this heterogeneity is the origin of some of the most interesting phenomena in network science.

2.1.3 Multilayer graphs

The methodology presented so far has always considered that all the links in a network represent the same kind of relationship. But, for instance, if we think of a social interaction network, it is clear that relationships can be of very diverse nature. Thus, given a diverse system, we can classify its interactions into groups according to their characteristics. This classification yields a set of networks, one for each interaction, related to each other. The way in which these networks are connected to one another, the entities their nodes represent and the way their relationships are represented, produce a new set of networks that goes beyond the concept of simple graphs. We call these structures *multilayer networks* [77]. When talking about multilayer networks, one often refers to the graphs presented so far as single-layer networks or monoplex networks.

In multilayer networks, we have an extra ingredient apart from nodes and links, *layers*, which contain the networks defined by each interaction type as discussed above. In its most general form, a multilayer network is composed by nodes that can be connected to other nodes in the same layer or to nodes in different layers. Then, depending on the specific

system under consideration, it is possible to have several types of multilayer networks. For instance, if nodes represent the same entity in all layers we say that we have a *multiplex* network (although the definition of multiplex network is slightly more general, as the main requirement is that all nodes have their counterparts in other layers, regardless of them representing the same entity or not). A paradigmatic example of such objects are social networks, where nodes represent individuals participating in social interactions. If an individual has relationships in two different contexts, we will find the node representing that person in two layers.

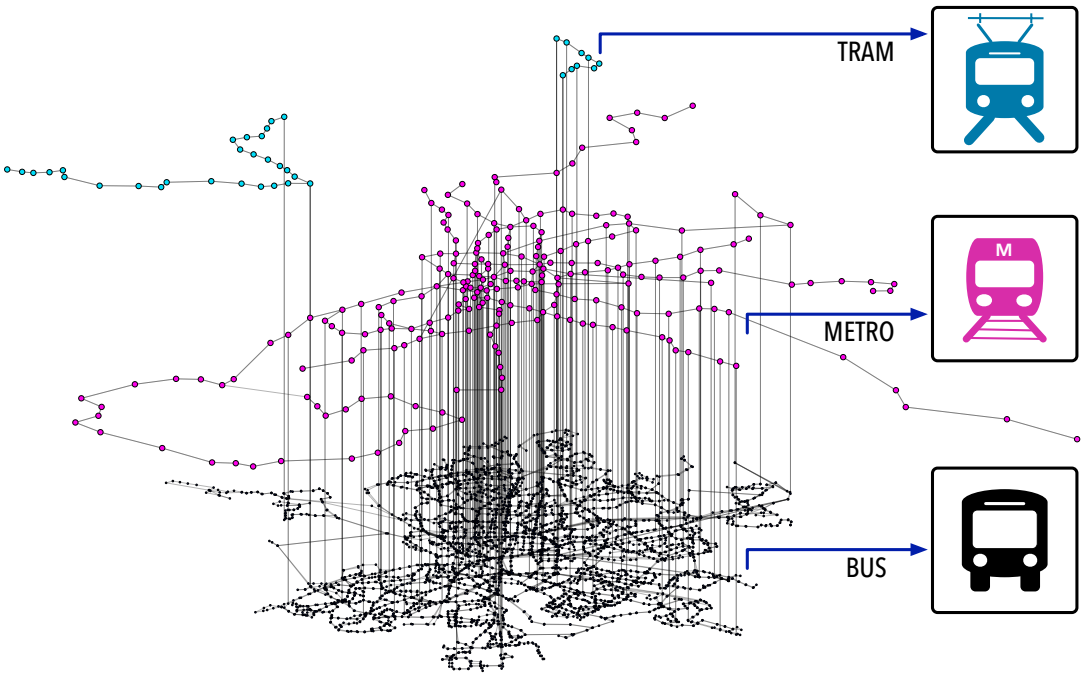

Figure 2.2: Multilayer representation of Madrid's public transportation system. Nodes in the bottom layer (black) represent bus stops, with a link between two nodes if at least one bus line connects them. In the middle layer (pink) nodes are metro stops and links their corresponding metro lines. Lastly, the top layer (blue) is composed by tram stops and their connections. Nodes in different layers are connected if they are within 100 meters.

Another system that can be effectively represented using multilayer networks is the public transportation system of a city. To illustrate this, in figure 2.2 we show the multilayer representation of Madrid's public transportation network. In this case, each transportation mode is encoded in a distinct layer, with nodes representing bus, metro and tram stops. To take into account the possibility of commuting, e.g., taking first the metro and later a bus, a link connecting nodes in different layers is set if they are within a reasonable walking distance. Thus, in this case, nodes do not represent exactly the same entity (although one could argue that they represent the same physical location) and links across layers have a clear physical meaning.

There are two main approaches to encode multilayer networks in a mathematical object: *tensors* and *supra-adjacency matrices* [81]. For simplicity, we will only present the latter approach, as it is straightforward to extend the notions of classical graph theory into multilayer graphs this way.

Suppose that we have two single-layer networks as depicted in figure 2.3A, encoded in their respective adjacency matrices, A_1 and A_2. These matrices contain information about the links that are set inside each layer, the *intra-layer links*. Now, suppose that these two networks are not isolated but are allowed to interact, or represent different parts of the same system. In such scenario we can build a *coupling matrix*, C, containing the links that connect nodes in different layers, the *inter-layer links*. These matrices allow us to define the supra-adjacency matrix as

$$A = \oplus_\alpha A_\alpha + C\,, \tag{2.12}$$

where α runs over the set of layers. This matrix is shown in figure 2.3B, where both layers interact via nodes $1-4$ and $2-5$.

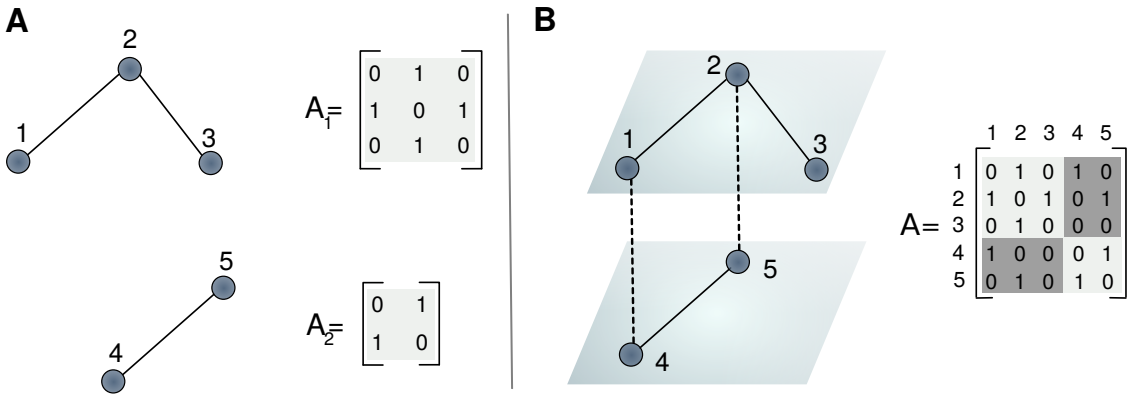

Figure 2.3: Schematic representation of multilayer networks. A) Two independent graphs with their respective adjacency matrices. B) A multilayer network made of both networks with its supra-adjacency matrix.

The same way we defined in monoplex networks the degree of a node as the sum of its links, in multilayer networks the degree of node i in layer α reads

$$k_i^\alpha = \sum_j a_{ij}^\alpha\,. \tag{2.13}$$

Note that this implies that the degree of a node is no longer a scalar but a vector $k_i = (k_i^1, \ldots, k_i^L)$. Nevertheless, we can recover the equivalent quantity of degree in monoplex networks denominated *degree overlap* as

$$o_i = \sum_\alpha k_i^\alpha\,. \tag{2.14}$$

A similar procedure can be followed to extend the notion of strength in the case of weighted multilayer networks.

To conclude this section, let us point out the existence of measures that only exist in multilayer networks, without single-layer counterparts. For instance, the *interdependence* of node i is defined as

$$\lambda_i = \sum_{i\neq j} \frac{\psi_{ij}}{\sigma_{ij}}\,, \tag{2.15}$$

where σ_{ij} is the total number of shortest paths between nodes i and j and ψ_{ij} is the number of shortest paths between those nodes that make use of links in two or more layers. Hence, the interdependence measures how dependent a node is on the multiplex structure in terms of reachability. To understand this property we can have a look again at figure 2.2. In that network, it is clear that to reach most nodes from any tram stop it is required to go through more than one layer, making tram stops quite interdependent. Conversely, it is

possible to reach almost all nodes from a bus stop without having to go through other layers. This example highlights that sometimes individual nodes are not that important, as the layer they are in already determines some of their properties. In this particular example, it is then useful to extend the definition of interdependence from nodes to layers to account for the importance of a given layer in the whole system [82].

2.2 The problem of null models

For some systems, networks naturally provide the skeleton on which dynamical processes can be studied, allowing us to predict the outcome of such processes under different conditions. However, in Jaynes' words [31], *prediction is only one of the functions of statistical mechanics. Equally important is the problem of interpretation; given certain observed behavior of a system, what conclusions can we draw as to the microscopic causes of that behavior?*

To this aim, one of the main focus of network science is to determine the empirical properties of a network that provide the maximum amount of information about the network itself or about the dynamical processes taking place in the system that the network represents [83]. In order to do so, a standard approach is to create a *null model* of the network that will act as a benchmark model. Then, it is possible to identify which properties deviate from what is "expected" in rigorous statistical terms. This, in turn, has several implications. First, it highlights which properties might bear important information about the system dynamics. Similarly, it can be used to detect the hidden mechanisms behind the formation of the network structure, by pointing out the properties that the model was not able to capture. Furthermore, it can be used to elucidate which characteristics of the network are truly important and which ones are just a consequence of lower-order attributes.

Indeed, suppose that we measure two properties of a real network, X and Y. Then, we build a null model of the network using only information about X. If most networks in the ensemble of randomized graphs exhibit property Y, we can state that X explains Y and thus any attempt of analyzing Y without considering X is futile. For instance, many real networks exhibit what is known as the small-world property, meaning that the average length of shortest paths in the network scales as $l \propto \ln(N)$ rather than with the number of nodes [57]. Yet, while in sparse networks like social networks it is a genuine feature of the system, in highly dense networks, such as the interarial cortical network in the primate brain where 66% of links that could exist do exist, it is just a consequence of the huge amount of links present in the network [84].

The previous example shows that sometimes Y can be just a consequence of a particular value of X. However, it is also possible to find properties Y that are always a direct byproduct of X, rendering them redundant. For instance, it has been observed that many ecological networks of mutualistic interactions between animals and plants are nested, that is, that interactions of a given node are a subset of the interactions of all nodes with larger degree. Yet, despite being a widely used metric in the field of ecology, it was recently demonstrated that the degree sequence of the network completely determines the nestedness of the system. Note that to calculate the nestedness of a network it is necessary to know the whole adjacency matrix, while the degree sequence is just the number of non zero entries per row. Hence, the nested pattern is not more informative of the behavior of real systems than their degree sequence alone. In other words, attempting to explain why an ecological system exhibits a given value of nestedness is pointless, as the focus should be on elucidating the mechanisms that give rise to a particular degree sequence [85].

The applications of null models go far beyond the determination of which properties of the network are important. For instance, when not all the information of a system is available, they can be used to infer missing information or to devise adequate sampling

procedures to reduce it [86, 87]. Another interesting and open problem is that network data may contain errors inherent to any experimental measurement. Null models can then be used to estimate the errors in the data and even to correct them [74].

However, as with any hypothesis test, the choice of the null model can directly affect the conclusions. Hence, care must be taken when deciding which one is the best suited for the system under consideration. There are two possible approaches to create null models of networks. The first one consists in building networks based on heuristic rules that are compatible with the specific characteristics of the real system. A classical example is that of gravitational models, widely used in economical and spatial systems way before the birth of network science [88, 89], in which an interaction between two entities depends inversely on their distance and proportionally to their "mass", be it products or population. This method can also be used to shed some light on the mechanisms driving the evolution of networks, rather than just as a tool for generating randomized networks. For instance, the Barabási-Albert model was proposed as a plausible explanation of the emergence of scale-free networks. This model is based on, at each iteration, adding a new node into the network and linking it to m existing nodes, chosen with probability proportional to their degree. This *preferential attachment* naturally leads to networks exhibiting the scale-free property [90].

These types of models, however, can be either too domain specific or require several refinements to give quantitatively accurate predictions [61]. Hence, frequently, a better approach is to identify a set of characteristic properties of the real network and then build a collection or ensemble of networks with those same properties but otherwise maximally random. Despite its conceptual simplicity, this approach has several subtleties that have to be properly handled, something that, as we shall see, is not systematically done. Additionally, we can distinguish two different families of models within this approach: the microcanonical, in which properties are exactly fixed, and the canonical, in which properties are preserved *on average*.

2.2.1 Microcanonical models

Microcanonical approaches are based on generating randomized variants of real networks in such a way that some properties are identical to the empirical ones. However, although their simplicity makes them quite appealing, most of them suffer from problems such as bias, lack of ergodicity or mathematical intractability [91]. Yet, they are widely used, even if not always correctly.

The simplest microcanonical ensemble of graphs that can be built is based on fixing the total number of links, L, and otherwise keeping the graphs completely random. This, as we show in section 2.1.2, defines the ER model. The problem of this model is that it is overly simplistic and real networks often exhibit much richer structures. Thus, a more common approach is to fix the degree of the nodes, also known as *degree-preserving randomization* or *configuration model*. To see why this is a good option, note that L can be obtained by adding up the whole degree sequence. Hence, as discussed previously, if we were able to reproduce the degree distribution of a network by just fixing L, it would not convey any valuable information about the network. However, we showed in section 2.1.2 that the degree distribution in the ER model is Poisson, whereas in most real networks it follows a power law distribution. In other words, the degree sequence often provides much more information about the system than just the number of links.

To build graphs with a fixed degree sequence, two different approaches can be followed: a bottom-up and a top-down approach. The work by Maslov et al. in 2002 on protein and gene networks [92] is often cited as one of the earliest examples of degree-preserving randomization. Another common reference (for instance, in the wikipedia entry on degree-preserving randomization [93]), is the work by Rao et al. in 1996 [94] on generating random $(0, 1)$-matrices with given marginals. Yet, both methods fall into the category of top-down

approaches but, historically, bottom-up approaches were introduced much earlier. In fact, this is one of those methods that have been independently rediscovered several times in different fields, usually with different terminology as well [95]. For instance, in sociology this approach can be traced back to 1938 [96] and in ecology to 1979 [97].

For historical reasons, let us begin with the bottom-up approach. The basic idea is to attach to each node as many "link stubs" as its degree in the original network. Then, pairs of stubs are randomly matched, generating graphs that preserve the original degree sequence. In the context of graph theory stub matching was introduced by Bollobás in 1980 [98]. He called each of the graphs extracted from the ensemble a "configuration", from which the term configuration model emerges, even though nowadays it is used to refer to any model that preserves the degree sequence of the network. Note also that for this approach it is not necessary to have knowledge of the whole adjacency matrix, as only the total addition of its rows and columns are needed. Thus, this procedure is suitable also to create graphs based on a degree sequence sampled from any degree distribution, and not only to randomize existing networks.

Unfortunately, stub matching presents several drawbacks. First, even if the original network is simple (and most real networks are), this method naturally allows both multilinks and self-loops. It is sometimes argued that this is not a problem as their number tends to zero as the size of the graphs grows [74]. However, this is not true in general for scale-free networks, which are the most common ones. Even more, real networks are always finite, and in some fields such as ecology rather small, making this method unsuitable for their randomization. A widely proposed solution is to reject multilinks or self-loops [61, 99], but this, again, has numerous issues. Indeed, when most stubs are already matched, it is possible that the remaining ones cannot be matched, as there might be already links between the nodes with remaining open stubs. Furthermore, this procedure no longer samples uniformly the space of graphs, introducing some biases. In the famous work by Catanzaro et al. on the generation of scale-free networks [100], it was shown that this bias can be solved by imposing the maximum degree in the network to be $k_{\max} \leq N^{1/2}$. Although this might be a suitable approach for creating random networks from scratch, it is clearly not valid for randomizing real networks, in which the maximum degree often exceeds such limit (see figure 2.4A). Several other modifications have been proposed to fix stub matching for simple networks, but there is not a clear solution that always works [95].

Conversely, the top-down approach starts with the full adjacency matrix and randomizes it. In this case, the idea is to take existing links and randomly change which nodes they are attached to. For this reason, this procedure is also known as *rewiring*. The straightforward approach would be to simply randomize the entries of the adjacency matrix, but this procedure would yield the same results as the ER model with fixed number of links [76]. A better idea is, thus, to take two pairs of connected nodes and swap their links, so that the degree sequence of the network is exactly preserved. Howbeit, this procedure is neither free of caveats. For instance, the number of necessary rewires to effectively randomize the network is unknown a priori, spaning several orders of magnitude depending on the network under consideration [102]. Even more important, swaps are usually proposed to be such that given two pairs of connected nodes, $\{a, b\}, \{c, d\}$, the rewiring results in $\{a, d\}, \{b, c\}$ [99]. However, note that the rewiring $\{a, c\}, \{b, d\}$ is perfectly valid too [103], but it is often concealed because if the two randomly selected links are $\{a, b\}, \{a, c\}$, this last swap would introduce a self-loop in the system. Similarly, multilinks can also appear in the graph, unless we strictly forbid them. Yet, if this is naively done biases will be introduced in the sampling procedure, although there are advanced techniques that can be applied to solve these problems in some cases [95, 103].

Yet, it is stricking that even though the biases of the naive link swapping have been known for years [104], this method is still widely used today. One could even argue that it

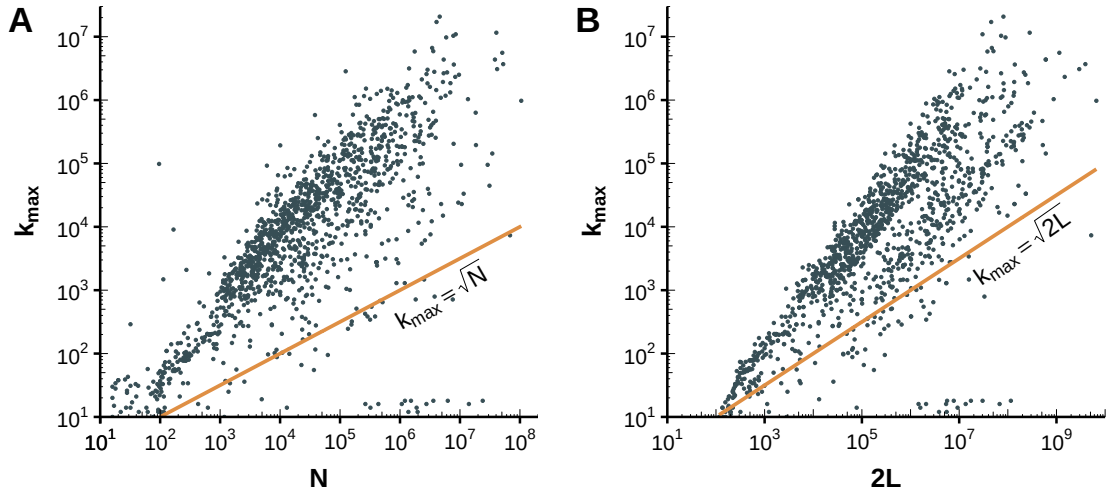

Figure 2.4: Topological properties of 1326 real networks. These networks represent the whole Koblenz Network Collection spanning over 24 different types of systems, from online to infrastructure or trophic networks [101]. In A the maximum degree of each network as a function of its number of nodes is shown. The line represents the structural cutoff $k_{\max} = \sqrt{N}$ of the uncorrelated configuration model [100]. In B the maximum degree is plotted against the total degree of the network (note that $k_{\text{tot}} = 2L$). The line represents the structural cutoff $k_{\max} = \sqrt{2L}$ above which equation (2.18) is not valid.

is actually the most common one, as it is recommended in recent books [76, 99] and some of the most widely used software packages for studying networks do not take into account the necessary corrections in their routines [105, 106]. In particular, it was shown [107] that naive rewiring sampling is only uniform as long as

$$\langle k^2 \rangle k_{\max} / \langle k \rangle^2 \ll N \,. \tag{2.16}$$

But we know that in scale-free networks $\langle k^2 \rangle$ diverges, making this approach ill-suited for strongly heterogeneous networks, the most common ones.

It is clear then, that even if the main ideas of microcanonical models are easy to understand, properly implementing them is not as straightforward as it may seem and multiple subtleties have to be taken into account. Furthermore, so far we have only discussed about binary networks as extending these concepts to weighted graphs is not an easy task either. For instance, one could propose that when swapping links its weight should be attached to it, as if we were directly randomizing the adjacency matrix. Another possible choice would be to only shuffle weights over links, keeping the latter fixed to their nodes. Or even creating stubs with fixed weights and only matching those with equal values, and many more [108]. Obviously, each choice might have its own advantages and drawbacks, but there is no doubt that using such null models should rise more questions than answers about the system.

2.2.2 Canonical models

In canonical models, rather than generating randomized networks, the main objective is to determine mathematical expressions for the expected topological properties as a function of the imposed constraints. Nevertheless, sampling random graphs from canonical ensembles is not only possible but often necessary, as we will see in section 2.5.1.

Focusing on binary graphs [91], since any topological property X is a function of the adjacency matrix A, the goal is to determine the probability of occurrence for each graph, $P(A)$. This allows the computation of the expected value of X as $\sum_A P(A)X(A)$ without

resorting to sampling adjacency matrices. More importantly, in canonical ensembles where the properties that we want to preserve are local, $P(A)$ factorizes to a product over the probability p_{ij} that nodes i and j are connected.

For instance, if we choose $p_{ij} = p \quad \forall i, j$ we obtain the canonical version of the ER model, which in the limit $N \to \infty$ is equivalent to the microcanonical setup in which the number of links is fixed. Note that in the former setting $p_{ij} = p$ is the same as fixing the number of links to be $p\binom{N}{2}$.

If, instead, one wishes to preserve the degree sequence of the network, the most popular choice of p_{ij} is

$$p_{ij} = \frac{k_i k_j}{k_{\text{tot}}} = \frac{k_i k_j}{2L} \tag{2.17}$$

which can be clearly related to the problem of stub matching described in section 2.2.1. Despite its popularity and widespread use, it is important to note - although often disregarded - that this expression is only valid for uncorrelated networks, i.e., as long as the largest degree in the network does not exceed the so-called structural cut-off [109],

$$k_{\text{max}} \leq \sqrt{k_{\text{tot}}} = \sqrt{2L}\,. \tag{2.18}$$

But, as we can see in figure 2.4B, most networks do not fulfill this condition. Hence, if we were to use equation (2.17) to determine if a given property of the network can be explained just by its degree sequence, we would be making a mistake as the own degree sequence already has a property that our model is not capturing. Thus, an expression that correctly captures degree correlations should be used instead. This is not to say that the reasons why correlations or, similarly, large degrees appear in the network are not important. On the contrary, they are, but if they are not added into the null model we will not be able to discern if a given feature of the network is truly meaningful or just a byproduct of the unavoidable correlations. In addition to all of this, equation (2.17) is not a correct probability as one can easily think of toy models in which $p_{ij} > 1$.

The problems induced by highly heterogeneous networks, the most common type of real networks, are clearly challenging most null models, both microcanonical and canonical. In what follows, we will introduce an approach that can solve all these problems and provide a unified approach to sample graph ensembles with local constraints in an unbiased way, the exponential random graph model.

2.3 Exponential random graphs

Suppose that we want to build a simple, binary and undirected graph based only on two macroscopic properties: its number of nodes, N, and its number of links, L. As in the classical problems of statistical mechanics, the number of microstates - graphs - that are compatible with those macroscopic quantities is quite large. Indeed, as we are considering simple graphs, the number of microstates can be obtained from counting the number of ways in which we can select L elements from the set of node pairs, C_2^N. As such,

$$\begin{aligned}\Omega = C_L^{C_2^N} &= \binom{\binom{N}{2}}{L} = \binom{\frac{N(N-1)}{2}}{L} = \frac{\frac{N(N-1)}{2}!}{L!\left(\frac{N(N-1)}{2} - L\right)!} \\ &\approx \exp\left[L \log\left(\frac{N(N-1)}{2L} - 1\right) - \frac{N(N-1)}{2} \log\left(1 - \frac{2L}{N(N-1)}\right)\right],\end{aligned} \tag{2.19}$$

where we have used Stirling's approximation to the factorials.

To illustrate, we can take the example of a seemingly small network, the neural network of the nematode *Caenorhabditis elegans* (or simply *C. elegans*) which was the first - and so far, only - animal to have the whole nervous system completely characterized [110]. This

little worm, of about 1 mm in length, possesses 302 neurons and about 5600 synapses (considering for simplicity both chemical and electrical synapses undirected) as measured by Brenner et al. in 1986 after 15 years of work [111] (the most recent measurements have increased this number slightly above 6000 [112]). Even though small in comparison with the neural system of organisms such as the fruit fly, with over 10^5 neurons, or the human, with over 10^{10} neurons, the total amount of compatible microstates of this system is

$$\Omega = \binom{\binom{302}{2}}{5600} \approx e^{16966} \sim 10^{7368} \,. \tag{2.20}$$

No wonder, then, that this network is still being intensely studied today. To asses even further the importance of this little network, suffice to say that the work by Brenner et al. not only was the starting point of the field of connectomics, but it also created a whole field of research around this organism from which three Nobel Prizes honoring eight scientists have been awarded to date [110].

This example clearly shows that a huge amount of different graphs can give rise to the same macroscopic properties. However, as described in section 2.2, this can actually be a powerful tool to asses what are the main topological characteristics of a network, or which ones are just a consequence of others. The problem is, then, on how to build an ensemble of networks using only partial information of the real one.

Fortunately, this question was already answered by Jaynes in the context of statistical mechanics using the results provided by information theory [31]. Indeed, given a certain quantity G which can take the discrete values $G_i (i = 1, \ldots, n)$ and the expectation value of the function $f(G)$,

$$\langle f(G) \rangle = \sum_{i=1}^{n} p_i f(G_i) \,, \tag{2.21}$$

where p_i is the probability of finding G_i in the ensemble. The only way to set these probabilities without any bias, while agreeing with the given information, is to maximize Shannon's entropy,

$$S = -\sum_{i=1}^{n} p_i \ln p_i \,, \tag{2.22}$$

subject to the appropriate constraints derived from the information. Applied to statistical mechanics, this meant that the old Laplacian principle of insufficient reason according to which, in absence of evidence, the same probability should be assigned to each possible event was not needed anymore (although it is still regarded as one of the basic postulates of statistical mechanics [113, 114]). Instead, the same proposition could be read in positive as maximizing the uncertainty in the probability subject to whatever is known, thus removing the apparent arbitrariness of the principle of insufficient reason. It is trivial to see that if there are not any constraints the maximum entropy is attained for the uniform probability distribution.

Back into the context of networks, following Park and Newman [55], suppose that the information we have is a collection of graph observables X_i, $i = 1, \ldots, r$ from which we have an estimate of their expectation value, $\langle X_i \rangle$, measured over a real network. Let G be a graph in the set of all simple graphs of N nodes, $\mathcal{G}$, and let $P(G)$ be the probability of finding that graph in the ensemble. According to our previous discussion, the most unbiased method to assign values to said probabilities given the available information is to maximize the entropy

$$S = -\sum_{G \in \mathcal{G}} P(G) \ln P(G) \tag{2.23}$$

subject to the constraints

$$\sum_{G \in \mathcal{G}} P(G) X_i(G) = \langle X_i \rangle \tag{2.24}$$

plus the normalization condition

$$\sum_{G\in\mathcal{G}} P(G) = 1\,, \tag{2.25}$$

where $X_i(G)$ denotes the value of the observable X_i in graph G.

To impose these constraints in the maximization process we can introduce the Lagrange multipliers α, θ_i, so that

$$\frac{\partial}{\partial P(G)}\left[S + \alpha\left(1 - \sum_{G\in\mathcal{G}} P(G)\right) + \sum_{i=1}^{r}\theta_i\left(\langle X_i\rangle - \sum_{G\in\mathcal{G}} P(G)X_i(G)\right)\right] = 0 \tag{2.26}$$

for all graphs G. This gives

$$\ln P(G) + 1 + \alpha + \sum_{i=1}^{r}\theta_i X_i(G) = 0\,, \tag{2.27}$$

or equivalently

$$P(G) = e^{-(1+\alpha+\sum_{i=1}^{r}\theta_i X_i(G))}\,. \tag{2.28}$$

Now, inserting this last equation into (2.25) we can define the variable Z as

$$\begin{aligned}\sum_{G\in\mathcal{G}} P(G) = \sum_{G\in\mathcal{G}} e^{-(1+\alpha+\sum_{i=1}^{r}\theta_i X_i(G))} = 1\\ \Rightarrow Z \equiv \sum_{G\in\mathcal{G}} e^{-\sum_{i=1}^{r}\theta_i X_i(G)} = e^{1+\alpha}\,.\end{aligned} \tag{2.29}$$

Hence, the exponential random graph model is completely defined by

$$P(G) = \frac{e^{-H(G)}}{Z}\,, \tag{2.30}$$

where $H(G) = \sum_{i=1}^{r}\theta_i X_i(G)$ is the graph Hamiltonian and $Z = \sum_{G\in\mathcal{G}} e^{-H(G)}$ is the partition function. Note that this expression is equivalent to the probability distribution for the canonical ensemble of statistical physics, which can actually be derived using the same procedure considering an isolated system in thermal equilibrium with the energy as observable [115]. Thus, it represents the graph analog of the Boltzmann distribution. Actually, the classical parametrization of the Boltzmann distribution, i.e., using β instead of θ, gave the name to the β-model, which is the exponential random graph model in the particular case of undirected graphs [116].

Using (2.30) it is then possible to measure the expected value of any graph property Y over the ensemble,

$$\langle Y\rangle = \sum_{G\in\mathcal{G}} P(G)Y(G)\,, \tag{2.31}$$

obtaining the best estimate of the unknown quantity Y given the set of known quantities X_i.

Before going any further, it might be enlightening to see a simple example. Suppose that we only know the average number of links, $\langle m\rangle$, of our network. In that case the Hamiltonian is just

$$H(G) = \theta m(G)\,. \tag{2.32}$$

Let A be the adjacency matrix of graph G with $N\times N$ nodes and elements $a_{ij} = 0, 1$. Then, the number of links is $m = \sum_{i<j} a_{ij}$ and the partition function is

$$\begin{aligned}Z = \sum_{G\in\mathcal{G}} e^{-H(\theta)} = \sum_{\{a_{ij}\}} e^{-\theta\sum_{i<j}a_{ij}} = \sum_{\{a_{ij}\}}\prod_{i<j} e^{-\theta a_{ij}}\\ = \prod_{i<j}\sum_{a_{ij}=0}^{1} e^{-\theta a_{ij}} = \prod_{i<j}(1+e^{-\theta}) = [1+e^{-\theta}]^{\binom{N}{2}}\end{aligned} \tag{2.33}$$

so that

$$P(G) = \frac{e^{-H(\theta)}}{Z} = \frac{e^{-\theta m}}{[1+e^{-\theta}]^{\binom{N}{2}}} = p^m (1-p)^{\binom{N}{2}-m}, \tag{2.34}$$

where we have defined $p \equiv (e^{\theta}+1)^{-1}$.

Recalling section 2.1 we can clearly see that equation (2.34) is just the well known Erdős and Rényi graph, or random graph. Thus, it seems that we have achieved our objectives, building a graph that satisfies our constraints but otherwise completely random, hence its name.

As an example of how to calculate expected values of graphs over this ensemble, we show how to obtain the expected degree, which we know should be equal to $p(N-1)$ (equation (2.7)),

$$\begin{aligned} \langle k \rangle &= \frac{2\langle m \rangle}{N} = \frac{2}{N}\frac{1}{Z}\sum_{G\in\mathcal{G}} m e^{-\theta m} = -\frac{2}{N}\frac{1}{Z}\frac{\partial Z}{\partial \theta} = \frac{2}{N}\binom{N}{2}\frac{1}{e^{\theta}+1} \\ &= \frac{2}{N}\binom{N}{2}p = \frac{2}{N}\frac{N(N-1)}{2}p = (N-1)p\,. \end{aligned} \tag{2.35}$$

2.4 Randomizing real networks

We have seen that equation (2.31) allows us to calculate expected values of graph observables over an ensemble of maximally random graphs restricted to the information we have of the system. However, in some occasions the observables we are interested in might not have an analytic closed-form solution. Or we might be interested in the effect of a given dynamical process on the network and we would like to know if in networks with equivalent macroscopic properties the outcomes would be similar. In those cases, the only thing we can do is to directly sample a large amount of networks from the constructed ensemble and measure the desired topological properties or implement the dynamical process on them.

In order to do so, following Squartini and Garlaschelli [117], we will focus on local topological properties of the networks, i.e., properties determined by moving only one step from a node. As we will see, this will factorize the ensemble probability, allowing us to independently sample each link of the network. In particular, we will focus on two of the most simple types of networks: binary undirected networks, characterized by the degree of their nodes, $k_i = \sum_{j\neq i} a_{ij}$; and weighted undirected networks, characterized by the strength of their nodes, $s_i = \sum_{j\neq i} w_{ij}$. In these cases, the graph probability, in which we will make explicit the dependency with $\vec{\theta} = \theta_i$, $i = 1, \ldots, r$, $P(G|\vec{\theta}) \equiv P(G)$ factorizes as

$$P(G|\vec{\theta}) = \prod_{i<j} P_{ij}(g|\vec{\theta})\,, \tag{2.36}$$

where g represents an element of the adjacency matrix of graph G, either a binary number $a_{ij} = \{0,1\}$ in the case of binary undirected networks or a real number $w_{ij} \in \mathbb{N}$ in the case of weighted undirected networks.

Besides, to obtain the adequate values of the Lagrangian multipliers, $\vec{\theta}$, we can maximize the log-likelihood

$$\mathcal{L}(\vec{\theta}) \equiv \ln P(G^*|\vec{\theta}) = -H(G^*, \vec{\theta}) - \ln Z(\vec{\theta})\,, \tag{2.37}$$

where G^* denotes the particular real network that we want to randomize. It has been shown that in the exponential random graph model the maximization of log-likelihood provides an unbiased method for estimating the values of $\vec{\theta}$ for which the constraints equal the empirical value measured on the real network, $\vec{\theta}^*$ [116].

Hence, the whole procedure for obtaining the maximum entropy ensemble of a network subject to local constraints is:

1. Specify the local constraints $\vec{X}$ and obtain the probability $P(G|\vec{\theta})$ using (2.26).
2. Numerically determine the parameters $\vec{\theta}^*$ by maximizing (2.37).
3. Use $\vec{\theta}^*$ to compute the ensemble average $\langle Y \rangle$ of any desired topological property Y or, alternatively, sample a large number of graphs from the ensemble.

2.4.1 Undirected binary networks

An undirected binary network is completely specified by a binary symmetric adjacency matrix A. Suppose that we have a real network which we want to randomize, A^*, and the quantity we want to preserve, besides the number of nodes N, is its degree sequence, $k_i = \sum_j a_{ij}$. Hence, equation (2.30) reads

$$\begin{aligned} P(A|\vec{\theta}) &= \frac{e^{-\sum_i \theta_i k_i(A)}}{\sum_A e^{-\sum_i \theta_i k_i(A)}} = \frac{e^{-\sum_{ij} \theta_i a_{ij}}}{\sum_A e^{-\sum_{ij} \theta_i a_{ij}}} = \frac{e^{-\sum_{i<j}(\theta_i+\theta_j)a_{ij}}}{\sum_{\{a_{ij}\}} e^{-\sum_{i<j}(\theta_i+\theta_j)a_{ij}}} \\ &= \prod_{i<j} \frac{e^{-(\theta_i+\theta_j)a_{ij}}}{\sum_{\{a_{ij}\}} e^{(-\theta_i-\theta_j)a_{ij}}} = \prod_{i<j} \frac{e^{-(\theta_i+\theta_j)a_{ij}}}{1+e^{-\theta_i-\theta_j}} \equiv \prod_{i<j} P_{ij}(a_{ij}|\vec{\theta})\,. \end{aligned} \tag{2.38}$$

Defining

$$x_i \equiv e^{-\theta_i} \tag{2.39}$$

and

$$p_{ij} \equiv \frac{e^{-\theta_i}e^{-\theta_j}}{1+e^{-\theta_i}e^{-\theta_j}} = \frac{x_i x_j}{1+x_i x_j}\,, \tag{2.40}$$

equation (2.38) can be expressed as

$$\begin{aligned} P(A|\vec{\theta}) &= \prod_{i<j} P_{ij}(a_{ij}|\vec{\theta}) = \prod_{i<j} \frac{\left(\frac{p_{ij}}{1-p_{ij}}\right)^{a_{ij}}}{1+\frac{p_{ij}}{1-p_{ij}}} \\ &= \prod_{i<j} p_{ij}^{a_{ij}}(1-p_{ij})^{1-a_{ij}}\,, \end{aligned} \tag{2.41}$$

which is simply the probability mass function of the Bernoulli distribution.

Given this last result, it is now clear how to sample in an unbiased way graphs from the ensemble. Indeed, as the graph probability is factorized, we can just sample a graph by sequentially running over each pair of nodes and implementing a Bernoulli trial with success probability p_{ij}, as defined in (2.40) [91]. This highlights one of the advantages of this framework over microcanonical methods, the hard core of the computation resides in obtaining the correct values of p_{ij} and is independent of the number of samples we want to extract. Even more, the sampling procedure is guaranteed to be $\mathcal{O}(N^2)$.

The only thing left is to devise a way of obtaining the link probabilities, p_{ij}. As previously discussed, this can be easily done by maximizing the log-likelihood (2.37) which in this particular case can be expressed as

$$\begin{aligned} \mathcal{L}(\vec{x}) &= \ln P(A^*|\vec{x}) = \ln \prod_{i<j} p_{ij}^{a_{ij}}(1-p_{ij})^{1-a_{ij}} = \sum_{i<j} \ln\left[p_{ij}^{a_{ij}}(1-p_{ij}^{1-a_{ij}})\right] \\ &= \sum_{i<j} a_{ij} \ln p_{ij} + \sum_{i<j}(1-a_{ij})\ln(1-p_{ij}) \\ &= \sum_{i<j} a_{ij} \ln \frac{x_i x_j}{1+x_i x_j} + \sum_{i<j}(1-a_{ij}) \ln \frac{1}{1+x_i x_j} \\ &= \sum_{i<j}(a_{ij}+a_{ji}) \ln x_i - \sum_{i<j} \ln(1+x_i x_j) \\ &= \sum_i k_i(A^*) \ln x_i - \sum_{i<j} \ln(1+x_i x_j)\,. \end{aligned} \tag{2.42}$$

As expected, the only quantity from the real network that is needed to obtain the values of x_i is the degree distribution. Besides, given the definition of these values, (2.39), these parameters vary in the region defined by $x_i \geq 0 \, \forall i$. Unfortunately, maximizing $\mathcal{L}(\vec{x})$ does not yield a closed-form expression for the values of x_i as a function of k_i and one would have to numerically solve the problem.

Nevertheless, the procedure for randomizing a real binary, undirected network is now completely clear. First, one should numerically solve equation (2.42) to obtain the parameters x_i. Then, the probability of a link existing between any pair of nodes i and j can be computed using (2.40). With these values the whole ensemble can be characterized using equation (2.41), from which we can then calculate the expected values of any observable using (2.31) or, conversely, sample as many graphs as required by going sequentially over each pair of nodes and performing a Bernoulli trial with success probability p_{ij}.

As a last remark, note that if $x_i x_j \ll 1$ the link probability can be expressed as $p_{ij} \approx x_i x_j$. If we set $x_i = k_i/\sqrt{2L}$, where L is the number of links such that $2L = \sum_i k_i$, we recover the results of the configuration model (equation 2.17). Thus, said model can be regarded as an approximation of the general exponential random graph with fixed degree sequence when $k_i k_j \ll 2L$ (the low heterogeneity regime).

2.4.2 Undirected weighted networks

An undirected weighted network is completely specified by a non-negative symmetric matrix W whose entry w_{ij} represents the weight of the link between nodes i and j, which we will consider to be an integer. Similarly to the previous section, suppose that we want to randomize the real network W^* while preserving its strength sequence, $s_i = \sum_j w_{ij}$. In this case, the probability of finding a graph in the ensemble, equation (2.30), reads

$$\begin{aligned} P(W|\vec{\theta}) &= \frac{e^{-\sum_i \theta_i s_i(W)}}{\sum_W e^{-\sum_i \theta_i s_i(W)}} = \frac{e^{-\sum_{i<j}(\theta_i+\theta_j)w_{ij}}}{\sum_{\{w_{ij}\}} e^{-\sum_{i<j}(\theta_i+\theta_j)w_{ij}}} \\ &= \prod_{i<j} \frac{e^{-(\theta_i+\theta_j)w_{ij}}}{\sum_{\{w_{ij}\}} e^{-(\theta_i+\theta_j)w_{ij}}} = \prod_{i<j} \frac{e^{-(\theta_i+\theta_j)w_{ij}}}{1 + e^{-(\theta_i+\theta_j)} + \ldots + e^{-(\theta_i+\theta_j)\infty}} \\ &= \prod_{i<j} \frac{e^{-(\theta_i+\theta_j)w_{ij}}}{\frac{1}{1-e^{-(\theta_i+\theta_j)}}} = \prod_{i<j} p_{ij}^{w_{ij}}(1-p_{ij}) \,, \end{aligned} \tag{2.43}$$

where, analogously to the previous case, we have defined

$$x_i \equiv e^{-\theta_i} \tag{2.44}$$

and

$$p_{ij} \equiv x_i x_j \,, \tag{2.45}$$

although in this case the partition function is only defined if $\theta_i > 0$ and thus $x_i \in [0, 1)$, ensuring that the probability p_{ij} is correctly defined.

With this formulation the probability of two nodes having weight w_{ij} is then

$$P_{ij}(w_{ij}|\vec{\theta}) = p_{ij}^{w_{ij}}(1-p_{ij}) \tag{2.46}$$

which equals the geometric distribution of a variable with success probability p_{ij} with $w_{ij} \in \{0, 1, 2, \ldots\}$ failures. Thus, graphs can be sampled drawing a link of weight w with geometrical distributed probability $p_{ij}^w(1-p_{ij})$. Note that with this formulation of the geometric distribution the possibility of $w = 0$ - the absence of a link - is included in the distribution. Alternatively, it is possible to follow a similar procedure as in the binary undirected case. To do so, one can connect two nodes with probability p_{ij} according to the Bernoulli distribution and repeat this process until the first failure is encountered [117].

To determine the correct value of p_{ij} for the network W^*, we can simply maximize the log-likelihood (2.37) which in this case reads

$$\begin{aligned}\mathcal{L}(\vec{x}) &= \ln P(W^*|\vec{x}) = \ln \prod_{i<j} p_{ij}^{w_{ij}}(1-p_{ij}) = \sum_{i<j} \ln\left[p_{ij}^{w_{ij}}(1-p_{ij})\right] \\ &= \sum_{i<j} w_{ij} \ln(x_i x_j) + \sum_{i<j} \ln(1-x_i x_j) \\ &= \sum_{i<j} (w_{ij}+w_{ji}) \ln x_i + \sum_{i<j} \ln(1-x_i x_j) \\ &= \sum_i s_i(W^*) \ln x_i + \sum_{i<j} \ln(1-x_i x_j)\,.\end{aligned} \tag{2.47}$$

This equation lacks a closed-form expression for its maxima. Consequently, the values of $\vec{x}$ have to be computed numerically with the constraint $x_i \in [0,1)$.

2.4.3 Fermionic and bosonic graphs

To finish this brief analysis of the exponential random graph model, we can compute two quantities that will highlight the similarities between this model and classical statistical mechanics. The purpose of this analysis is not to claim that graphs behave exactly as physical particles. On the contrary, the mechanisms behind the growth of networks are completely different from the physical principles underlying quantum statistics. Yet, this mathematical resemblance highlights once again how powerful the statistical mechanics formalism is, and also points out again into Jaynes vision of statistical mechanics as a general problem of inference from incomplete information [118].

In the exponential random graph model we can regard node pairs (i,j) as energy levels that can be occupied by links. Defining $\Theta_{ij} \equiv \theta_i + \theta_j$, in the case of undirected binary graphs, the average number of links between i and j is simply given by

$$\begin{aligned}\langle n_{(i,j)} \rangle = \langle a_{ij} \rangle = p_{ij} &= \frac{x_i x_j}{1+x_i x_j} = \frac{e^{-\Theta_{ij}}}{1+e^{-\Theta_{ij}}} \\ &= \frac{1}{e^{\Theta_{ij}}+1}\,,\end{aligned} \tag{2.48}$$

where $\Theta_{ij} \in \mathbb{R}$ as θ_i was defined as a real number. Thus, in the limits where $\Theta_{ij} \to \infty$ we have that the average occupation is 0. Conversely, when $\Theta_{ij} \to -\infty$ the average occupation is 1. In other words, this formulation is equivalent to the Fermi-Dirac statistic of non-interacting fermions [55], which is not surprising as by construction we only allowed at most one link to be in each energy level.

Similarly, in the case of undirected weighted graphs the average number of links in state (i,j) is given by

$$\begin{aligned}\langle n_{(i,j)} \rangle = \langle w_{ij} \rangle = \sum_{w=0}^{\infty} w p_{ij}^w (1-p_{ij}) = \frac{p_{ij}}{1-p_{ij}} = \frac{x_i x_j}{1-x_i x_j} = \frac{e^{-\Theta_{ij}}}{1-e^{-\Theta_{ij}}} \\ = \frac{1}{e^{\Theta_{ij}}-1}\,,\end{aligned} \tag{2.49}$$

where $\Theta_{ij} > 0$ as for weighted graphs $\theta_i > 0$. Thus, in this case, when $\Theta_{ij} \to 0$ the average occupation tends to ∞, whereas if $\Theta_{ij} \to \infty$ the average occupation tends to 0. Hence, equation (2.49) behaves as the Bose-Einstein distribution for bosons. Again, this was expected as in this case we allow any number of links to connect each pair of nodes. In fact, this analogy can go even further as there are dynamical models for network growth that show properties compatible with Bose-Einstein condensation [119].

2.5 Anomalies in transportation networks

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.

("On the Method of Theoretical Physics", Albert Einstein)

This well-known quotation by Einstein - usually paraphrased as, everything should be made as simple as possible, but not simpler - highlights one of the basic problems of network science. Given a set of data, what are the minimum elements we have to take into account into our analysis? In other words, what is the proper (network) model for it?

The simplest network model one can think of is the undirected binary graph. Given a set of elements, a link is established between each pair that interacts somehow, regardless of the type or length of interaction or the characteristics of the elements themselves. This simple model has many advantages: it is often analytically tractable, or at least is easier to work with it than with more complex models; it is easier to analyze as there are less ingredients into play, specially if we are studying complex dynamics; and they can be built with very few data. Yet, there are several cases in which this model is clearly not enough. For instance, as we shall see in section 3.3.3, when one studies disease dynamics on networks, it is often assumed that the contact between two individuals is undirected, but there are several examples of diseases whose spreading is clearly not undirectional, such as HIV which has a male-to-female transmission ratio 2.3 times greater than female-to-male transmission [120]. This symmetry breaking clearly invalidates a model that explicitly assumes symmetrical interactions, such as the undirected network.

In this section we will address the question of whether a model is simpler than it needs to be in the context of transportation networks. Our starting observation is the report of betweenness centrality anomalies in the worldwide air transportation network [121]. Guimerà et al. built the network considering cities as nodes (joining together airports belonging to the same city) and establishing a link between them if there was at least one direct flight connection. Then, they studied the betweenness centrality of the nodes and found that cities with large number of connections were not necessarily the ones with largest betweenness. Conversely, they observed that cities with few connections could have high values of betweenness. This contrasts sharply with the fact that the betweenness centrality of a node is directly proportional to its degree in synthetic scale-free networks [122] as well as in several real networks [123]. The emergence of these anomalies has been attributed to the multi-community structure of the network and to spatial constraints [121, 124, 125].

However, given our previous discussion, one may wonder whether the anomalies are truly there or if they are just a consequence of using a network model that is too simple. Indeed, using a binary representation implies that two cities connected with one flight per year are as close as two cities with several hundreds of weekly flights. At first glance, this seems a too far-fetched assumption, specially if we are interested in studying city centrality. Thus, we propose that an undirected weighted network might be more suited for this specific problem. This, as we will see, will make the anomalies disappear.

2.5.1 Null models for undirected weighted networks

As discussed in section 2.2, the most common approach to determine if the properties of a network are out of the ordinary is to reshuffle the connections of the network to build a null model, extract the average properties of the latter and compare them. In addition to the caveats of this procedure that have been already described, there is another obvious issue: it is not clear how to extend this procedure to weighted networks. Indeed, one possibility would be to extract the weight distribution of the links, reshuffle the network and then randomly assign weights to the links according to said distribution. Another

option could be to preserve the total strength of a node and then share it evenly across the reshuffled links, or according to some distribution. Or even attaching weights to their links and then reshuffle them preserving their weights. In any case, it is clear that whichever procedure we choose, there will be several implicit assumptions that will reduce the universality of the results. Fortunately, there is a better solution: exponential random graphs.

Initially, as we are working with a weighted network, one might be inclined to use the formalism we presented in section 2.4.2 that preserves the weights of the nodes. However, it has been shown that the strength sequence of a network is often less informative than its degree sequence [126]. In particular, synthetic networks built preserving the strength sequence tend to be much denser that their real counterparts. For this reason, we propose that our null model should be an exponential random graph preserving both the degree and strength sequences.

Following [126], suppose that our real network is described by the symmetric matrix W^* of size $N \times N$. First, we want to obtain the probability of finding any compatible graph in the ensemble, $P(W)$, imposing as constraints the degree, $k_i(W) = \sum_j a_{ij} = \sum_j 1 - \delta(w_{ij})$, and strength, $s_i(W) = \sum_j w_{ij}$, sequences, the Hamiltonian of the graph reads

$$H(W|\vec{\theta},\vec{\sigma}) = \sum_i \theta_i k_i + \sum_i \sigma_i s_i \tag{2.50}$$

and thus the probability of finding any graph W in the ensemble is

$$\begin{aligned} P(W|\vec{\theta},\vec{\sigma}) &= \frac{e^{-\sum_i \theta_i k_i - \sum_i \sigma_i s_i}}{\sum_W e^{-\sum_i \theta_i k_i - \sum_i \sigma_i s_i}} = \frac{e^{-\sum_{i<j}(\theta_i+\theta_j)a_{ij} - \sum_{i<j}(\sigma_i+\sigma_j)w_{ij}}}{\sum_{\{w_{ij}\}} e^{-\sum_{i<j}(\theta_i+\theta_j)a_{ij}) - \sum_{i<j}(\sigma_i+\sigma_j)w_{ij}}} \\ &= \prod_{i<j} \frac{e^{-\theta_i a_{ij}} e^{-\theta_j a_{ij}} e^{-\sigma_i w_{ij}} e^{-\sigma_j w_{ij}}}{1 + e^{-\theta_i-\theta_j} \sum_{w_{ij}=1}^{\infty} e^{-(\sigma_i+\sigma_j)w_{ij}}}, \quad \text{defining} \left[\begin{array}{l} x_i \equiv e^{-\theta_i} \\ y_i \equiv e^{-\sigma_i} \end{array} \right], \\ &= \prod_{i<j} \frac{(x_i x_j)^{a_{ij}} (y_i y_j)^{w_{ij}} (1 - y_i y_j)}{1 - y_i y_j + x_i x_j y_i y_j} = \prod_{i<j} P_{ij}(w_{ij}|\vec{\theta},\vec{\sigma}) \,. \end{aligned} \tag{2.51}$$

To obtain the appropriate values of $\vec{\theta}$ and $\vec{\sigma}$ we can numerically maximize the log-likelihood function,

$$\begin{aligned} \mathcal{L}(\vec{\theta},\vec{\sigma}) &= \ln P(W^*|\vec{\theta},\vec{\sigma}) = \ln \prod_{i<j} \frac{(x_i x_j)^{a_{ij}} (y_i y_j)^{w_{ij}} (1 - y_i y_j)}{1 - y_i y_j + x_i x_j y_i y_j} \\ &= \sum_i \left[k_i(W^*) \ln x_i + s_i(W^*) \ln y_i \right] + \sum_{i<j} \ln \left(\frac{1 - y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j} \right) , \end{aligned} \tag{2.52}$$

where $x_i \geq 0 \, \forall i$ and $y_i \in [0, 1)$.

Lastly, to sample networks from this distribution we must divide the process into two parts. First, note that the probability of not having a link, $w_{ij} = 0$, is

$$P_{ij}(w_{ij} = 0|\vec{\theta},\vec{\sigma}) = \frac{1 - y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j} \equiv 1 - p_{ij} \,. \tag{2.53}$$

Hence, we can perform a Bernoulli trial with probability

$$p_{ij} = \frac{x_i x_j y_i y_j}{1 - y_i y_j + x_i x_j y_i y_j} \tag{2.54}$$

to establish a link. Then, if it is successful, note that for $w_{ij} > 0$ the probability reads

$$P_{ij}(w_{ij} > 0|\vec{\theta},\vec{\sigma}) = p_{ij} (y_i y_j)^{w_{ij}-1} (1 - y_i y_j) \,, \tag{2.55}$$

which is the success probability, p_{ij}, times a geometric distribution with parameter $y_i y_j$. Thus, we can simply extract the weight associated to the node from said distribution. Note that as one trial was already successful, it is necessary to add 1 to the number extracted from the distribution.

Iterating this process (Bernoulli trial plus weight from geometric distribution) we can extract unbiased graphs from the ensemble systematically. In this study this will be of outermost importance as there is not an analytical expression to calculate the betweenness of a node in a graph from its adjacency matrix. Thus, to test if these two ingredients (degree and strength) are enough to explain the anomalies of our network, we will have to sample a large amount of graphs from the ensemble, compute the betweenness of their nodes and lastly compare their distribution with the one of the real network.

2.5.2 The worldwide air transportation network

We will begin our analysis using the same network as in the paper that inspired this work [121], the worldwide air transportation network. Howbeit, to facilitate the interpretation of the results, we will use countries as nodes, rather than cities. This will dramatically reduce the number of nodes, making the network more manageable. Nonetheless, in section 2.5.3 we will analyze the whole network as it was presented in the original paper.

The geographical location of the airports, the city and country they belong to and the routes connecting them were obtained from the Open Flights database [127]. After aggregation, we end up with a network of 224 nodes (countries) and 2,903 undirected links (unique routes). This data, however, does not include any information about the number of flights, which we need to establish the weight of the routes. Hence, we collected data from an online flight tracking website [128] in a period between May 17, 2018 and May 22, 2018.

Our first step is to reproduce the results of Guimerà et al. but, for consistency, using the undirected exponential random graphs presented in section 2.4.1. This is an important step for three reasons: first, their data was collected in the period from November 1, 2000 to October 21, 2001, 18 years before ours; second, because to randomize the network they preserve exactly the degree sequence; and third, because we have used countries rather than cities as nodes. The results are shown in figure 2.5.

In spite of all the methodological differences between our work and Guimerà et al. we do find a similar pattern of anomalies. For instance, Christmas Island is one of the countries with highest betweenness even though is the one with the lowest degree. Conversely, Montenegro, with degree 11, has one of the lowest values of betweenness. This clearly contradicts the results obtained for the graphs sampled from the ensemble built preserving, on average, the degree sequence of the network (shaded gray area). Indeed, in random networks the betweenness centrality tends to be proportional to the degree of the node.

Even more, using countries as nodes instead of cities reveals a clear pattern. We can see that most countries with betweenness below the expectation are European. On the other hand, most nodes laying above the expectation are islands. This can clearly be seen in figure 2.5B, where we show a representation of the network embedded in space. As we can see, Europe is much more densely connected that the rest of the world, due to having a large population shared in several countries. Thus, there are lot of connections between them that increase their degree but do not make them more central globally. The opposite effect is observed in islands and large countries, which act as bridges connecting either large territories or hardly accessible ones, increasing the number of shortest paths that go through them.

These observations also agree with our hypothesis. We expect that if weights - number of flights - are included in the network, the centrality of small, fairly disconnected islands will be reduced while the highly active airports in Europe will increase theirs. Thus, we will now add the number of flights between any two routes as weight. Note, however, that the calculation of shortest paths (the key ingredient of the betweenness centrality) tries to

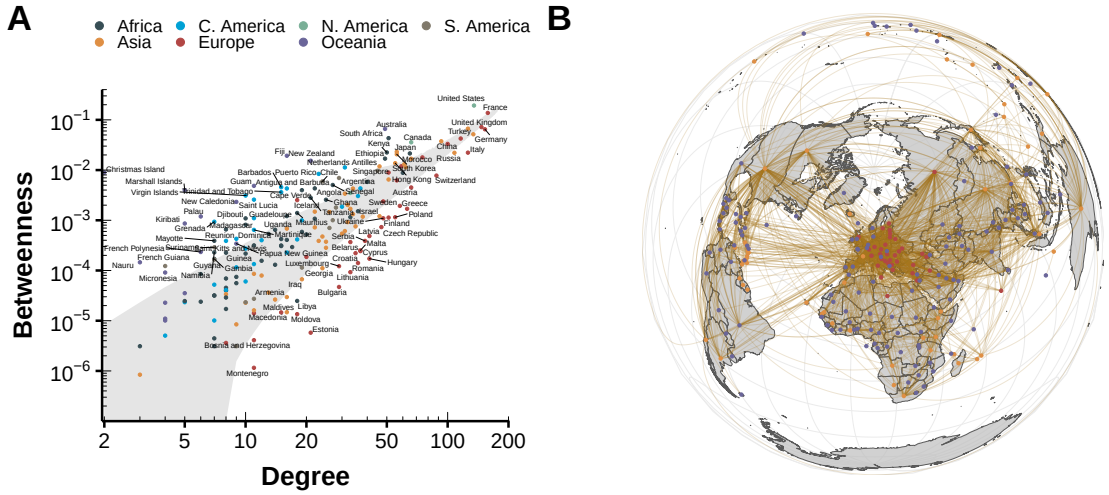

Figure 2.5: Analysis of the worldwide air transportation network with countries as nodes. In panel A we show the betweenness of each country in the network as a function of its degree. The shaded gray area represents where are 95% of the nodes belonging to 10^4 random graphs sampled from the ensemble. In other words, countries outside that region can be considered abnormal. Labels have been attached only to said countries. In panel B we show the geographical projection of the network, with those countries with betweenness higher than expected in orange and with those countries with lower values than expected in red.

minimize the total weight of a path. Hence, a larger weight would make the connection less desirable, which is the opposite of what we want as a larger number of flights makes it easier to use that connection. For this reason, we will initially set the number of flights as weight, then we will randomize the network using the method presented in section 2.5.1 and, lastly, to compute the betweenness we will set the weights to be the inverse of the number of flights, i.e. $w'_{ij} = w_{ij}^{-1}$. The results are shown in figure 2.6.

At first glance, we observe that the number of nodes in the betweenness plot is much smaller than before (figure 2.6A). The reason is that there are several nodes, both in the real network and in the randomized graphs, that have 0 betweenness. In other terms, there are not any shortest paths going through them. This was somehow expected given that in figure 2.5A there were several small countries which clearly do not have as many daily flights as the biggest hubs in the world, for instance the islands in Oceania. Nonetheless, now most islands are compatible with the randomized graphs, as well as most European countries.

To facilitate the comparison of both approaches, in figure 2.6B we show the percentage of nodes that are not compatible with 95% of the randomized graphs. In the case of the undirected network, we find over 40% of nodes out of the area covered by the randomized graphs. Contrariwise, when weights are added to the network, the amount of anomalous nodes goes below the 5% mark. Thus, it is clear that the anomaly that was observed in the undirected case is solved once one considers weights. We will not try to answer why the weights are distributed the way they are, as it is probably a mixture of geographical, economical and political reasons, out of the scope of this analysis. Nevertheless, this result partially answers our initial question. Sometimes anomalies can be a byproduct of using a model that is simpler than it should be. To increase the validity of this statement, in the following section we will repeat the analysis on other types of transportation networks.

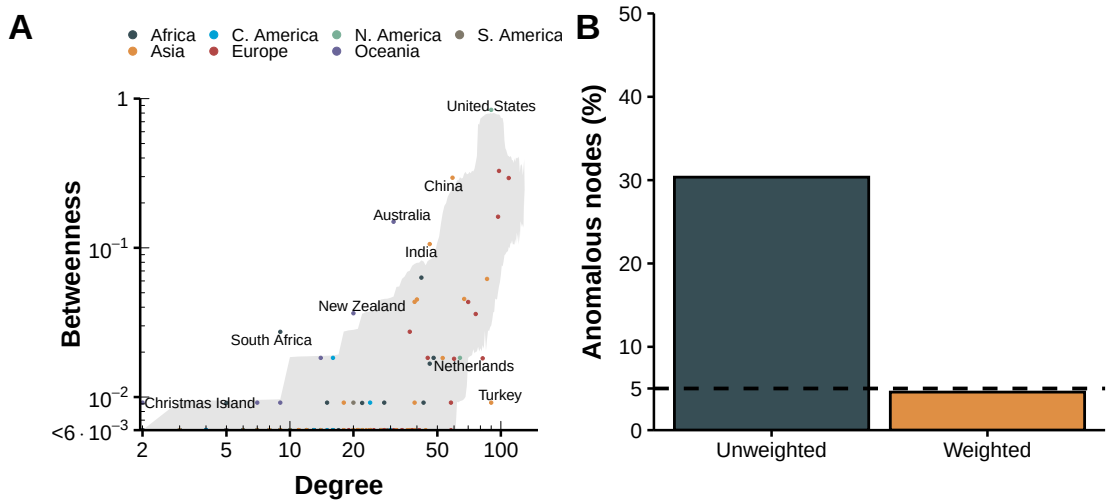

Figure 2.6: Analysis of the weighted worldwide air transportation network with countries as nodes. In panel A we show the betweenness of each country in the network as a function of its degree. The shaded gray area represents where are 95% of the nodes belonging to 10^4 random graphs sampled from the ensemble. Labels have been attached only to countries outside the expectation. In panel B we show the fraction of nodes that can be considered anomalous in the undirected network (left) and in the weighted network (right). The dashed line indicates the 5% threshold.

2.5.3 Other transportation networks

This analysis can be extended to other transportation networks. In particular, we have obtained the inter-city bus transportation networks of Brazil, Great Britain and Spain. Besides, we will now consider the full air transportation network, with cities as nodes, as in the work by Guimerà et al. [121]. The networks are represented in figure 2.7.

Data from the Brazilian inter-city routes was obtained from the Brazilian National Land Transportation Agency (ANTT) [129]. The data corresponds to the period between January 2005 and December 2014 with a monthly resolution and includes more than 19,000 unique routes connecting 1,786 cities. The geographical location of the nodes was determined using data from the Brazilian Institute of Geography and Statistics (IBGE) [130].

Similarly, data from the British inter-city routes was contained in the National Public Transport Data Repository (NPTDR) maintained by the Department of Transport. The data corresponds to the period between October 4, 2010 to October 10, 2010, with an hourly resolution. This dataset was complemented with the National Coach Services Data (NCSD) distributed also by the Department of Transport [131]. The total number of nodes is 279 nodes with almost 4,000 unique routes.

For the case of Spain, however, there is no public repository containing the data in a suitable format to be used. Thus, we had to scrap the data from a website maintained by the Spanish Ministry of Development that offers information of all the bus connections between municipalities in Spain except for the province of Girona [132]. The data corresponds to the period between January 1, 2017 and December 31, 2017. The total number of nodes is 1,435 with over 20,000 unique routes. As nodes in this network are municipalities instead of cities, we aggregated in all networks their bus stops into their corresponding municipalities.

Next, we repeat the analysis outlined in section 2.5.2 on these 4 networks. To sum up, we first build the unweighted version of the networks and randomize them using the formalism of undirected binary graphs presented in section 2.4.1. We then compare the betweenness of the real nodes to their randomized counterparts. If their betweenness is

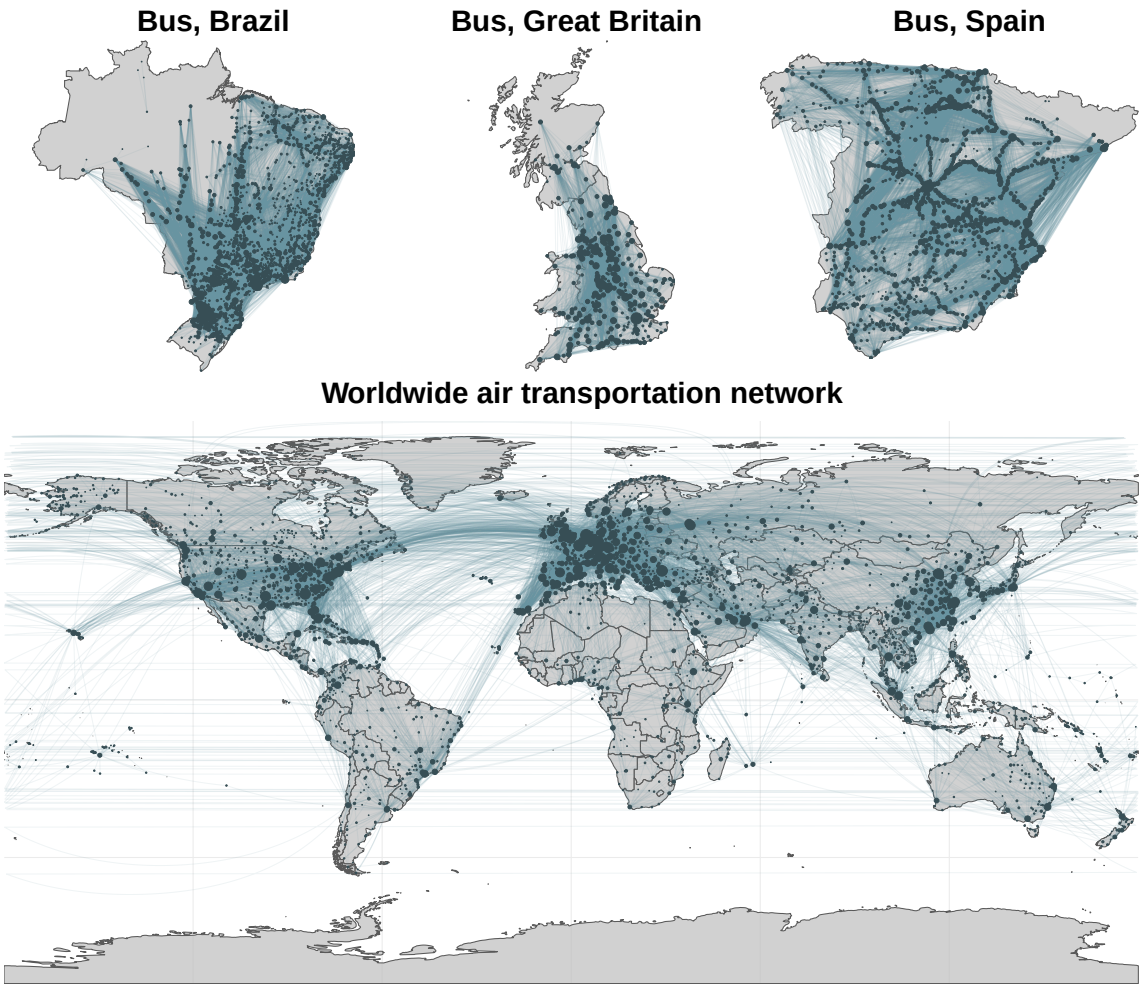

Figure 2.7: Spatial representation of transportation networks. The four networks under consideration are represented embedded in their corresponding spaces. Nodes are located in the center of their administrative region in the case of bus networks and in the coordinates of the city they belong to in the case of the air transportation network. The size of the nodes is proportional to their degree.

higher than their randomized versions in 97.5% of the samples, or if it is lower than in 2.5% of the samples, we flag them as abnormal nodes. We measure the amount of nodes that can be regarded as abnormal in all networks, and then repeat the whole process with the weighted versions of the networks, using the appropriate null model (section 2.5.1). The results are depicted in figure 2.8.

Several observations are in order. First, we not only find the expected anomalies in the worldwide air transportation network, but also in the Brazilian and Spanish bus networks. In the case of Great Britain we do not observe anomalies, although this network has some special characteristics that might explain this fact. Indeed, as we can see in figure 2.7 large cities are situated in the middle, with few small cities in the north and some more in the south. This centralization might explain why we do not see small nodes with large values of betweenness. Besides, municipalities are larger than in other countries and as a result the number of nodes in this network is one order of magnitude below the Spanish and Brazilian networks. This reduces the amount of unique routes and tends to centralize them, preventing routes in isolated areas that would increase the centrality of small nodes (which would actually be even smaller if we had not aggregated them). In any case, regardless of the specific reasons for this observation, it seems that for the particular case of the Great

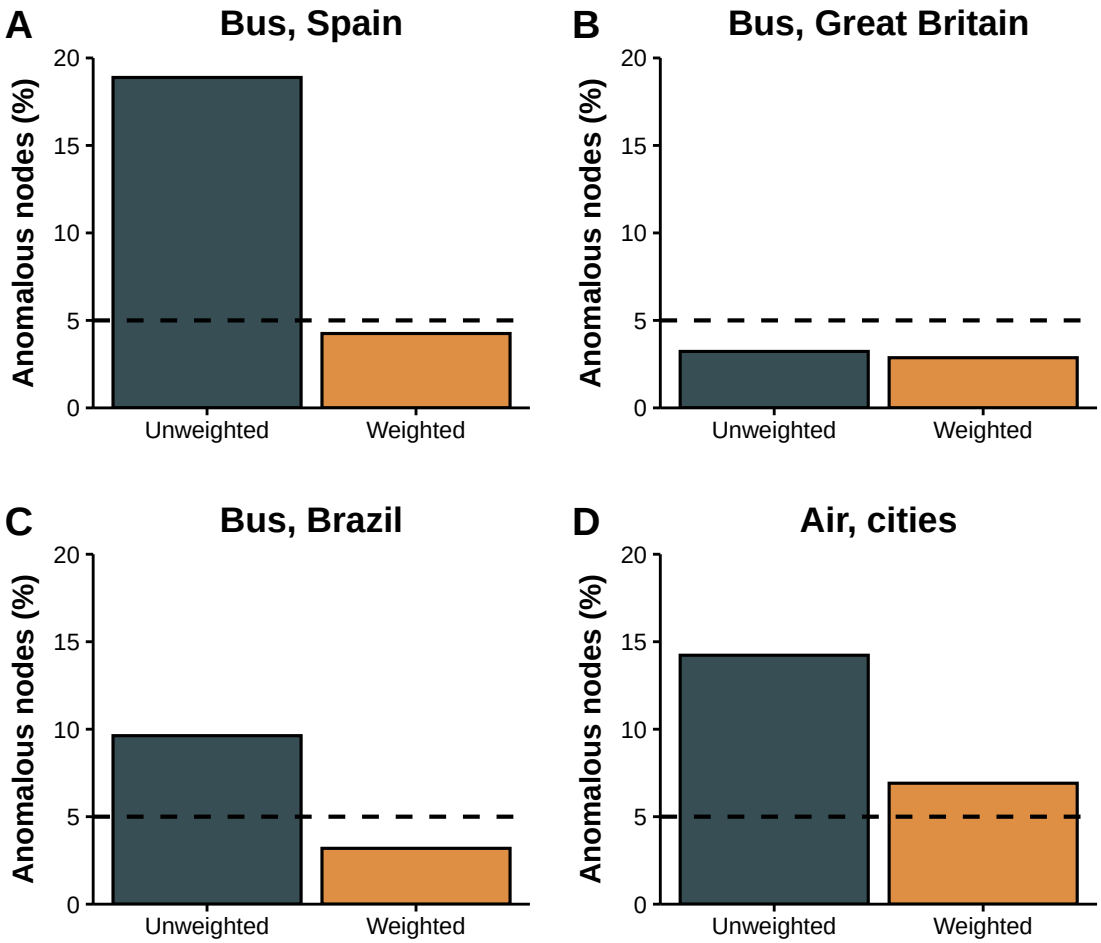

Figure 2.8: Fraction of anomalous nodes in transportation networks. Percentage of anomalous nodes in each network for the unweighted and weighted versions of the networks. The anomaly is verified for all unweighted networks, except for Great Britain. In contrast, in the weighted version the fraction of anomalous nodes is of the order of the false discovery rate, i.e., approximately 5%.

Britain network it is enough to consider only the degree of the nodes to avoid anomalies.

Nevertheless, in the weighted versions of the networks we can see that systematically the amount of anomalous nodes is reduced in all networks, even in the one from Great Britain (although both the weighted and unweighted amounts are below the 5% line and thus we cannot state that the reduction is significant). This result suggests that the existence of centrality anomalies in transportation networks might be a consequence of neglected, but functionally crucial, information about the system.

2.5.4 Conclusions

The exponential random graph framework allows us to build very general null models in which only the information we provide is fixed, the rest being maximally random. This is specially important when one wants to analyze if the characteristics of a real network are different from what would be expected from alike graphs. Indeed, if all the assumptions implicit in a null model are not under control, it is possible to find anomalies that might be just a byproduct of the null model itself, rather than coming from the real system.

In this particular case, we have seen that when one only gathers information about the degree of the nodes in a transportation network, their centrality does not behave as

expected. However, by adding a small piece of information - the weight of the links - the network starts to behave as its random counterparts. Hence, this work highlights the fact that when some anomalies are found in a network, care must be taken in order to determine if those anomalies are really important or just a consequence of not using the proper amount of information.

Even more, we propose that this technique might be useful to determine how much data of a system is needed, or how complex its network model has to be. In the particular case of transportation networks the use of weights is rather straightforward. However, in systems in which the choice might not be so obvious, hunting for anomalies and looking for network characteristics that can explain them, might be the key to determine the most important elements of the system.

2.6 Generating data-driven contact networks

What about the children?!
Won't somebody please think of the children!?
Helen Lovejoy

Contact networks are graphs whose nodes represent individuals and their links represent some kind of close interaction between them. This kind of networks are of particular interest in the field of epidemiology, as there are several diseases which can only be transmitted from person to person, such as influenza, tuberculosis or HIV. Thus, the contact patterns of the population can provide the necessary information to implement preventive measures [133]. For instance, by analyzing who are the most central nodes in the network, it is possible to devise efficient vaccination strategies [134].

Yet, even for small populations, obtaining the whole contact network of the population is really hard. A prominent example can be found in the case of sexually transmitted diseases. It is clear that in those particular diseases it should be possible, in general, to trace back the whole chain of infections and, ideally, reconstruct the contact network [135]. However, this process is full of biases, such as individuals not being willing to provide detailed information about their sexual partners [136, 137]. In practice, it is only possible to trace back the whole chain of infections for very small epidemic outbreaks [138].

Hence, to study large scale epidemics one usually resorts to building synthetic contact networks. In some cases, they can be based on some ad hoc assumptions specific for the disease under study [139, 140]. In others, a small portion of the population is thoroughly investigated and then its characteristics are extended to the whole population [141]. Another possibility is to use aggregated statistics about the population such as household size or age distributions to build them [142]. The framework that we will develop in the following pages will follow the latter.

One of the key elements defining contact networks is the social structure of the population. It has been observed that individuals have very different mixing patterns depending on their age, and even that they vary greatly from country to country [143]. The usual approach to implement this in epidemic spreading models is to consider that the whole population is in contact with each other (resembling a fully connected graph) and weighting the transmission probability between two individuals according to their respective ages [144, 145]. This, however, completely neglects the network structure of the population.

In this work we propose a methodology to build contact networks based on the exponential random graph model using socio-demographic data. In particular, we will take into account the demographic structure of the population as well as their mixing patterns. As previously mentioned, these patterns are highly dependent on the age of the individual. Thus, age will be the key ingredient in this model. Besides, we will arrange individuals

with similar age in layers, building a multilayer network. As we shall see in section 3.4 this will facilitate the study of disease spreading on these networks. However, in this section we will only focus on the methodology used to build said networks.

2.6.1 Theoretical framework

Demographic studies usually classify individuals into age groups, also known as age brackets, instead of using their exact value. Hence, mixing patterns are often given in the form of number of contacts between individuals in age group X with individuals in age group Y. For this reason, we will consider that the population is divided into L layers, one for each age group. The number of nodes in each layer will be set according to the demographic structure of the population under study. We will denote by n_α the set of individuals contained in layer α or, equivalently, with age contained in age bracket α (henceforth, individual with age α). Thus, the total number of nodes will be $N = \sum_{\alpha=1}^{L} |n_\alpha|$.

Besides the demographic structure of the population, we will suppose that the only information we have is the number of (undirected) links of individual i in layer α to individuals in layer β. As a consequence, the quantity that we want to preserve is the layer-to-layer degree, $k_{i,\alpha}^{\beta} = \sum_{j \in n_\beta} a_{ij}$ (the procedure to extract this information from real data will be described in section 2.6.2). An schematic representation of the system is depicted in figure 2.9.

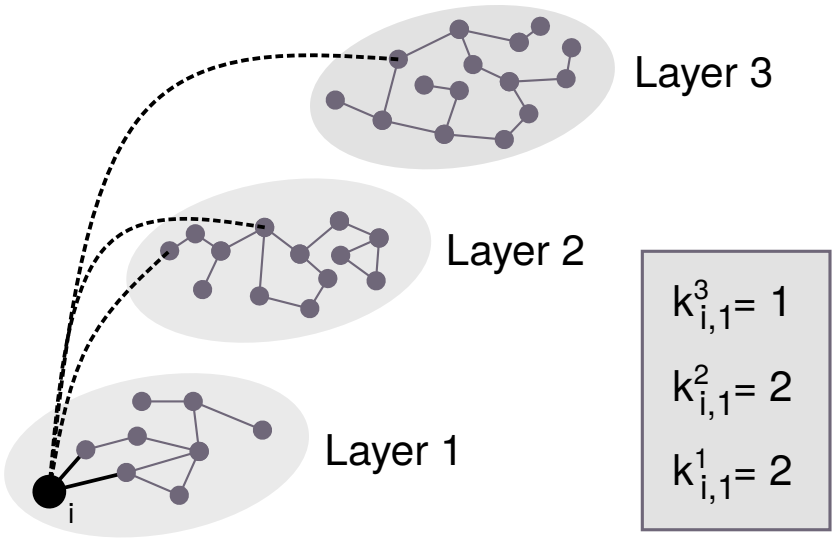

Figure 2.9: Schematic of a multilayer network with age-dependent contacts. Each layer represents one age bracket $[a_{\text{lmin}}, a_{\text{lmax}})$, so that individuals set in layer l are between a_{lmin} and a_{lmax} years old. Links can be inside layers, representing contacts with individuals within the same age bracket, and across layers, representing contacts with individuals of different age. For instance, node i has two links in layer 1, two links to layer 2 and one link in layer 3 so that $\{k_{i,\alpha}^{\beta}\} = \{k_{i,1}^{\beta}\} = \{2, 2, 1\}$.

Under these assumptions, the Hamiltonian of the graph reads

$$H(G, \theta) = \sum_{\alpha=1}^{L} \sum_{\beta=1}^{L} \sum_{i \in n_\alpha} \theta_{i,\alpha}^{\beta} k_{i,\alpha}^{\beta} \,, \tag{2.56}$$

which can be rewritten in a more useful way after some algebra,

$$\begin{aligned} H(G,\theta) &= \sum_{\alpha=1}^{L}\sum_{\beta=1}^{L}\sum_{i\in n_\alpha}\theta_{i,\alpha}^{\beta}\sum_{j\in n_\beta}a_{ij} = \sum_{\alpha=1}^{L}\sum_{i\in n_\alpha}\sum_{j=1}^{N}\theta_{i,\alpha}^{\beta}a_{ij} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\theta_{i,\alpha}^{\beta}a_{ij} = \sum_{i<j}(\theta_{i,\alpha}^{\beta}+\theta_{j,\beta}^{\alpha})a_{ij}\,. \end{aligned} \tag{2.57}$$

In the derivation of equation (2.57) we have used the fact that node i only belongs to layer α. This is not the case in general multilayer networks, where one node can be present in more than one layer. However, in this particular case, as an individual can only have one age, it can only belong to one layer. Note that this also implies that this problem can be mapped into the one of building single layer exponential random graphs with fixed community structure and degree sequence [146].

Now, the probability of finding any graph G in the ensemble will be

$$\begin{aligned} P(G|\theta) &= \frac{e^{-\sum_{i<j}(\theta_{i,\alpha}^{\beta}+\theta_{j,\beta}^{\alpha})a_{ij}}}{\sum_G e^{-\sum_{i<j}(\theta_{i,\alpha}^{\beta}+\theta_{j,\beta}^{\alpha})a_{ij}}} = \prod_{i<j}\frac{\left(e^{-\theta_{i,\alpha}^{\beta}}e^{-\theta_{j,\beta}^{\alpha}}\right)^{a_{ij}}}{1+e^{-\theta_{i,\alpha}^{\beta}}e^{-\theta_{j,\beta}^{\alpha}}} \\ &= \prod_{i<j}\frac{(x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha})^{a_{ij}}}{1+x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}} = \prod_{i<j}p_{ij}^{a_{ij}}(1-p_{ij})^{1-a_{ij}}\,, \end{aligned} \tag{2.58}$$

where, similarly to previous sections, we have defined the auxiliary variable $x_{i,\alpha}^{\beta}\equiv e^{-\theta_{i,\alpha}^{\beta}}$ and the probability of having a link between node i and j as

$$p_{ij}\equiv\frac{x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}}{1+x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}}\,. \tag{2.59}$$

Note that the shape of equation (2.58) is equivalent to the one of undirected binary networks, section 2.4.1. Hence, to sample networks from this ensemble we can simply perform sequential Bernoulli trials on each pair of nodes.

Lastly, to obtain the value of $x_{i,\alpha}^{\beta}$ we only need to maximize the log-likelihood which in this case reads

$$\begin{aligned} \mathcal{L}(\theta) &= \ln\prod_{i<j}p_{ij}^{a_{ij}}(1-p_{ij})^{1-a_{ij}} \\ &= \sum_{i=1}^{N}\sum_{j=i+1}^{N}a_{ij}\ln\frac{x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}}{1+x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}} + (1-a_{ij})\ln\frac{1}{1+x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}} \\ &= \sum_{i=1}^{N}\sum_{j=i+1}^{N}a_{ij}\ln\left(x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}\right) - \ln\left(1+x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}\right) \\ &= \sum_{i=1}^{N}\sum_{\beta=1}^{L}k_{i,\alpha}^{\beta}\ln x_{i,\alpha}^{\beta} - \sum_{i=1}^{N}\sum_{j=i+1}^{N}\ln\left(1+x_{i,\alpha}^{\beta}x_{j,\beta}^{\alpha}\right)\,. \end{aligned} \tag{2.60}$$

Thus, the process is analogous to the one we have seen in sections 2.4.1, 2.4.2 and 2.5.1. Indeed, once the constraints have been established, we need to maximize the log-likelihood to obtain the appropriate values of $\theta_{i,\alpha}^{\beta}$. Then, we can get the probability of two nodes being connected from equation (2.59). Lastly, we can perform measurements over the whole ensemble using the full $P(G|\theta)$ or sample networks using independent Bernoulli trials on each pair of nodes. In the following section we will describe how we can obtain the values of the constraints, $k_{i,\alpha}^{\beta}$, from real data.

2.6.2 Data description

The key ingredient of the proposed methodology is the number of links a node i has within its layer, α, and to each of the other layers, β, denoted as $k_{i,\alpha}^{\beta}$. In order to generate realistic distributions of this quantity, we would need the probability distribution of a node with age a to contact k individuals of age a', $P(k, a, a')$. Fortunately, there are some empirical studies that provide enough information to do so, such as the POLYMOD study [143].

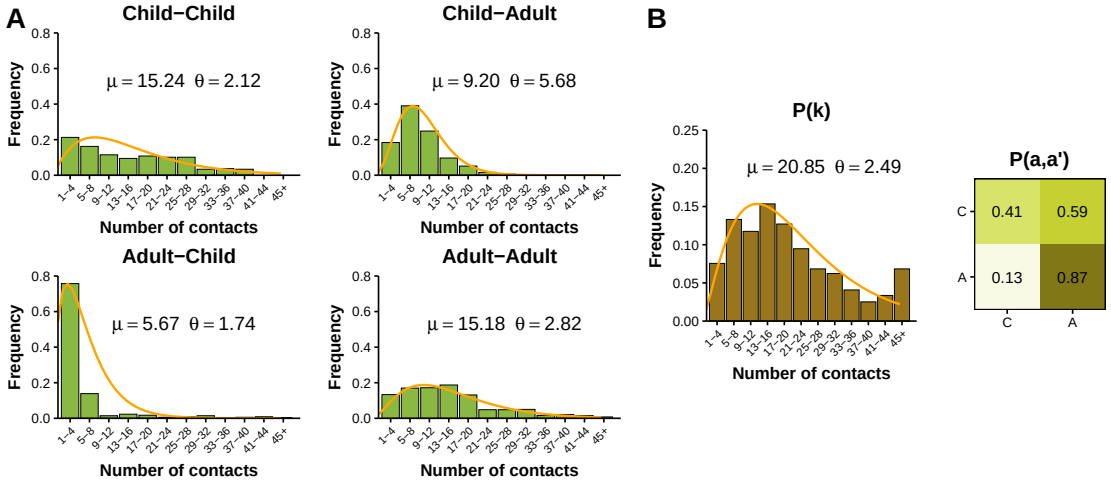

Figure 2.10: Mixing patterns of the Italian population. The population has been divided into two groups: children (individuals aged 18 or less) and adults (individuals aged 19 or more). In A the distribution of the number of contacts between the different groups is shown. The fitted curve corresponds to a negative binomial distribution parameterized with mean μ and shape θ. In B the total number of contacts, regardless of the age of the individual, is shown. As in A, the fitted curve corresponds to a negative binomial distribution. Besides, the matrix on the right represents the probability that an individual in age group a (rows) contacts with an individual in age group a' (columns).

In the POLYMOD study volunteers from eight different European countries were asked to provide information about who they contacted during a given period of time. In particular, the information that is relevant for our interests is the age of the participant, a, and the age of the person contacted, a'. Then, aggregating this data, we can obtain the distribution of the number of contacts that an average person of age a has with people of age a', $P(k, a, a')$. In figure 2.10A we show this distribution for the case of Italy, which was one of the most complete surveys of the study.

To facilitate the visualization of the results, we have chosen to classify individuals, following previous literature [144], into two age groups: children (aged 18 or less) and adults (aged 19 or more). Then, we have measured the number of contacts each child had with other children (top-left distribution) and with adults (top-right distribution). Similarly, on the bottom row we show the contact distribution of adults with children (bottom-left) and of adults with other adults (bottom-right). All empirical distributions have been fitted to a negative binomial distribution, which is one of the most common methods for analyzing count response surveys[1] [147].

However, this data has some drawbacks, such as the problem of reciprocity. In a completely closed population (and with perfect measurement instruments), for each contact

[1] Count data is usually modeled in physics as a Poisson distribution, whose mean equals its variance. In count data involving humans, however, this is not usually the case. Thus, negative binomial distributions are used instead, as their variance can be larger than the mean. This special characteristic of these distributions is known as overdispersion.

that a child reports with an adult we should get a report of and adult contacting a child. But, as in this survey the population was not closed, that is, people participating in the survey could report contacts with people not participating in it, the reciprocity condition is not always satisfied. This is a well known problem in the study of mixing patterns of the population and there are several techniques used to solve it, such as the pairwise correction [148]. Thus, we will perform an assumption that will allow us to leverage all the techniques that have already been developed to work with age-contact matrices.

Indeed, from now on, we will suppose that the probability of having k contacts is independent of the age of both individuals involved in it. That is, $P(k, a, a') \approx P(k)P(a, a')$. Note that with this assumption we have separated the mixing patterns of the population, $P(a, a')$, from the main topological characteristic of networks, the degree distribution $P(k)$. This has two main advantages. First, this will allow us to directly input into the model mixing contact matrices, which are more common in the literature than the whole survey data. Second, we will be able to test the effect of different degree distributions. This situation will be further analyzed in section 3.4.

In figure 2.10B we show the necessary information to build $P(k, a, a')$ under this assumption. As we can see, the average number of contacts of any individual in the Italian population is close to 21. Interestingly, this number can vary greatly. For instance, while Italy was the country with the highest number of reported contacts, Germany was the one with the fewest daily contacts, 8. Regarding $P(a, a')$, it is important to note that the matrix is not symmetric per se, but this is due to the fact that age groups are not equally sized. Indeed, if we suppose that all groups have the same degree distribution, the reciprocity condition implies that $N_a P(a, a') = N_{a'} P(a', a)$, where N_a represents the number of individuals with age a.

Summing up, to obtain $k^{\beta}_{i,\alpha}$ we simply need to sample k from the desired degree distribution for each node i of age α. Then, we can distribute this value across layers using a multinomial distribution with probability parameters $P(\alpha, \beta)$.

2.6.3 Age contact networks

We will finish this section demonstrating the full process of generating one age contact network, which can be divided in three steps:

1. **Data collection and standardization**: the first step is to obtain demographic structure, mixing patterns and degree distribution of the population under consideration. Standardization in this context refers to deciding the number of age groups in the system and modifying the distributions accordingly.

   As in the previous section, we will focus on Italy in the year 2005, which is when the POLYMOD study took place. Besides, we will consider that there are still only two groups, children and adults. Under these circumstances, according to the demographic structure of the country, 18% of the nodes must be set in the age group of children while the other 82% must belong to the adults group. The $P(a, a')$ matrix as well as the degree distribution will be the ones shown in figure 2.10B.

   Next, we need to decide the number of nodes in our network. In this case, we will set this number equal to $N = 10^3$. Hence, there will be 180 nodes and 819 nodes in layers 1 and 2 respectively. Once this is set, we can extract suitable values of $k^{\beta}_{i,\alpha}$ for each node i from the $P(k, a, a')$ distribution as explained in the previous section. These will be the constraints in our model.

2. **Determination of the ensemble**: once the constraints are set, the next step is to maximize the log-likelihood defined in (2.60) to obtain the proper values of $x^{\beta}_{i,\alpha}$. These can be used to build the ensemble of networks using (2.58).

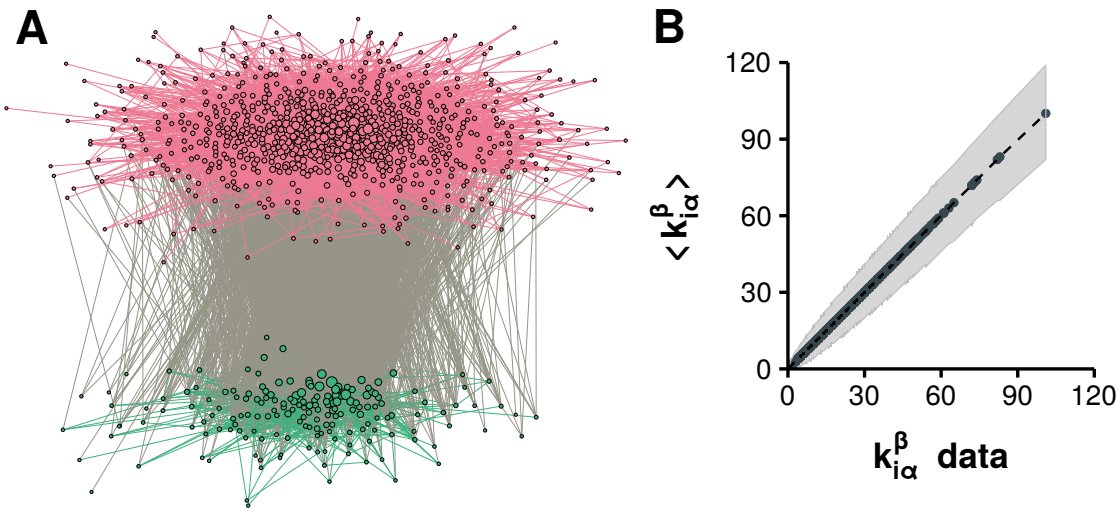

Figure 2.11: Synthetic Italian age contact network. A) depicts a random network obtained from the ensemble. The size of the nodes is proportional to their degree. The layer on the bottom corresponds to nodes classified as children, while the layer on top corresponds to adults. The difference in group size is due to the demographic structure of the country. B) Comparison of the average degree of each node obtained from averaging 10^4 networks sampled from the ensemble to the imposed constraint. The 95% confidence interval of the distribution is shown in gray for each node.

3. **Network sampling**: the last step is to randomly sample as many networks as desired from the ensemble. This can be easily done by performing a Bernoulli trial over each pair of nodes i and j with probability of success given by p_{ij} as defined in equation (2.59).

For this example, we have sampled 10^4 networks from the ensemble. In figure 2.11A we show one of the sampled networks. As it can be seen, the network is divided in two layers. Nodes in the bottom layer represent children while the ones on top adults. As imposed by the demographic constraints, there are many more nodes in the adults layer than in the children layer. To confirm that the sampled networks successfully preserve the rest of the constraints, in figure 2.11B we plot the average value of $k^{\beta}_{i,\alpha}$ in the set of 10^4 networks (y axis) against the value imposed by the data (x axis). Clearly, all nodes are close to the identity line, indicating that the conditions were successfully maintained.

This example demonstrates that we now have all the tools required to generate realistic contact networks with age related data. In section 3.4 we will use this framework to build networks for different countries and periods of time, and we will measure the impact that incorporating this kind of data can have on disease dynamics.

3 The law of mass action: animals collide

Infectious diseases have been an unpleasant companion of humankind for millions of years. Yet, crowd epidemic diseases could have only emerged within the past 11,000 years, following the rise of agriculture. The ability to maintain large and dense human populations, as well as the close contact with domestic animals, allowed the most deadly diseases to be sustained unlike when human populations were sparse [149].

Perhaps the most-well documented epidemic outbreak in ancient times is the plague of Athens (430-427 BCE) that caused the death of Pericles and killed around 30% of Athens population [150]. The fact that some diseases were contagious was probably well-known way before that. For instance, it has been claimed that in the 14th century BCE the Hittites sent rams infected with tularemia to their enemies to weaken them [151] and there are evidences of quarantine-like isolation of leprous individuals in the Biblical book of Leviticus. Yet, it was thought that diseases were caused by *miasma* or "bad air" for over 2,000 years. It was not until the end of the XIX century that it was finally discovered that microorganisms were the cause of diseases [152].

The advent of modern epidemiology is usually attributed to John Snow who in the mid of the XIX century traced back the origin of a cholera epidemic in the city of London [153]. However, mathematical methods were not firmly introduced until the beginning of the XX century[1]. Already in 1906 Hamer showed that "an epidemic outbreak could come to an end despite the existence of large numbers of susceptible persons in the population, merely on a *mechanical theory of numbers and density*" [156]. Although it was thanks to the works by Ross, Kermack and McKendrick that finally a mechanistic theory of epidemics was developed as an analogy to the law of mass-action. In particular, it was McKendrick who gave the title to this chapter when he said in a lecture in 1912: "consider a type of epidemic which is spread by simple contact from human being to human being [...] The rate at which this epidemic will spread depends obviously on the number of infected animals, and also on the number of animals that remain to be infected - in other words the occurence of a new infection depends on a *collision between an infected and uninfected animal*" [157].

The next 50 years were mostly devoted to establishing the mathematical foundations of epidemiology. The problem was that this process was mostly done by mathematicians and statisticians, who were more interested in the theoretical implications of the models rather than in their application to data [158]. This situation changed during the 1980s when Anderson and May, coming from a background in zoology and ecology respectively, started to collaborate with biologists and mathematicians, bridging the gap between data and theory [63]. During the 1990s graphs where introduced in epidemiological models, challenging the classical assumption of homogeneous mixing (that we will discuss in section 3.1), and brought physicists into the field attracted by the similarity of some concepts with phase transitions in non-equilibrium systems [159].

[1] A noteworthy exception is the work by Daniel Bernoulli in 1766, although he was just too ahead of his time. Furthermore, some authors claim that the credit for creating the first modern disease transmission model should go to Pyotr Dimitrievich Enko. Unfortunately, even though he published his work as early as 1889, he wrote it in russian and thus it became largely unnoticed for the majority of the scientific community until it was translated by the middle of the XX century (see [154, 155] for a nice introduction to the early days of mathematical epidemiology).

The latest developments of epidemic modeling are based on incorporating more and more data. For instance, the Global Epidemic and Mobility (GLEaM) framework incorporates demographic data of the whole world with short-range and long-range mobility data as the basis of its epidemic model, allowing for the simulation of world-wide pandemics [160]. Similarly, to properly study the spreading of Zika virus it is necessary to take into account the dynamics of mosquitoes, temperature, demographics and mobility, attached to the disease dynamics of the own virus [161]. Multiple sources of data are also being used in the study of vaccination, either to devise efficient administration strategies, including economic considerations [162], or to properly understand how they actually work [163]. Even more, of particular interest nowadays is analyzing the interplay between processes that have been deeply studied in complex systems such as game theory, behavior diffusion and epidemic processes.

Herd immunity, a term coined by Topley and Wilson in 1923 albeit with the completely opposite meaning to the current one [164], refers to the fact that it is possible to have a population where diseases cannot spread even if only a fraction of the individuals are immune to it. Firstly calculated theoretically in 1970 by Smith [165], it has been the subject of great interest as it allows to completely immunize a population even if there are members who cannot be administered a vaccine due to their medical conditions [166]. Unfortunately, the great successes achieved by vaccination are now endangered by people who refuse to vaccinate their children, which also affects those kids who cannot be vaccinated but should have been protected by herd immunity [167].

For instance, measles requires 95% of the population to be vaccinated for herd immunity to work. This was achieved in the U.S. by the end of the past century, being declared measles free in 2000. Similarly, the UK was declared measles free in 2017. Yet, the World Health Organization (WHO) removed this status from the UK in August 2019 [168], and the U.S. is facing a similar fate as so far in 2019 they have reported the greatest number of cases since 1992 [169]. Both phenomena have been attributed to anti-vaccine groups, whose behavior can be studied from the point of view of game theory. But there are more ingredients into play. In particular, if the risk of infection is regarded low, maybe thanks to herd immunity, the motivation to become vaccinated can decrease. This behavior can then be spread among adults, a process that will be clearly coupled with the disease dynamics. Thus, a holistic view of the whole problem is needed, something that can only be done under the lenses of complex systems [134].

In this context, rather than extending the mathematical formalism that is already well established, we pushed forward our knowledge about disease dynamics by adding data and revisiting some of the assumptions classically made either for simplicity or lack of information. For this reason, rather than giving a whole mathematical introduction and then visiting each contribution, we will organize them in a way that roughly follows the historical development of mathematical epidemiology, explaining in each section the basic ideas and then showing how we challenged those assumptions.

We will begin in section 3.1 with the most basic approach to disease dynamics. That is, humans are gathered in closed populations in which every individual can contact every other, very much like particles colliding in a box. This simple premise, known as *homogeneous mixing*, can be slightly improved by considering individuals to be part of smaller groups, with correspondingly different patterns of interaction. This is the classical approach to introduce the age structure of the population, for which experimental data exist. We will, however, go one step further and analyze the problem of projecting this data into the future, taking into account the demographic evolution of society. This part of the thesis will be thus based on the publication:

- S. Arregui, A. Aleta, J. Sanz, and Y. Moreno, Projecting social contact matrices to different demographic structures, *PLoS Comput. Biol.*, vol. 14, pp. 1–18, Dec 2018.

Our next step will be to introduce, in section 3.2, one of the cornerstone quantities of modern epidemiology, the basic reproduction number, R_0. We will revisit its original definition and challenge it using data-driven population models, demonstrating that some of the assumptions that have been made since its conception are not entirely correct. This corresponds to the work:

- Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani, Measurability of the epidemic reproduction number in data-driven contact networks, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, pp. 12680–12685, Dec 2018

Then, in section 3.3 we will finally introduce networks into the picture. We will show some of the counter-intuitive consequences of this and, again, challenge some of the most basic assumptions. In particular, disease dynamics are often implemented on single layer undirected networks, but we will show that directionality can play a crucial role on the dynamics, with particular emphasis on multilayer networks. We will follow the article

- X. Wang, A. Aleta, D. Lu, and Y. Moreno, Directionality reduces the impact of epidemics in multilayer networks, *New J. Phys.*, vol. 21, p. 093026, Sep 2019

of which I am first co-author.

We will finish this chapter in section 3.4 analyzing the age-contact networks that we generated in section 2.6, chapter 2. The objective of this part will be to show the different approaches than can be followed depending on the available data and their impact in the outcome of the dynamics. This will be based on the work

- A. Aleta, G. Ferraz de Arruda, and Y. Moreno, Generating data-driven age contact networks, *In preparation*, 2019

3.1 A basic assumption: homogeneous mixing

The starting point of this discussion is going to be precisely the own introduction of the paper by Kermack and McKendrick published in 1927 that is regarded as the starting point of modern epidemiological models [172]. Even if over 90 years have passed, any text written today about the subject would start roughly in the same way:

"The problem may be summarised as follows: One (or more) infected person is introduced into a community of individuals, more or less susceptible to the disease in question. The disease spreads from the affected to the unaffected by contact infection. Each infected person runs through the course of his sickness, and finally is removed from the number of those who are sick, by recovery or by death. The chances of recovery or death vary from day to day during the course of his illness. The chances that the affected may convey infection to the unaffected are likewise dependent upon the stage of the sickness. As the epidemic spreads, the number of unaffected members of the community becomes reduced. Since the course of an epidemic is short compared with the life of an individual, the population may be considered as remaining constant, except in as far as it is modified by deaths due to the epidemic disease itself. In the course of time the epidemic may come to an end. [...] [This] discussion will be limited to the case in which all members of the community are initially equally susceptible to the disease, and it will be further assumed that complete immunity is conferred by a single infection."

For the shake of clarity we can summarize some of the implicit assumptions in the previous paragraph, plus some more that were introduced in other parts of the paper, as [173]:

1. The disease is directly transmitted from host to host.

2. The disease ends in either complete immunity or death.

3. Contacts are according to the law of mass-action.

4. Individuals are only distinguishable by their health status.

5. The population is closed.

6. The population is large enough to be described with a deterministic approach.

In this section we will explore the effect of relaxing assumptions 3 and 4. Note that these two assumptions can be regarded as an approximation when no sufficient data about the whereabouts of the population are known. Now, however, we have much more data available than they did and thus in section 3.2 we will be able to completely remove assumption 4. Similarly, in sections 3.3 and 3.4 we will suppress assumptions 2 and 3. Besides, except for this introduction, throughout the chapter we will disregard assumption 6, but we will always respect the 1st and 5th ones.

With modern terminology, models in which individuals are only distinguishable by their health status are known as *compartmental models*. In these models, it is supposed that each individual belongs to one and only one *compartment* (class, in Kermack and McKendrick terms). Compartments are a tool to encapsulate the complexity of infections in a simple way. Hence, an individual that is completely free from the disease but can be infected is said to be in the *susceptible* state (S), one that can spread the disease is said to be *infected* (I) and one that can neither be infected nor infect is said to be *removed* (R) either because is immune or dead. This classification is known as the SIR model. This framework, however, is quite flexible and it is possible to incorporate as many compartments as needed, depending on the disease under study, reaching hundreds of compartments in the most sophisticated models [174]. In particular, in section 3.1.2, we will introduce the *exposed* (E) state to classify individuals that have been infected but are not yet infectious.

The six assumptions, after some algebra, lead to the original equation proposed by Kermack and McKendrick (albeit with slightly updated notation),

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = S(t)\int_0^\infty A(\tau)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau\,, \tag{3.1}$$

where $S(t)$ denotes the number of individuals in the susceptible compartment - henceforth number of susceptibles - at time t and $A(\tau)$ is the expected infectivity of an individual that became infected τ units of time ago [172, 173].

To obtain $A(\tau)$, we define $\phi(\tau)$ as the rate of infectivity of an individual that has been infected for a time τ. Similarly, we define $\psi(\tau)$ as the rate of removal, either by immunization or death. Let us denote by $v(t,\tau)$ the number of individuals that are infected at time t and have been infected for a period of length τ. If we divide time into separate intervals Δt, such that the infection takes places only at the instant of passing from one interval to the next, the following relation holds:

$$\begin{aligned} v(t,\tau) &= v(t-\Delta t,\tau-\Delta t)(1-\psi(\tau-\Delta t)) \\ &= v(t-2\Delta t,\tau-2\Delta t)(1-\psi(\tau-\Delta t))(1-\psi(\tau-2\Delta t)) \\ &= v(t-\tau,0)B(\tau)\,, \end{aligned} \tag{3.2}$$

so that, if Δt is small enough,

$$\begin{aligned} B(\tau) &= (1-\psi(\tau-\Delta t))(1-\psi(\tau-2\Delta t))\dots(1-\psi(0)) \\ &\approx e^{-\psi(\tau-\Delta t)}e^{-\psi(\tau-2\Delta t)}\dots e^{-\psi(0)} \\ &\approx e^{-\int_0^\tau \psi(a)\mathrm{d}a}\,. \end{aligned} \tag{3.3}$$

Hence,

$$A(\tau) = \phi(\tau)B(\tau) = \phi(\tau)e^{-\int_0^\tau \psi(a)\mathrm{d}a}\,, \tag{3.4}$$

which defines the original shape of the Kermack and McKendrick model.

However, in the literature it is common to present as the Kermack and McKendrick model the special case they analyze in their paper in which both the infectivity and removal rates are constant. Indeed, if we set $\phi(\tau) = \beta$ and $\psi(\tau) = \mu$,

$$A(\tau) = \beta e^{-\int_0^\tau \mu \mathrm{d}a} = \beta e^{-\mu\tau}\,, \tag{3.5}$$

and defining the number of infected individuals at time t as

$$I(t) \equiv -\frac{1}{\beta}\int_0^\infty A(\tau)\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau\,, \tag{3.6}$$

equation 3.1 reads

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta I(t)S(t)\,. \tag{3.7}$$

If we now derive expression (3.6), using Leibniz's rule,

$$\begin{aligned}\frac{\mathrm{d}I(t)}{\mathrm{d}t} &= -\frac{\mathrm{d}S(t)}{\mathrm{d}t} - \int_{-\infty}^{t}\frac{\partial}{\partial t}e^{-\mu(t-\tau)}\frac{\mathrm{d}S(\tau)}{\mathrm{d}t}\mathrm{d}\tau \\ &= -\frac{\mathrm{d}S(t)}{\mathrm{d}t} + \mu\int_0^\infty e^{-\mu\tau}\frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t}\mathrm{d}\tau \\ &= \beta I(t)S(t) - \mu I(t)\,,\end{aligned} \tag{3.8}$$

together with the fact that the population, N, is closed, $S(t) + I(t) + R(t) = N$, we obtain the system of equations

$$\begin{cases}\dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta I(t)S(t) \\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta I(t)S(t) - \mu I(t) \\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t} = \mu I(t)\end{cases} \tag{3.9}$$

which is the model that is usually introduced as the Kermack-McKendrick, even though we have seen that their original contribution was much more general [175].

Equation (3.9) is also often used to introduce epidemic models in the literature as it constitutes one of the most basic models. As we are considering that every individual can contact every other this model is also known as the *homogeneous mixing* model [159, 176]. However, it should be noted that sometimes a slightly different version of this set of equations is presented. Indeed, if we define the fraction of susceptible individuals in the population as $s(t) \equiv S(t)/N$, and similarly with the others, note that the expression for the evolution of infected individuals is

$$\frac{\mathrm{d}i(t)}{\mathrm{d}t} = \beta N i(t)s(t) - \mu i(t)\,. \tag{3.10}$$

Hence, the larger the population, the faster the spreading. This is known as the *density dependent* approach. However, we can formulate a very similar model in which we define the infectivity rate as $\phi(\tau) = \beta/N$, so that

$$\frac{\mathrm{d}i(t)}{\mathrm{d}t} = \beta i(t)s(t) - \mu i(t) \tag{3.11}$$

is independent of N. This latter approach is called *frequency dependent* and is probably the most common one in the literature of epidemic processes on networks. Both approaches are valid and depend on the specific disease that is being modeled, see [177] for a deeper discussion of this matter.

Despite the simplicity of this model, it provides two very powerful insights about disease dynamics. The first one is related to the reasons that account for the termination of an epidemic. Until the publication of this model, the most accepted explanations in medical circles were that an epidemic stopped either because all susceptible individuals had been removed or because during the course of the epidemic the virulence of the organism causing the disease decreased gradually [178]. Yet, this model shows that with a fixed virulence (β) it is possible to reach states in which the epidemic fades out even if there are still susceptible individuals. To demonstrate this, although some approximations can be done to show this behavior (there is no closed form solution of the model), for our purposes it suffices to show a numerical solution, figure 3.1A.

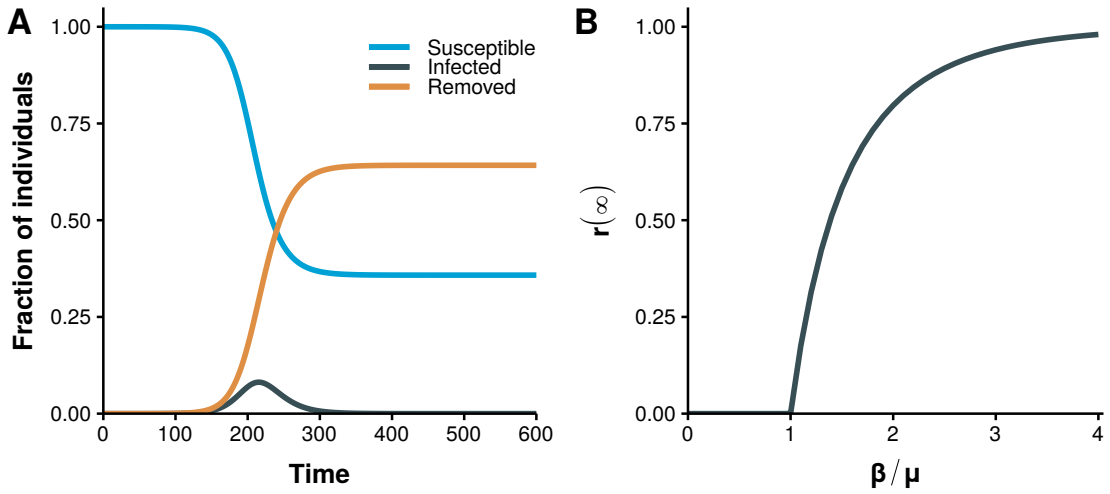

Figure 3.1: Basic results of the homogeneous mixing model. In panel A the evolution of the set of equations (3.9) as a function of time with $\beta = 0.16$ and $\mu = 0.10$ is shown. It is possible to reach a disease free state with a fraction of susceptible individuals larger than 0. In panel B the total fraction of recovered individuals in equilibrium conditions as a function of β/μ is shown. For simplicity the frequency dependent approach has been used so that the threshold is 1.

At this point a clarification might be in order. During the introduction we said that Hamer had already showed in 1906 that it was possible for an epidemic to end despite the existence of large number of susceptible persons in the population. However, the difference resides in that Hamer proposal was based on data about measles, while this model is formulated without any specific disease in mind. Indeed, although clearly influenced by Hamer's and Ross' works, one of the great achievements of Kermack and McKendrick was to establish a formulation based only on mechanistic principles, regardless of the specific properties of the disease. Nonetheless, the most important result of this model has not been discussed yet, the *epidemic threshold*.

Suppose that in a completely susceptible population we introduce a tiny amount of infected individuals so that $s(t=0) \equiv s_0 = 1 - \epsilon$ and $i(t=0) \equiv i_0 = \epsilon$ with $\epsilon \to 0$. If we linearize equation (3.10) around this point, we have

$$\frac{\mathrm{d}i(t)}{\mathrm{d}t} \approx \beta N i_0 s_0 - \mu i_0 \,, \tag{3.12}$$

which only grows if $\beta N - \mu > 0$. Hence, there exists a minimum susceptible population at the initial state below which an epidemic cannot take place, the epidemic threshold:

$$N_c > \frac{\mu}{\beta} \,. \tag{3.13}$$

Note that the formulation of this threshold can vary slightly according to the characteristics of the model. For instance, in the frequency dependent approach, equation (3.11), the epidemic threshold is defined by

$$\frac{\beta}{\mu} > 1\,, \tag{3.14}$$

which is independent of N. The existence of this threshold is numerically demonstrated in figure 3.1B, where the final fraction of recovered individuals as a function of the ratio β/μ is shown. Regardless of the specific shape of the condition, the message is that it is possible to explain why an epidemic might not spread in a population of fully susceptible individuals. Moreover, it also provides a mechanism to fight diseases before they spread. Indeed, in equation 3.13 we have simply considered that $S_0 = N$, but if we were able to immunize a fraction of the population so that $S_0 < \mu/\beta$ then the epidemic could not take place. In other words, we would have conferred the population the herd immunity discussed in the beginning of the chapter.

3.1.1 Introducing the age compartment

Since the establishment of epidemiology as a science, a lot of attention has been devoted to the study of measles as its recurring patterns puzzled physicians and mathematicians alike. The distinguishing characteristic of measles epidemics is that they had a very regular temporal pattern with periodic outbreaks of the disease, as shown in figure 3.2. As this disease affects specially the children and also conveys permanent immunity to those who have suffered it, analyzing the time evolution over large time-scales to obtain the patterns required the inclusion of age in the models. Nevertheless, with the basic model that we have analyzed we can already propose a plausible explanation for this behavior. Indeed, we know that if the amount of susceptibles in the population is below a given threshold, the epidemic cannot take place. Thus, it seems reasonable to think that once the epidemic fades-out, there is a period in which there are not enough susceptibles for it to appear again. Yet, when new children are born, the amount of susceptibles will increase, possibly going above the threshold and therefore allowing a new outbreak.

A similar explanation was already proposed by Soper in 1929 [179], although it only matched the observations qualitatively, not quantitatively. It was Barlett who, in 1957, finally provided a quantitative explanation of the phenomenon [180]. Besides the details that we have already discussed, in his proposal he added a new factor that we have not mentioned yet. He proposed that the problem of previous models was that they were deterministic, an approximation that is only valid in very large populations. However, it was observed that the periodicity of measles not only depended on the size of the city, but it was specially so in small towns. In physical terms, we would say that there were finite size effects, tearing down assumption 6 (see section 3.1). Thus, he proposed to use a stochastic model for which he could not obtain a closed form solution, so he had to resort to an "electronic computer". Nowadays the use of stochastic computational simulations are much more common than the deterministic approach. The reasons why this approach is more favorable are out of the scope of this thesis (see, for instance [154, 181, 182, 183] for a discussion) but we will leverage this opportunity to say that in the following sections we will mostly work with stochastic simulations, rather than deterministic approaches. Before concluding the discussion about Barlett's paper, we find worth highlighting that it was presented during a meeting of the Royal Statistical Society, in 1956, after which a discussion followed. In said discussion, Norman T. J. Bailey said *"One of the signs of the times is the use of an electronic computer to handle the Monte Carlo experiments. Provided they are not made an excuse for avoiding difficult mathematics, I think there is a great scope for such computers in biometrical work"*. And indeed there was, as 25 years later Mr. Bailey was appointed Professor of Medical Informatics [184].

Returning to our discussion, it is not surprising, then, that McKendrick already

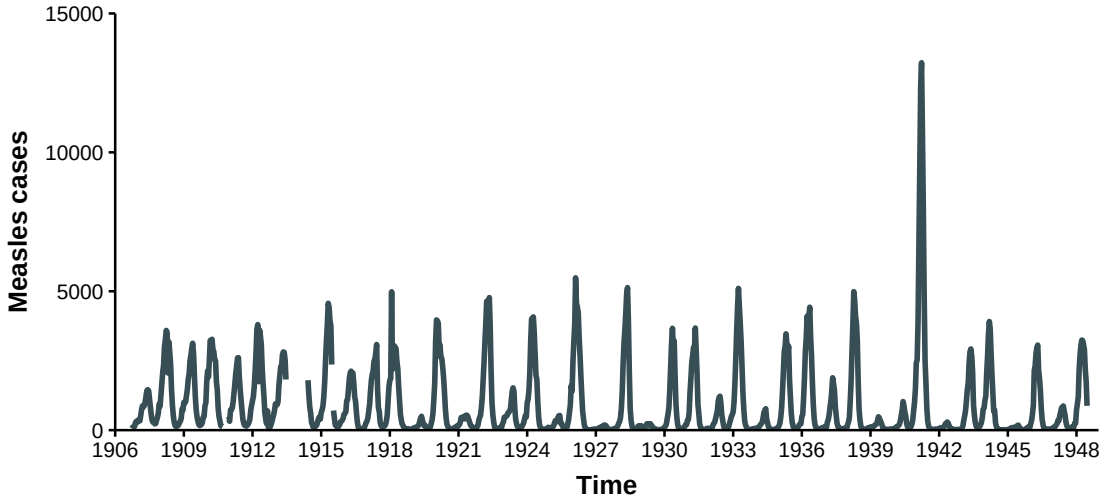

Figure 3.2: Measles epidemics in New York from 1906 to 1948. This figure represents the number of reported cases of measles in the city of New York from 1906 to 1948 with a biweekly resolution. There are some gaps due to missing reports. Data obtained from [185].

introduced age in his models in 1926 [186], one year before the publication of the full model that we have already explored. However, to introduce age we will use a slightly more modern formulation that will simplify the analysis. In particular, we need to revisit assumption 4, i.e., individuals are only distinguishable by their health status.

Let us state that individuals can now be identified both by their health status and their age. Hence, we have to add more compartments to the model, one for each age group and health status combination. In other words, rather than having three compartments, S, I, R, we now have 3 times the number of age brackets considered, i.e. S_a, I_a, R_a being a the age bracket the individuals belong to (see 2.6 for the definition of age bracket). Moreover, we will suppose that the disease dynamics is much faster than the demographic evolution of the population. The only thing left is to decide how to go from one compartment to another:

- For the rate of infectivity, we will define an auxiliary expression that will facilitate the discussion. By inspection of equation (3.9), we can define the *force of infection* [178] as

$$\lambda(t) \equiv \phi(\tau) I(t) = \beta I(t)\,, \tag{3.15}$$

  which does not depend on any characteristic of the individual. Hence, we can simply incorporate age by modifying the force of infection so that

$$\lambda(t, a) = \sum_{a'} \phi(\tau, a, a') I_{a'}(t)\,. \tag{3.16}$$

  This way, both the age of the individual that is getting infected (a) and the age of all other individuals ($\sum_{a'}$) are taken into account. Furthermore, we can separate $\phi(\tau, a, a')$ into two components: one accounting for the rate of contacts between individuals of age a and a' and another one accounting for the likelihood that such contacts lead to an infection. Hence,

$$\phi(\tau, a, a') \equiv C(a, a')\beta(a, a')\,. \tag{3.17}$$

  Recalling section 2.6, the term $C(a, a')$ can be obtained from the contact surveys that we have already studied. On the other hand, we will suppose that the likelihood of infection is independent of the age so that $\beta(a, a') = \beta$.

- For the rate of recovery, we will assume that it is independent of the age of the individual, i.e. $\mu(a) = \mu$.

Under these assumptions, the homogeneous mixing model with age dependent contacts reads

$$\begin{cases} \dfrac{\mathrm{d}S_a(t)}{\mathrm{d}t} = -\displaystyle\sum_{a'} \beta C(a,a') I_{a'}(t) S_a(t) \\ \dfrac{\mathrm{d}I_a(t)}{\mathrm{d}t} = \displaystyle\sum_{a'} \beta C(a,a') I_{a'}(t) S_a(t) - \mu I_A(t) \\ \dfrac{\mathrm{d}R_a(t)}{\mathrm{d}t} = \mu I_a(t) \end{cases} \tag{3.18}$$

Despite its simplicity, this model is still widely used today, specially in the context of metapopulations[2] [144]. Even more, as in all compartmental models, it is straightforward to extend it to include more complex dynamics. For instance, we can add the exposed state so that individuals that get infected remain in a latent state for a certain amount of time before showing symptoms and being able to infect others. This model, known as the *SEIR* model, can be used to describe influenza dynamics [188]

$$\begin{cases} \dfrac{\mathrm{d}S_a(t)}{\mathrm{d}t} = -\displaystyle\sum_{a'} \beta C(a,a') I_{a'}(t) S_a(t) \\ \dfrac{\mathrm{d}E_a(t)}{\mathrm{d}t} = \displaystyle\sum_{a'} \beta C(a,a') I_{a'}(t) S_a(t) - \sigma E_a(t) \\ \dfrac{\mathrm{d}I_a(t)}{\mathrm{d}t} = \sigma E_a(t) - \mu I_a(t) \\ \dfrac{\mathrm{d}R_a(t)}{\mathrm{d}t} = \mu I_a(t) \end{cases} . \tag{3.19}$$

The new parameter, σ, accounts for the rate at which an individual from the latent state goes to the infectious state, in a similar fashion as μ does for the transition from I to R. This model will be the focus of the last part of this section.

3.1.2 Changing demographics

> *I call myself a Social Atom - a small speck on the surface of society.*
> ("Memoirs of a social atom", William E. Adams)

As we discussed earlier, for a long period of time the developments in mathematical epidemiology were disconnected from data, at least until Anderson and May arrived to the field in the late 1980s. It is not so surprising, then, that even though age was incorporated into models since the beginning of the discipline, we had to wait until the late 1990s to get experimental data of age mixing patterns.

The first attempt to quantify the mixing behavior responsible for infections transmitted by respiratory droplets or close contact (which are the ones best suited to be studied with homogeneous mixing models) was the pioneering work by Edmunds et al. in 1997 [189]. Their results, however, can hardly be extrapolated as they only analyzed a population consisting of 62 individuals coming from two British universities. The first large-scale experiment to measure these patterns was conducted by Mossong et al. in 2008 [143]. In their study, they measured the age-dependent contact rates in eight European countries (Belgium, Finland, Germany, Great Britain, Italy, Luxembourg, Netherlands and Poland),

[2] A metapopulation is a set of populations that are spatially separated but can exchange individuals. Within each subpopulation any plausible disease dynamics can be implemented, although usually the homogeneous mixing approach is used [187].

as part of the European project Polymod, using contact diaries. In the next years other authors followed the route opened by Mossong et al. and measured the age-dependent social contacts of countries such as China [190], France [191], Japan [192], Kenya [193], Russia [194], Uganda [195] or Zimbabwe [196], as well as the Special Administrative Region of Hong Kong [197], greatly expanding the available empirical data.

These experiments provide us with the key ingredient required for the introduction of age compartments into the models, the age contact matrix, C. There are, however, a couple of ways of defining this matrix that are equivalent under certain transformations. We define the matrix in *extensive scale*, C, as the one in which each element $C_{i,j}$ contains the total number of contacts between two age groups i and j. It is trivial to see that given this definition there must be reciprocity in the system, i.e.,

$$C_{i,j} = C_{j,i} \tag{3.20}$$

A similar definition can be obtained if instead of accounting for all the contacts between two groups we want to capture the average number of contacts that a single individual of group i will have with individuals in group j:

$$M_{i,j} = \frac{C_{i,j}}{N_i}\,, \tag{3.21}$$

where N_i is the number of individuals in group i. We call the matrix in this form the *intensive scale*. This is the usual format in which this matrix is given. In this case, reciprocity is fulfilled if

$$M_{i,j}N_i = M_{j,i}N_j\,. \tag{3.22}$$

This last expression rises an interesting question. The reciprocity relation depends on the population in each age bracket, N_i. Thus, if the matrix M was measured in year y, we have that $M_{i,j}(y)N_i(y) = M_{j,i}(y)N_j(y)$. However, if we want to use this matrix in a different year, that is, with a different demographic structure due to the inherent evolution of the population, reciprocity will no longer be fulfilled, i.e.

$$M_{i,j}(y)N_i(y') \neq M_{j,i}(y)N_j(y')\,, \tag{3.23}$$

unless the population has not changed. This is a major problem because there are diseases whose temporal dynamics are comparable to the ones of the demographic evolution. For instance, Tuberculosis is a disease in which age is particularly important and the incubation period ranges from 1 to 30 years [198]. Hence, to properly forecast the evolution of Tuberculosis in a population it is strictly necessary to project somehow these age-contact matrices into the future [148]. Even for diseases that have much shorter dynamics, such as influenza, this is a relevant problem because given how costly these experiments are, it is unpractical to repeat them every few years to obtain updated matrices. As a consequence, if we simply want to study the impact of influenza this year, more than 10 years after the work by Massong et al., we need to devise a way to properly update them.

Figure 3.3 exemplifies, for Poland and Zimbabwe, the error we would make if we do not adapt M and blindly use it with demographic structures that are different than the original. We define the reciprocity error as

$$E = \frac{\sum_{i,j>i} |C_{i,j} - C_{j,i}|}{0.5 \cdot \sum_{i,j} C_{i,j}} = \frac{\sum_{i,j>i} |M_{i,j}N_i - M_{j,i}N_j|}{0.5 \cdot \sum_{i,j} M_{i,j}N_i}\,, \tag{3.24}$$

to quantify the fraction of links that are not reciprocal. The two countries under consideration have very different demographic patterns, both in the past and in the future, and yet we can see that the error is quite large in both of them.

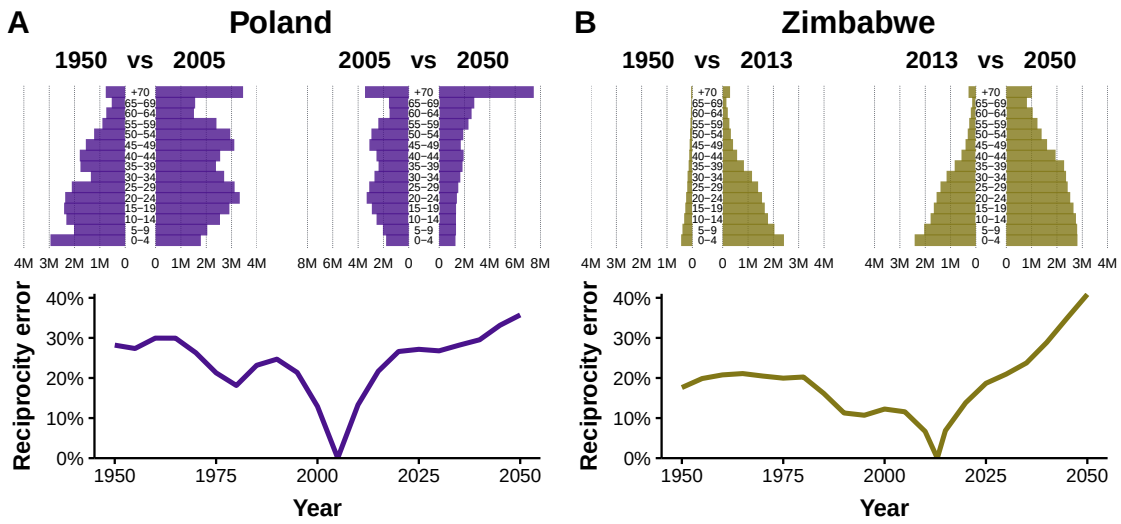

Figure 3.3: Reciprocity error as a function of time in Poland and Zimbabwe. For each country, in the top plots the demographic structures of 1950 and 2050 are compared to the one existing when the contact matrices were measured. In the bottom plot the reciprocity error as a function of time is shown. For the matrix to be correct in different years the error should be 0, but that only happens in the year when the data was collected.

The problem is that both $C_{i,j}$ and $M_{i,j}$ implicitly contain information about the demographic structure of the population at the time they were measured. To solve this problem, we define the *intrinsic connectivity matrix* as

$$\Gamma_{i,j} = M_{i,j}\frac{N}{N_j}\,. \tag{3.25}$$

This matrix corresponds, except for a global factor, to the contact pattern in a "rectangular" demography (a population structure where all age groups have the same density). Hence, it does not have any information about the demographic structure of the population.

In figure 3.4A we show the instrinsic connectivity matrices for each of the 16 regions enumerated previously. Interestingly, the contact patterns are quite different from region to region. To facilitate the comparison, in figure 3.4B we plot the fraction of connectivity that corresponds to young individuals (less than 20 years old) as a function of the assortativity of each matrix as defined by Newman [74] (this quantity is an adaptation of the Pearson correlation coefficient so that it is equal to 1 if individuals tend to contact those who are like them, -1 in the opposite case and 0 if the pattern is completely uncorrelated). We can see that regions with similar demographic structures and culture tend to cluster together, although it is not possible to disentangle which is the precise cause leading to one pattern or the other.

With this matrix we can now easily compute M at any other time, as long as we know the demographic structure of the population at that time:

$$M_{i,j}(y') = \Gamma_{i,j}\frac{N(y')}{N_j(y')} = M_{i,j}(y)\frac{N(y)N_j(y')}{N_j(y)N(y')}\,. \tag{3.26}$$

In our case, we will obtain this data from the UN population division database, which contains information of both the past demographic structures and their projections to 2050 for the whole world [199].

We conclude this section addressing how this correction impacts disease modeling. To this end, we simulate the spreading of an influenza-like disease both with and without

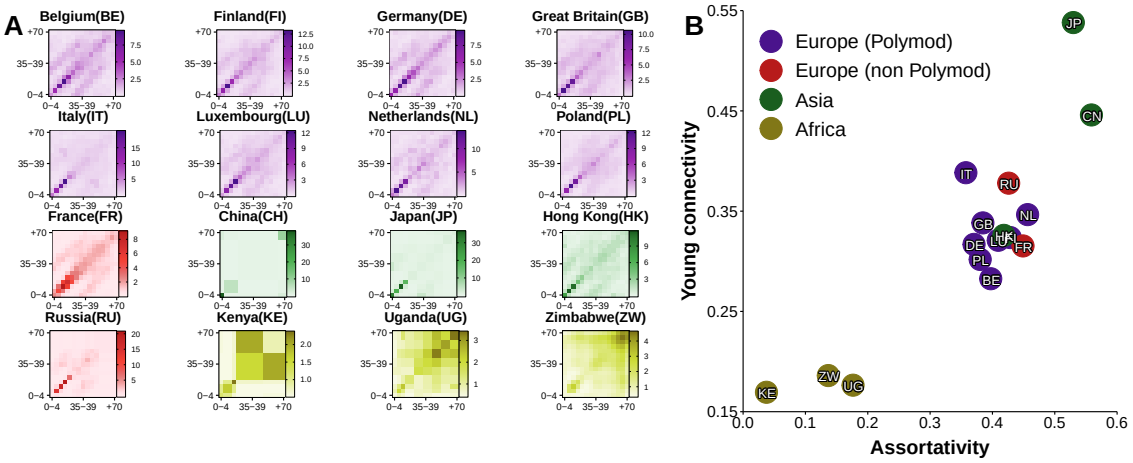

Figure 3.4: Age contact matrices from 16 regions. A) Intrinsic connectivity matrix, $\Gamma_{i,j}$, of each region. There is not a standard definition of age brackets and thus each study had its own definition. For comparison purposes, we have adapted the data to 15 age brackets: $[0,5),[5,10),\ldots,[65,70),+70$. B) Proportion of connectivity corresponding to individuals younger than 20 versus the assortativity coefficient of each matrix.

corrections on the matrix. We choose influenza because it is a short-cycle disease so that we can assume that the population structure is constant during each simulated outbreak. Besides, it can be effectively modeled using the SEIR model presented in 3.1.1.

To parameterize the model we use the values of an influenza outbreak that took place in Belgium in the season 2008/2009 [144]. Thus, individuals can catch the disease with transmissibility rate β per-contact with an infectious individual. The value of β is determined in each simulation so that the basic reproductive number is equal to 2.12 using the next generation approach (this procedure will be explained in more detail in section 3.2) [144, 200]. Once infected, individuals remain on a latency state for $\sigma^{-1} = 1.1$ days on average. Then, they become infectious for $\mu^{-1} = 3$ days on average, period when they can transmit the infection to susceptible individuals. After that, they recover and become immune to the disease. We use a discrete and stochastic model with the population divided into 15 age classes, whose mixing is given by the age contact matrix M. To sum up:

- The probability of an individual belonging to age group i to get infected is

$$p_{S\to E} = \beta \sum_j \frac{M_{i,j}}{N_j} I_j \,. \tag{3.27}$$

- Once in the latent state, the probability of entering the infected state is

$$p_{E\to I} = \sigma \,. \tag{3.28}$$

- Finally, an infected individual will recover with probability

$$p_{I\to R} = \mu \,. \tag{3.29}$$

Under these conditions, we compute the predicted size of the epidemic, i.e., $R(t\to\infty)$, in years 2000 and 2050. In figure 3.5 we present the results. In particular, in A we show the difference between the predicted size of the epidemic in 2050 versus the one in 2000 using the same M matrix in both years. In almost all countries the final size of the epidemic is smaller in 2050, except for China and the African countries in which it increases. However, in B we repeat the analysis but using the adapted values of $M(y')$ obtained using (3.26). In

this case, in general, the situation is reversed. Most countries have larger epidemics, except for the African ones. Even more, in countries such as China and Japan the difference is quite large, close to 20%.

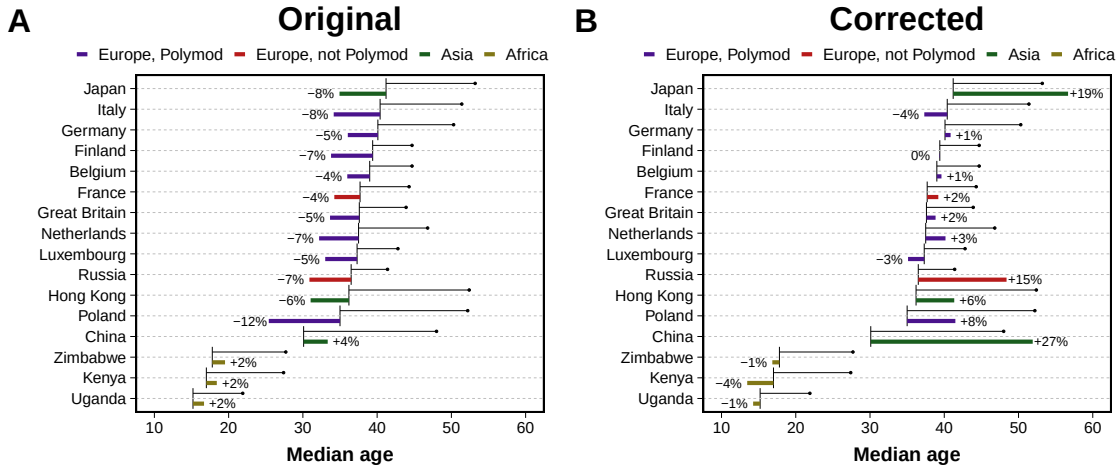

Figure 3.5: Predictions of influenza incidence in 2050 with demographic corrections. In both plots the black horizontal line starts at the median age of each region in the year 2000 and ends with a bullet point with the predicted value in 2050. Color bars denote the relative variation of incidence over the same period. In A the predictions are computed using the original contact matrices collected from the surveys. In B the proposed demographic corrections are applied to the matrices.

Summarizing, to create more realistic models we need to incorporate empirically measured data. However, blindly using data without thinking whether it can be applied to the specific system we are studying is not adequate. In the particular case of social mixing matrices, we have seen that even if we keep studying the same country, just moving a few years away from the moment in which the experiment took place dramatically affects the reciprocity of the contacts. This, in turn, leads to important differences in the global incidence for influenza-like diseases, as we have shown in our analysis of the SEIR model. Even more, since there are different intrinsic connectivity patterns across countries, it is possible that there exists a time evolution of this quantity. Indeed, if we believe that the intrinsic pattern is a consequence of the culture of the country, it seems logical to think that an evolving culture will also have evolving intrinsic connectivity patterns. Although predicting how society will change in the future is currently impossible, this should be taken into account as a limitation in any forecast for which heterogeneity in social mixing is a key element.

3.2 The basic reproduction number

One of the cornerstones of modern epidemiology is the *basic reproduction number*, R_0, defined as the expected number of individuals infected by a single infected person during her entire infectious period in a population which is entirely susceptible. From this definition, it is clear that if $R_0 < 1$, then, each infected individual will produce, on average, less than one infection. Therefore, the disease will not be able to be sustained in the population. Conversely, if $R_0 > 1$ the disease will be able to propagate to a macroscopic fraction of the population. Hence, this simple dimensionless quantity is informing us of three key aspects of a disease: (1) whether the disease will be able to invade the population, at least initially; (2) a way to determine which control measures, and at what magnitude, would be the most effective, i.e., which ones will reduce R_0 below 1; (3) to gauge the risk of an epidemic in emerging infectious diseases [201].

Interestingly, despite its importance, this quantity was not originated in epidemiology. The concept of R_0, and its notation, was formalized by Dublin and Lotka, in 1925, in the context of demography[3] [203]. The similitude of the concept in both fields is obvious, in one it measures the number of new infections per infected while in the other the number of births per female. Yet, in epidemiology the concept was mostly unknown until Anderson and May popularized it 60 years later in the Dahlem conference [204] (see [205] for a nice historical discussion on why it took so long for this concept to mature in epidemiology).

It might be enlightening to introduce the mathematical definition from the point of view of demography. Consider a large population. Let $F_d(a)$ be the survival function, i.e., the probability for a new-born individual to survive at least to age a, and let $b(a)$ denote the average number of offspring that an individual will produce per unit of time at age a. The function $n(a) \equiv b(a)F_d(a)$ is called the reproduction function. Hence, the expected future offspring of a new-born individual, R_0, is [206]

$$R_0^{demo} \equiv \int_0^\infty n(a)\mathrm{d}a = \int_0^\infty b(a)F_d(a)\mathrm{d}a\,. \tag{3.30}$$

The translation of this definition to epidemiology is straightforward. First, note that the reproduction function at age a is equivalent to the expected infectivity of an individual who was infected τ units of time ago, $A(\tau)$ (see equation 3.4). There is, however, one crucial difference. While in demography it is possible to "create" new individuals regardless of the size of the rest of the population, in epidemiology the creation of new individuals depends both on the infectivity and on the amount of susceptible individuals in the population. Hence,

$$R_0(\eta) \equiv \int_\Omega S(\xi)\int_0^\infty A(\tau,\xi,\eta)\mathrm{d}\tau\mathrm{d}\xi\,. \tag{3.31}$$

This rather cryptic expression is the most general definition of this quantity [207], although we will see in a moment simpler ones. The expression should be read as follows: the value of R_0 for individuals in an infectious state η is equal to the sum of all individuals in a susceptible state characterized by ξ, of size Ω, times the infectivity of individuals in said state η who where infected τ steps ago and can infect individuals in state ξ.

In the particular case of the SIR model under the density dependent approach, equation (3.9), the basic reproduction number is simply

$$\begin{aligned} R_0 &= S_0\int_0^\infty A(\tau)\mathrm{d}\tau = S_0\int_0^\infty \beta e^{-\mu\tau}\mathrm{d}\tau \\ &= \frac{\beta S_0}{\mu}\,, \end{aligned} \tag{3.32}$$

where S_0 is the number of susceptible individuals at the beginning of the infection, which in the absence of immunized individuals is equal to N. Recall that the linear stability analysis of the SIR model (3.12) yields,

$$i(t) = i_0 e^{\beta N-\mu}\,, \tag{3.33}$$

which only grows if $\beta N > \mu$. In other words, R_0 defines precisely the epidemic threshold that we found in the previous section, i.e.,

$$R_0 = \frac{\beta N}{\mu} > 1\,, \tag{3.34}$$

[3] In the original paper at a certain point the authors expand a quantity using Taylor's theorem an introduce the notation $R_n = \int_0^\infty a^n p(a)m(a)\mathrm{d}A$. Then, they determine the ratio between the total births in two successive generations to be $\int_0^\infty p(a)m(a)\mathrm{d}A$, which is equal to R_0 according to the previous definition. Thus, the subscript 0 historically would represent the 0-*th* moment of the distribution. In modern literature, however, the subscript is interpreted as referring to the "very beginning" of the epidemic [154]. Note also that R_0 is usually pronounced *R naught* in Britain and *R zero* in the U.S. [202].

as heuristically discussed at the beginning of this section. In the frequency dependent approach, which is more common in the network science literature as we shall see in 3.3, the basic reproduction number[4] is

$$R_0 = \frac{\beta}{\mu} > 1\,. \tag{3.35}$$

To obtain an explicit expression for R_0 with more complex compartmental models, an alternative approach to linear stability analysis is the *next generation matrix* as proposed by Diekmann in 1990 [207] and further elaborated by van den Driessche and Watmough in 2002 [200]. Briefly, the idea is to study the stability of the disease free state, x_0. To do so, we restrict the model to those compartments with infected individuals and separate the evolution due to new individuals getting infected, $\mathcal{F}$, and the transitions resulting for any other reason,

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = \mathcal{F}_i(x) - \mathcal{V}_i(x)\,, \tag{3.36}$$

where $x = (x_1, \ldots, x_m)$ denotes the m infected states in the model. If we now define

$$F = \left[\frac{\partial \mathcal{F}_i(x_0)}{\partial x_j}\right] \quad \text{and} \quad V = \left[\frac{\partial \mathcal{V}_i(x_0)}{\partial x_j}\right]\,, \tag{3.37}$$

the next generation matrix is FV^{-1} and the basic reproductive number can be obtained as

$$R_0 = \rho(FV^{-1}) \tag{3.38}$$

where ρ denotes the spectral radius [208]. In particular, for the model considered in section 3.1.2, the next generation matrix reads

$$K_{i,j} = \frac{\beta}{\mu}\frac{M_{i,j}}{N_j}\,. \tag{3.39}$$

This expression can be used to ensure that, regardless of the values of $M_{i,j}$ and N_j, the starting point of the dynamics is the same. For this reason, when we wanted to address the differences in incidence consequence of the changing demographics, we fitted β so that the spectral radius of (3.39) was always $R_0 = 2.12$.

It is worth pointing out that R_0 clearly depends on the model we choose for the dynamics. As a consequence, even though its epidemiological definition is completely independent from models (number of secondary infections per infected individual in a fully susceptible population), its mathematical formulation is not univocal. Ideally, if in a disease outbreak we knew who infected whom, we would be able to obtain the exact value of R_0. In reality, however, this information is seldom available. Hence, to compute it, one often relies on aggregated quantities, such as β and μ in equation (3.35). The problem is that if we assume that a disease can be modeled within a specific framework, we cannot directly compare the value obtained for R_0 with the ones measured for other diseases unless the exact same model has been used to obtain it. This is one of the observations that will motivate our work, which we will describe in section 3.2.3.

3.2.1 Measuring $\mathbf{R_0}$

Measuring R_0 is not an easy task, specially in the case of emergent diseases for which fast forecasts are required. An accurate estimation of its value is crucial to planning for the

[4] Although the notation R_0 is well established, there is not an agreement on how to call this quantity. Exchanging the word reproduction for *reproductive* is common in the literature, as well as using *rate* or *ratio* instead of number. Although the differences are minimal, note that *rate* is clearly wrong as it is not a rate but a (dimensionless) ratio [206].

control of an infection, but usually the only available information about the transmissibility of a new infectious disease is restricted to the daily count of new cases. Fortunately, it is possible, under certain conditions, to obtain an expression for R_0 as a function of that data.

Following [209], we will assume that in the beginning of a disease outbreak the growth of the number of infected individuals is exponential. Hence, the number of new infected individuals at time t will be equal to the number of new infected individuals τ time units ago, multiplied by the exponential growth,

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \frac{\mathrm{d}S(t-\tau)}{\mathrm{d}t} e^{r\tau} \tag{3.40}$$

where r denotes the growth rate. Inserting this expression in (3.1) with $t \to 0$,

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = S(t=0) \int_0^\infty A(\tau) \frac{\mathrm{d}S(t)}{\mathrm{d}t} e^{-r\tau} \mathrm{d}\tau \Rightarrow 1 = S_0 \int_0^\infty A(\tau) e^{-r\tau} \mathrm{d}\tau \tag{3.41}$$

At this point it might be enlightening to return once again to the demographic simile. In equation (3.30) we saw that the total number of offspring of a person could be obtained integrating $n(a)$ (rate of reproduction at age a) over the whole lifespan of the individual. Thus, we can define the distribution of the age a person has when she has a child as

$$g'(a) = \frac{n(a)}{\int_0^\infty n(a)\mathrm{d}a} = \frac{n(a)}{R_0^{demo}} \,. \tag{3.42}$$

If we take the "age" of an infection to be the time since the infection, we can define an analogous quantity in epidemiology,

$$g(\tau) = \frac{S_0 A(\tau)}{R_0} \,, \tag{3.43}$$

called *generation interval distribution*. In this case this distribution is the probability distribution function for the time from infection of an individual to the infection of a secondary case by that individual. Going back to (3.41) we now have

$$\frac{1}{R_0} = \int_0^\infty g(\tau) e^{-r\tau} \mathrm{d}\tau \,. \tag{3.44}$$

According to this last expression, the shape of the generation interval distribution determines the relation between the basic reproduction number R_0 and the growth rate r. In all the models explored so far, we assumed that both the rate of infection β and the rate of leaving the infectious stage μ were constant. Hence, it follows that the duration of a generation interval is specified as an exponential distribution with mean $Tg = 1/\mu$. Under these assumptions the basic reproduction number is then

$$\begin{aligned} R_0 &= \left(\int_0^\infty \mu e^{-\mu\tau} e^{-r\tau} \mathrm{d}\tau \right)^{-1} = \left(\frac{\mu}{r+\mu} \right)^{-1} \\ &= 1 + rTg \,. \end{aligned} \tag{3.45}$$

This relation between the growth rate and the generation time was already proposed by Dietz in 1976 [210], although only in the specific case of the SIR model. Equation (3.44), however, allows for the calculation of R_0 in more complex scenarios, such as non constant μ [209]. Despite its limitations, this expression is widely used in the literature due to its simplicity. Indeed, Tg is often considered to be simply the inverse of the recovery rate, which is relatively easy to measure. Thus, r can be obtained by fitting a straight line to the cumulative number of infections as a function of time, see (3.40).

There are, however, several problems with this procedure. First, we stated that the exponential growth is valid during the early phase of an outbreak, but there is no way to know how long is that in general. As a consequence, when one fits a straight line to the data, some heuristics have to be used to determine which points to use. Even more, if the dynamics is really fast there might be just a few valid points. Given the stochasticity of the process, this might lead to poor estimates of the growth rate.

Besides, there are some caveats on the exponential growth assumption. For instance, it has been observed that for some diseases such as AIDS/HIV the early growth is sub-exponential [211]. Likelihood based methods in which the early exponential growth is not needed were thus proposed [212, 213]. But even for diseases in which it might be a good approximation, there is the problem of susceptible depletion. Indeed, if the population is infinite, each infected individual will be always able to reach an infinite amount of susceptibles. But this is not true in real situations, forbidding the exponential growth to be sustained for too long. Hence, methods that account for this depletion during the initial phase had also to be developed [214, 215].

It should be clear by now that despite the widespread use of this parameter, it is far from being perfectly understood, specially in the presence of real world data. Yet, we can go one step further and generalize the definition of R_0 to the *effective reproduction number*, $R(t)$.

3.2.2 The effective reproduction number

The effective reproduction number, $R(t)$, is defined as the average number of secondary cases generated by an infectious individual at time t. Hence, we are relaxing the hypothesis of a fully susceptible population that we gave at the beginning of section 3.2.

This parameter is obviously better suited for studying the impact of protection measures taken after the detection of an epidemic, as it can be defined at any time. If $R(t) < 1$, it seems reasonable to say that the epidemic is in decline and may be regarded as being under control at time t. Furthermore, in section 3.1.1 we saw that diseases such as measles have periodic outbreaks and also convey immunity to those who have suffered it. Thus, when a new outbreak starts the population is not completely susceptible, invalidating one of the conditions in the definition of R_0 [216].

To provide a mathematical definition of $R(t)$ [217], we can revisit equation (3.32) and define

$$R(t) = S(t) \int_0^\infty A(\tau) \mathrm{d}\tau \,, \tag{3.46}$$

which leads to

$$R(t) = \frac{S(t)}{S_0} R_0 \,. \tag{3.47}$$

According to this expression, in a closed population, the value of $R(t)$ should monotonically decrease. As expected, this has been observed in computational simulations [216], even in the case of sub-exponential growth [218]. However, this is not always the case if one tries to obtain $R(t)$ from real data [219]. In particular, Walling and Teunis studied the severe acute respiratory syndrome (SARS) epidemic from 2003 and observed several local maxima in the evolution of the effective reproduction number, which they attributed to "super-spread events" in which certain individuals infected unusually large numbers of secondary cases [220]. This was also found in other diseases, signaling that the whole complexity of real systems cannot be completely captured with simple homogeneous models [221]. For this reason, the next section will be devoted to our contribution in the study of the effect that more heterogeneous population distributions have on the reproduction numbers.

3.2.3 Measurability of the epidemic reproduction number

Of course, there had been plenty of diseases, long before humans had been around. But humans had definitely created Pestilence. They had a genius for crowding together, for poking around in jungles, for setting the midden so handily next to the well. Pestilence was, therefore, part human, with all that this entailed.
("Thief of time", Terry Pratchett)

The fundamental role that households play in the spreading of epidemics has been acknowledged for a long time. Early estimations of influenza spreading already showed that the probability of getting infected from someone living in your household or someone from the community were quite different. Even more, it was shown that children were twice more likely to get the infection from the community than adults, signaling that the places that children visit and the own heterogeneity of household members are fundamental in the disease dynamics [222]. In a more recent study, data from a real epidemic in a small semi-rural community was analyzed, with schools added explicitly into the picture. As expected, it was observed that their role is key in the spreading of the disease. But, even more interesting, the authors calculated a reproduction number for each population structure and found it to be smaller or of the order of 1, meaning that for an outbreak to be sustained a complex interplay between those structures must take place [223].

In order to introduce the concept of households into the models analyzed so far, we need to revisit once again the assumption of full homogeneity. Most theoretical approaches in this line, since the seminal work of Ball et al. in 1997 [224], have focused on what is known as models with two levels of mixing. In these models, a local homogeneous mixing in small environments (such as households) is set over a background homogeneous mixing of the whole population. This can be further extended by adding other types of local interactions, such as schools or workplaces. An individual can thus belong at the same time to two or more local groups. For this reason, they are also known as overlapping group models [225]. This allows for the definition of several basic reproduction numbers, one for the community and the rest for local interactions, which in turn can be used to devise efficient vaccination strategies [226, 227]. Other studies have also proposed that the generation time can differ from within households and the community [228].

However, theoretical studies have been mostly focused on the early phase of the epidemics because it is more mathematically tractable. Yet, we have seen in the previous section that $R(t)$ can provide very important insights to understand the dynamics of real diseases. For this reason, statistical methods have been developed to analyze $R(t)$ [138, 229]. Unfortunately, for these methods, disentangling the role that each structure of the system plays is challenging due to the lack of microscale data on human contact patterns for large populations. Note also that due to the typically small size of households, stochastic effects are highly important.

In this work our objective is to shed some light into the mechanisms behind disease dynamics in heterogeneous populations. To do so, we study the evolution of $R(t)$ and Tg with data-driven stochastic micro-simulations of an influenza-like outbreak on a highly detailed synthetic population. The term "micro" refers to the fact that we will keep track of each individual in the population, allowing us to reconstruct the entire transmission chain. The great advantage of this method is that it allows for the computation of $R(t)$ from its own epidemiological definition, without requiring any mathematical approximation.

Our synthetic population is composed by 500,000 agents, representing a subset of the Italian population. This population model, developed by Fumanelli et al. [145], divides the system into the four settings where influenza transmission occurs, namely households, schools, workplaces and the general community [230]. Henceforth we will refer to these settings as *layers* for the similarity of this construction to the multilayer networks we saw

in section 2.1.3[5]; a visualization of the model is shown in 3.6A. The household layer is composed by n_H disconnected components, each one representing one household. The amount of individuals inside each household is determined by sampling from the actual Italian household size distribution, as well as their age. Then, by sampling from the multinomial distribution of schooling and employment rates by age, each individual might be also assigned to a school or workplace. Both the number and size of workplaces and schools is also sampled from the actual Italian distribution. As in the household layer, each of the n_S schools and n_W workplaces are disconnected from the rest in their respective layers. Lastly, all individuals are allowed to interact with each other in the community layer, encapsulating the background global interaction. To highlight the heterogeneity of the system, in figure 3.6B the size distribution of the places each individual belongs to is shown. Note that while most households contain 2-3 individuals and most schools are close to 1,000 students, workplaces cover a much wider range of sizes.

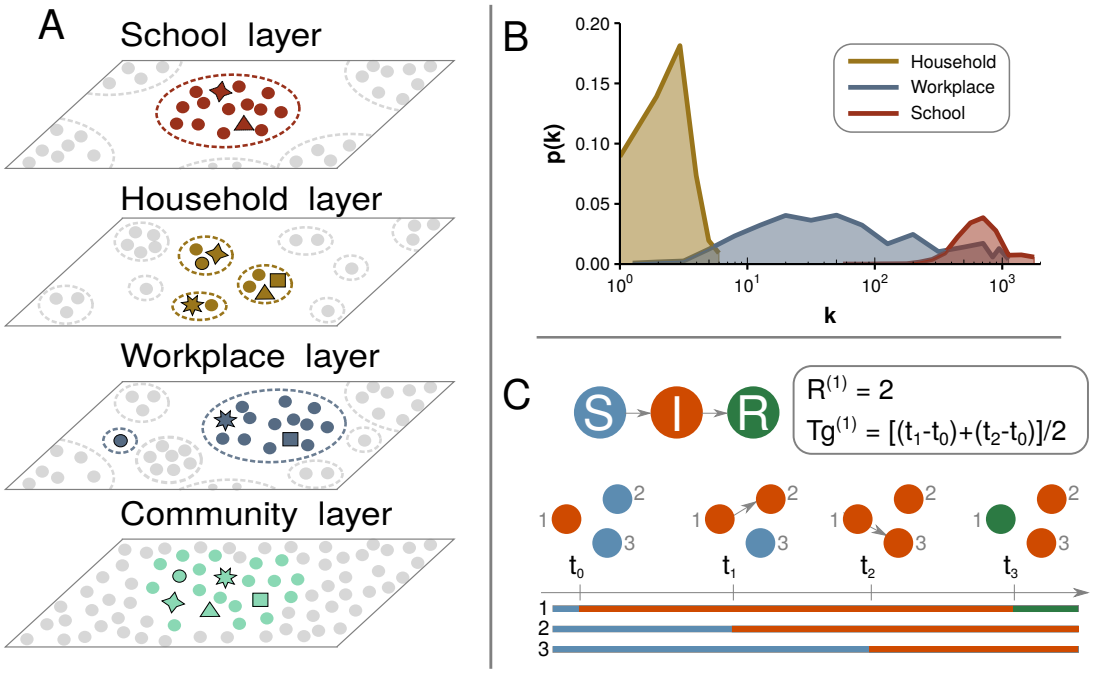

Figure 3.6: Model structure of a synthetic population organized in schools, households and workplaces. A) Visualization of the overlapping system, with individuals being able to interact locally in multiple contexts. B) Distribution of the structure size each individual belongs to. C) Illustration of the transmission process with an example of how to calculate the reproduction number and generation interval.

The influenza-like transmission dynamics are defined through the susceptible, infected, removed (SIR) compartmental model that we have been analyzing under diverse assumptions. We simulate the transmission dynamics using a stochastic process for each individual, keeping track of where she contracted the disease, who is in contact with and so on. In order to resemble an influenza-like disease, the local spreading power in each layer is calibrated in such a way that the fraction of cases in the four layers is in agreement with literature values (namely, 30% of all influenza infections are linked to transmission occurring in the household setting, 18% in schools, 19% in workplaces and 33% in the community [231]).

[5]Even though we have not discussed yet on how to introduce networks in epidemic modeling, note that an homogeneous population is equivalent to a complete network in which every node is connected to every other node. Hence, these overlapping structures can also be regarded as a multiplex network in which each node can be in more than one layer

Hence, the probability that individual j infects i in layer l is

$$\beta = w_l \,, \tag{3.48}$$

as long as j is infected, i is susceptible and both belong to the same component in layer l. Moreover, we set the values of w_l such that the basic reproduction number is $R_0 = 1.3$ [232]. Finally, the removal probability, μ, is set so that the removal time is 3 days [233].

The epidemic starts with a fully susceptible population in which we set just one individual as infected. Thanks to the microscopic detail of the model, we can thus compute the basic reproduction number directly as the number of infections that said first individual produces before recovery. Its counterpart over time, the effective reproduction number, $R(t)$, is measured using the average number of secondary cases generated by an infectious individual at time t. Similarly, we also define the effective reproduction number in layer l, $R_l(t)$, as the average number of secondary infections generated by a typical infectious individual in layer l:

$$R_l(t) = \frac{\sum_{i\in\mathcal{I}(t)} D_l(i)}{|\mathcal{I}(t)|} \,, \tag{3.49}$$

where $\mathcal{I}(t)$ represents the set of infectious individuals that acquired the infection at time t and $D_l(i)$ the number of infections generated by infectious node i in layer l with $l \in L = \{H, S, W, C\}$. With this expression we can obtain the overall reproductive number as

$$R(t) = \sum_{l\in L} R_l(t) \,. \tag{3.50}$$

The generation time Tg is defined as the average time interval between the infection time of infectors and their infectees. Hence, analogously to the reproduction number, we define the generation time in layer l as

$$Tg_l = \frac{\sum_{i\in\mathcal{I}(t)} \sum_{j\in\mathcal{I}'_l(i)} (\tau(j) - t)}{\sum_{i\in\mathcal{I}(t)} D_l(i)} \,, \tag{3.51}$$

where $\mathcal{I}'_l(i)$ denotes the set of individuals that i infected in layer l and $\tau(j)$ is the time when node j acquired the infection. Therefore, the overall generation time $Tg(t)$ reads

$$Tg(t) = \frac{\sum_{l\in L} \sum_{i\in\mathcal{I}(t)} \sum_{j\in\mathcal{I}'_l(i)} (\tau(j) - t)}{\sum_{l\in L} \sum_{i\in\mathcal{I}(t)} D_l(i)} \,. \tag{3.52}$$

A schematic illustration of the transmission dynamics is shown in figure 3.6C. In that case, individual 1 gets infected at $t = t_0$, while individuals 2 and 3 are still susceptible. During the course of her disease individual 1 infects individual 2 at $t = t_1$ and individual 3 at $t = t_2$ before finally getting recovered at $t = t_3$. Thus, her reproduction number is equal to 2 and her generation time is 1.5, supposing that $t_{i+1} - t_i = 1$. In fact, in our simulations we will set $\Delta t = 1$ day. Due to the stochasticity of the process, each realization might result in an outbreak of different length. Hence, the time evolution of each simulation is aligned so that the peak of the epidemic is exactly at $t = 0$. The results for the reproduction number and generation time are shown in figure 3.7.

We find that $R(t)$ increases over time in the early phase of the epidemic, starting from $R_0 = 1.3$ to a peak of about 2.5 (figure 3.7A). In contrast, in the homogeneous model (dashed line), which lacks the typical structures of human populations, $R(t)$ is nearly constant in the early epidemic phase and then rapidly declines before the epidemic peak ($t = 0$), as predicted by the classical theory. The non-constant phase of $R(t)$ implies that R_0 loses its meaning as a fundamental indicator in favor of $R(t)$. In figure 3.7B we show an analogous analysis of the measured generation time in the data-driven model. In this

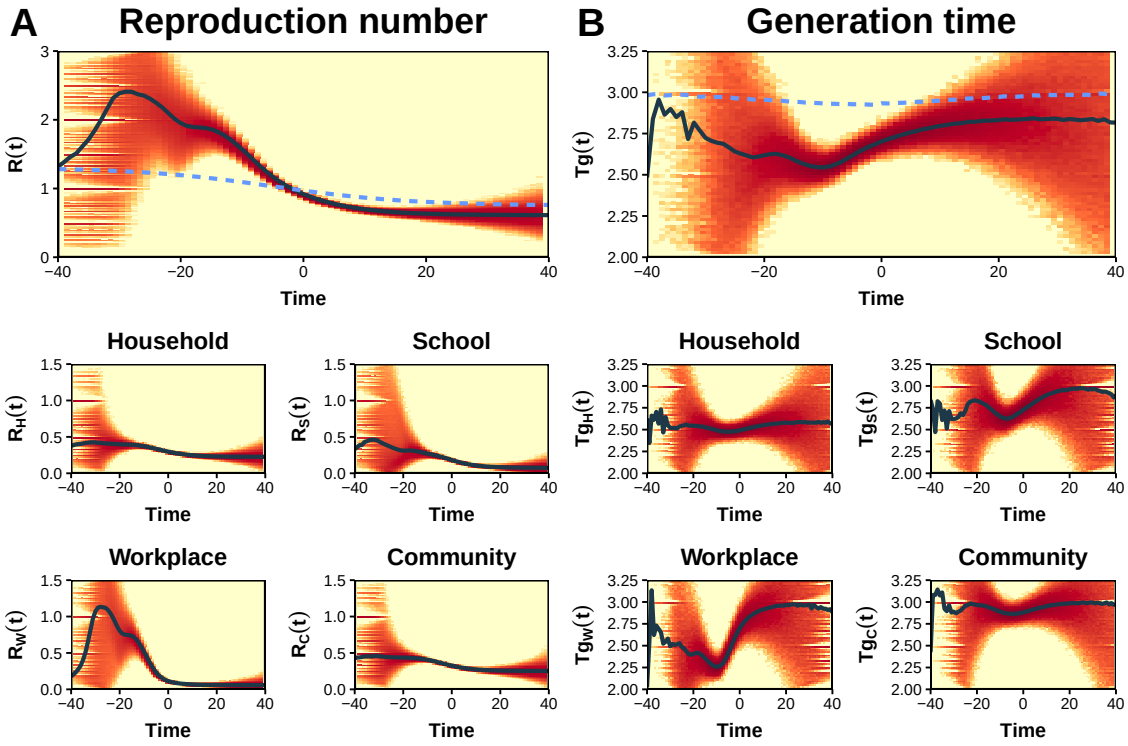

Figure 3.7: Fundamental epidemiological indicators. A) Top: mean $R(t)$ of data-driven model (solid line) compared to the solution under a completely homogeneous population (dashed line). The colored area shows the density distribution of $R(t)$ values obtained in single realizations of the model. Bottom: the reproductive number is broken down in the four layers. B) As A but for the generation time. In all cases the simulations have been aligned at the peak of the epidemic.

case, we find that Tg is considerably shorter than the infectious period (3 days), with a more marked shortening near the epidemic peak. Once again, in the homogeneous model (dashed line) the behavior predicted by the classical theory is recovered.

A closer look at the transmission process in each layer helps to understand the origin of the deviations from classical theory. Specifically, we see that $R(t)$ tends to peak in the workplace layer, and to some extent also in the school layer. In the community layer, on the other hand, the behavior is much closer to what is expected in a classical homogeneous population. We also find that Tg is remarkably shorter in the household layer than in all other layers. This could simply be due to a depletion of susceptibles. To illustrate this, suppose that an infected individual in a household of size 3 infected one of the other two. Then, during the next time step both will compete to infect the last susceptible, something that does not happen in large populations. This would lead to a shorter generation time simply because she is unable to infect other members, even if she is still infected. This evidence calls for considering within household competition effects when analyzing empirical data of the generation time.

To further understand the reasons of the diverse trends observed in each layer, in figure 3.8 we analyze the effect that the size of the components has on the dynamics. In figure 3.8A we study the attack rate (final fraction of removed individuals, i.e., individuals that suffered the infection at some point) as a function of the site size, distinguishing the three layers. The results indicate that the spreading is much more important in large buildings, but we know that they are scarce (see figure 3.6B). Hence, it seems that the initial growth of the epidemic might stop once the big components have been mostly infected. This is

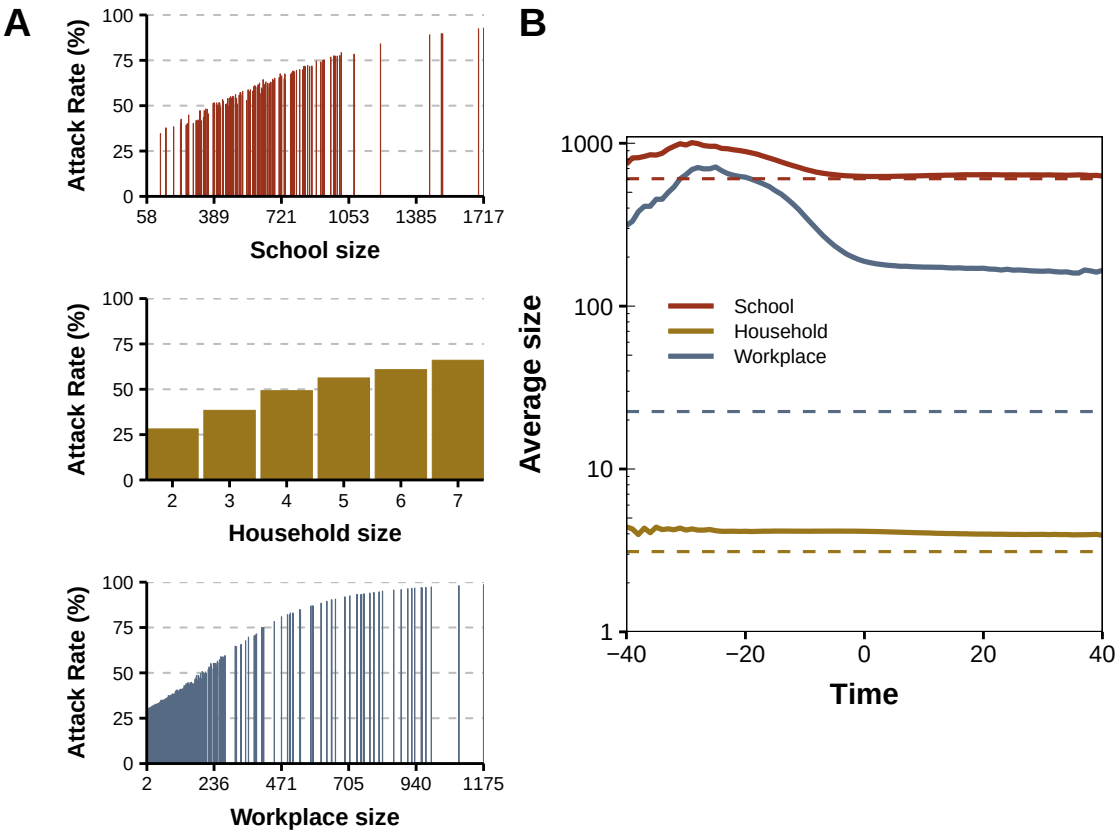

Figure 3.8: Attack rate as a function of site size. A) Fraction of individuals belonging to each place that contracted the disease, not necessarily in said setting. B) Solid line: average size of places in which there is at least one new infection in each time step, broken down in three layers. Dashed line: expected size if there is at least one infection in every place.

corroborated in figure 3.8B, where the average size of buildings with at least one infection is shown. The situation is thus clear. In the classical model not only it is assumed that all population is initially susceptible, but also that it is in contact with the first infected since the beginning. In heterogeneous populations, however, the first infected individual has only a handful of local contacts, diminishing its infectious power. Then, as the epidemic progresses more and more susceptibles enter into play, increasing the amount of individuals that can be infected. Yet, sooner or later the components will run out of susceptibles, even if there is still a large fraction available in the rest of the system. This, in turn, leads to a more abrupt descent than what is expected in the classical approximation.

These results clearly highlight how the heterogeneity of human interactions (i.e., clustering in households, schools and workplaces) alters the standard results of fundamental epidemiological indicators, such as the reproduction number and the generation time. Furthermore, they call into question the measurability of R_0 in realistic populations, as well as its adequacy as an approximate descriptor of the epidemic dynamics. Lastly, our study suggests that epidemic inflection points, often ascribed to behavioral changes or control strategies, could also be explained by the natural contact structure of the population. Hopefully, this analysis will open the path to developing better theoretical frameworks, in which some of the most fundamental assumptions of epidemiology have to be revisited.

3.3 The epidemic threshold fades out

In epidemiology attention has historically been restricted to biological factors. We began this chapter stating that individuals were just indistinguishable particles interacting according to the mass action law. However, throughout the following sections, we have shown that when this oversimplification is relaxed many interesting phenomena arise. In this section we shall go one step further and completely remove what we called assumptions 3 and 4: mass action and indistinguishability. To distinguish individuals, we will assign to each of them an index, i. Then, we will allow individuals to spread the disease only to those with whom they have some kind of contact (e.g. they are friends), which we will encode in links. In other words, we are finally going to introduce networks into the picture.

It is rather difficult to establish the origin of what we may call disease spreading on networks or network epidemiology. Probably, one of the earliest attempts is the work by Cochran, in 1936 [234], in which he studied the propagation of a disease in a small plantation of tomatoes. Although his work might be better described as statistical analysis, the reason to consider it one of the precursors of the spreading on networks is that, as Kermack and McKendrick had done roughly 10 years before, his assumptions were all mechanistic rather than based on the knowledge of the particular problem. This is clearly seen in how he introduced the model: *"We suppose that in the first [day] each plant in the field has an equal and independent chance p of becoming infected, and that in the second [day] a plant which is next to a diseased plant has a probability s of being infected by it, healthy plants which are not next to a diseased plant remaining healthy"*. In modern terminology, the plants were arranged in a *lattice* structure and could infect their first neighbors with probability s. The assumptions are particularly strong because he knew that the disease was propagated by an insect, but decided to create a very general model.

During the next couple of decades, lattice systems were quite popular in physics, geography and ecology. Then, in the 1960s the interest in studying the spatial spread of epidemics started to grow (see the introduction of [235] for a nice overview) with three main approaches. In the first, the agents that could be infected were set in the center of a certain tessellation of space and could only infect/be infected by their neighbors. This is the closest approach to the modern study of epidemics on networks, but it was not so popular as it was mainly used to study plant systems [236]. The most popular approach was to distribute individuals continuously in space with some chosen density. This led to the study of diffusion processes, focusing on the interplay between velocity and density [237]. The third approach was based on what we briefly defined in section 3.1.1 as metapopulations. Recall that in a metapopulation individuals are arranged in a set of sub-populations. Within each sub-population usually homogeneous mixing is applied and it is also allowed to have individuals migrating from one sub-population to another. Hence, the idea was to simulate the fact that people (or animals) live in a certain place where they can contract the disease, and then travel to a different area and spread it to its inhabitants [183]. Note, however, that in none of these methods we are taking into account any sociological factors of the population.

In the beginning of the 1980s some results pointing into the direction of introducing more complex networks started to appear. In particular, von Bahr and Martin-Löf in 1980 [238] and Grassberger in 1983 [239] (in the context of percolation) showed that the classical SIR model on a homogeneous population could be related to the ER graph model that we saw in section 2.1.2. Indeed, suppose that we have a set of individuals under the homogeneous mixing approach and simulate an epidemic. Next, if an individual i infects another individual j, we establish a link between them. If the probability of this event is really low, the epidemic will not take place. Conversely, for large probabilities most nodes will be randomly connected. This is precisely how we defined the ER model, with the only difference being that the probability p of establishing a link will be a function of both the

probability of infecting someone, β, and that of getting recovered, μ. Note, however, that they did not implement a disease dynamics on a network, but rather extracted the network from the results of the dynamics.

Roughly at the same time, the gonorrhea rates rose precipitously. To understand the dynamics of this venereal disease, it was recognized that some sort of nonrandom mixing of the population had to be incorporated to the models. The first attempts were based on separating the contact process and the spreading process [240]. Indeed, going back to the definition of the SIR model, we defined the rate of infectivity (3.5) as $\phi(\tau) = \beta/N$ (under the frequency approach). We can simply define $\beta = c\beta'$ where c is the contact rate between individuals and β' the probability of spreading the disease given that a contact has taken place. For simplicity, we can remove the apostrophe and simply write $\phi(\tau) = c\beta/N$. With this definition the epidemic threshold would read

$$\frac{c\beta}{\mu} > 1 \Rightarrow \frac{\beta}{\mu} > \frac{1}{c}\,. \tag{3.53}$$

This expression is giving us a very powerful insight about how to combat diseases. Supposing that β and μ are fixed, as they mostly depend on the characteristics of the pathogen, the best way to prevent and epidemic is to reduce the number of contacts as much as possible. A similar result was obtained in the context of vector-borne diseases, in which the number of vectors (e.g. mosquitoes) play the role of the number of contacts [241].

The next step was the introduction of mixing matrices, the same approach that we followed to incorporate age into the SIR model in section 3.1.1. Recall that the idea was to divide the population into smaller groups according to some characteristics (e.g. gender or age) and establish some rules governing the interaction between those groups encoded in a contact matrix (hence the name of mixing matrices). Typically, both the group definitions and the mixing function were very simple. In the context of venereal diseases, the most common characteristic used to form groups was activity level. This approach was popularized by Hethcote and Yorke in 1984 [242] in their modeling of gonorrhea dynamics using the core group: a group of highly sexually active individuals who are efficient transmitters interacted with a much larger noncore group. They showed that with less than 2% of the population in the core group, this model lead to 60% of the infections to be caused directly by core members. Yet, the world of epidemiology was about to be shaken by a new virus that would defy all these assumptions, HIV.

3.3.1 The decade of viruses

The emergence of HIV in the early 1980s forced scientists to pay even more attention to the role of selective mixing. In 1986, in one of the earliest attempts to model this disease [243], Anderson and May summarized the challenges that sexually transmitted diseases (STDs) presented in contrast to other more common infectious diseases such as measles:

1. For STDs only sexually active individuals need to be considered as candidates of the transmission process, in contrast to simple "mass action" transmission models.

2. The carrier phenomenon, in which certain individuals harbor asymptomatic infection, is important for many STDs.

3. Many STDs induce little or no acquired immunity. In the case of HIV, the situation is probably more complex since persistence without symptoms might be lifelong.

4. The transmission of most STDs is characterized by a high degree of heterogeneity generated by great variability in sexual habits among individuals within a given community.

They concluded their introduction with a sentence that we would like to highlight, although its full meaning will not be understood until the end of this section: "This set of characteristics - *virtual absence of a threshold density of hosts for disease-agent persistence*, long-lived carriers of infection, absence of lasting immunity, and a great heterogeneity in transmission - give rise to infectious diseases that are well adapted to persist in small low-density aggregations of people".

Clearly, the homogeneous mixing approach was not valid anymore and heterogeneity had to enter into play. However, there was a huge problem, they did not have any data. Up to that point, epidemiologists had focused on the biological factors of diseases, completely ignoring the heterogeneity of human interactions. Yet, the importance that they had in HIV transmission sprouted a series of studies that would finally shed some light in the contact patterns of human populations. In particular, during the first years of the HIV epidemic, it was observed that homosexual males accounted for 70-80% of the known cases, and thus most efforts were devoted to study said community. The earliest studies found that the distribution of the number of sexual partners had a high mean, but also a very large variance[6].

This observation led them to the formulation of a model that could account for this huge heterogeneity (a variance much larger than the mean). They focused on a closed population of homosexual males and divided it into sub-groups of size N_i, whose members on average had i new sexual partners per unit time. Under these assumptions, the SIR model reads

$$\begin{cases} \dfrac{\mathrm{d}S_i(t)}{\mathrm{d}t} = -i\lambda S_i \\ \dfrac{\mathrm{d}I_i(t)}{\mathrm{d}t} = i\lambda S_i - \mu I_i \\ \dfrac{\mathrm{d}R_i(t)}{\mathrm{d}t} = \mu I_i(t) \end{cases}, \tag{3.54}$$

where the infection probability per partner, λ, was given by

$$\lambda = \beta \frac{\sum_i i I_i}{\sum i N_i}. \tag{3.55}$$

At first glance it might seem that the model has not changed that much, as it is just the standard SIR model with i contacts. However, considering that the population is divided into several groups with heterogeneous contact patterns leads to a result that 15 years latter would become one of the cornerstones of network science. According to Anderson and May, in this model the early rate of exponential growth, Λ, is defined as

$$\Lambda = \beta \frac{\langle i^2 \rangle}{\langle i \rangle} - \mu. \tag{3.56}$$

Hence, the epidemic only grows if

$$\frac{\beta}{\mu} > \frac{\langle i \rangle}{\langle i^2 \rangle}, \tag{3.57}$$

which we can arrange to look like equation (3.53),

$$\frac{\beta}{\mu} > \frac{1}{c'} \quad \text{with } c' = \frac{\langle i^2 \rangle}{\langle i \rangle}. \tag{3.58}$$

Thus, for epidemiological purposes, the effective value of the average number of new partners per unit time is not the mean of the distribution, but rather it is the ratio of

[6] Actually, the shape of the distribution was the same as the ones we saw in figure 2.10 when we studied age contact patterns. The high heterogeneity of human interactions is clearly not restricted to sexual activity.

the mean square to the mean. In other words, this result reflects the disproportionate role played by individuals in the most active groups, who are both more likely to acquire infection and more likely to spread it [244].

In parallel, Boltz, Blancard and Krüger started to develop in 1986 a series of complex computational models that allowed them to introduce many more factors into the dynamics [245]. They said that the principal weaknesses of the standard epidemiological models when applied to HIV infection were:

- They describe, through mass action dynamics, permanent potential contact between all members of the groups involved.
- The behavior is uniform over each group. To account for nonuniform behavior, a subdivision into sub-groups has to be done at the prize of a higher dimensionality of the systems of differential equations.
- They cannot directly take into account partners.
- Time delays, age dependencies and time dependent rates are not easily incorporated.
- They do not represent the true contact structure of the population.

Hence, they proposed to use *"random graphs and discrete time stochastic processes to model the epidemic dynamics of sexually transmitted diseases"*. This is one of the earliest (if not the first) attempts to clearly study disease dynamics on networks.

Their models were highly detailed. For instance, they considered eleven groups of individuals: homosexual males, bisexual males, heterosexual males, heterosexual females, heterosexual females having contact with bisexual males, male intravenous drugs users, female intravenous drug users, prostitutes, prostitutes who are also intravenous drug users and hemophiliacs (clearly way beyond the simple model of only homosexuals that was being studied analytically in those days). Even more, they could track the behavior of single individuals. But this level of detail also posed the problem of acquiring a huge amount of data to parameterize the models that, admittedly, they did not have at that time. In any case, this represented a huge step forward in the direction of highly detailed computational models, as the one we presented in section 3.2.3.

Yet, data was about to arrive. Already in 1985 Klovdahl [246], inspired by the networks that sociologist had been studying since the 1930s, studied the information provided by a small sample of 40 patients with AIDS and reconstructed their social network. He proposed that a "network approach" could be a key element to understand the nature and spread of the disease. Some years later, in 1994 [247], together with some collaborators, he designed a larger experiment to obtain the social network of a small city in Colorado. Their approach was to initially target individuals with higher risk of contracting HIV, such as prostitute women and injecting drug users, and trace as many contacts (of any kind) as they could. Their results showed that a lot of people were much closer to HIV infection than expected, implying that a small change (e.g. reducing condom use) could quickly reach individuals who were not directly connected with people infected with HIV. To demonstrate the importance of a network perspective in epidemiology they gave a very simple example: suppose that one individual is infected with HIV and reports sexual relationships with two people and a social relation with another one. Commonly, health professionals would only worry about the first two, disregarding the latter. However, if that third individual happens to by highly *central* in the network (i.e., having a large degree or betweenness, using the terminology of chapter 2), an eventual infection could lead to an "explosion" of disease. Hence, their claim was that under a network perspective addressing the distance to the disease and the centrality of individuals was as important as being actually infected with HIV.

Clearly the concept of networks was starting to gain momentum in epidemiology, although it was not always clear what was part of the epidemic model and which elements came just from the topology of the network (see [57] and the references therein). A noteworthy exception is the work by Andersson in 1997 [248] in which he studied an epidemic process on a random graph, this time from an analytical point of view, and concluded that the basic reproduction number was

$$R_0 = p_1 \left(\frac{\langle X^2 \rangle}{\langle X \rangle} - 1 \right) , \tag{3.59}$$

where X denotes the number of links a certain node has. Thus, there is a clear dependency on the topology of the networks, regardless its specific shape. The similarity with the expression that Anderson and May obtained 10 years before is clear (equation (3.57)), the only difference is the -1 factor (and the fact that he set $\mu = 1$). The reasons why these two expressions are so similar, and why this expression contains a -1 will come clear in a moment. But first, we need to go back in time a little bit to have a look at what was happening in the emerging field of *cybersecurity*.

On November 3, 1983, the first computer virus was conceived as an experiment to be presented at a weekly seminar on computer security. A virus, in this context, is a computer program that can "infect" other programs by modifying them to include a possibly evolved copy of itself. Thus, if the programs communicate with other computers, the viruses can spread through the whole network. During this decade, the internet started to grow, connecting more and more computers through a huge network. Hence, viruses imposed a clear threat [249]. Soon after, in 1988, Murray proposed that the propagation of computer viruses could be studied using epidemiology tools[7] [250]. Then, in 1991, Kephart and White tried to apply epidemic models to study the propagation of viruses, but quickly realized that the homogeneous mixing approach was not suited for computers, as they were connected through networks [251]. Hence, they applied the model on a random directed graph and showed that the probability of having an epidemic depended on the connectivity of the network. When connectivity was high, they recovered the classical results of Kermack and McKendrick obtained under the homogeneous mixing approach. Conversely, when connectivity was really low the probability quickly decreased. However, they were unable to mathematically show this and had to rely on simulations (we can hypothesize that if they had been aware of the results found for venereal diseases they might had been able to do it, because their observation is essentially explained by equation (3.53), but it seems highly unlikely that computer scientists were interested in such a specific sub-field back then). Yet, there was something odd. During the 1990s Kephart and collaborators collected virus statistics from a population of several hundred thousand PCs. They observed that there was a huge amount of viruses that survived for a really long time, but also that their spreading was quite low and reached only a small fraction of the population, contrary to the expected exponential growth predicted by epidemiology (see section 3.2.1) [252]. The only possibility, according to the theory, was that the infection and recovery parameters associated with each virus were really close to the epidemic threshold, yielding low growth but still some persistence. Yet, this regularity seemed highly unlikely and they advocated to further account the network patterns as possible reasons behind this discrepancy.

With this context, we will now finally be able to realize why disease dynamics has been one of the main areas of research in network science and, at the same time, why

[7] Despite the great leap forward that his proposal represented, one could argue that he might have been a bit naive when he concluded the paper saying: *"All this having been said, I am sanguine. God is still in his heaven. The environment is generally benign. The community is resilient. Most individuals are acceptably polite, orderly and well behaved. On the list of vulnerabilities in our complex society, this one is distinguished primarily by its novelty. Unlike some of the more intractable ones, this one will yield to good will. In the face of genuine evil intent, I prefer it to plastic explosives in power plants."*

networks became so successful in a relatively short period of time. In 1999 Albert et al. measured the topology of the World Wide Web, considering each document as a node and their hyperlinks to other documents as links. Surprisingly, the degree distribution of such network did not follow the Poisson distribution of random graphs but rather a power law distribution [253]. In section 2.1.2 we defined scale-free networks as those networks whose degree distribution followed a power law. Moreover, we showed that if the exponent of the distribution is in the interval $\gamma \in (2, 3)$, the average of the distribution is finite but the second moment of the distribution diverges if the system size tends to infinity.

In 2001 Pastor-Satorras and Vespignani studied the behavior of an SIS model on a network, in an attempt to answer the questions posed by Kephart [254]. The SIS model, which we have not talked about yet, is simply the SIR model but once an individual is cured instead of going into the removed compartment she is sent back to the susceptible one. Hence, the equation describing the process is

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \frac{\beta}{N} S(t)I(t) - \mu I(t)\,, \tag{3.60}$$

with $S(t) + I(t) = N$. However, as they wanted to apply it on a network they had to keep track of the state of each node, yielding a system of N coupled equations,

$$\frac{\mathrm{d}p_i(t)}{\mathrm{d}t} = \beta[1 - p_i(t)] \sum_j a_{ij} p_j(t) - \mu p_i(t)\,, \tag{3.61}$$

where $p_i(t)$ denotes the probability that node i is infected at time t. This system does not have a closed form solution and, besides, it depends on the adjacency matrix of the network (the term a_{ij}). Hence, they followed a mean-field approach and supposed that the behavior of nodes with the same degree k would be similar. Under this assumption, the system is reduced to

$$\frac{\mathrm{d}\rho_k(t)}{\mathrm{d}t} = \beta[1 - \rho_k(t)]k\Theta - \mu\rho_k(t)\,, \tag{3.62}$$

where ρ_k is the density of infected individuals with degree k and Θ is the probability that any given link points to an infected node.

The stationary solution (i.e. $d_t\rho_k(t) = 0$) yields

$$\rho_k = \frac{k\beta\Theta}{\mu + k\beta\Theta}\,, \tag{3.63}$$

denoting that the higher the node connectivity, the higher the probability to be infected. Now, in the absence of correlations the probability that a randomly chosen link in the network is attached to a node with s links is proportional to $sP(s)$, where $P(s)$ denotes the probability of having degree s. Hence,

$$\Theta = \sum_k \frac{kP(k)\rho_k}{\sum_s sP(s)}\,. \tag{3.64}$$

Combining (3.63) and (3.64),

$$\Theta = \sum_k \frac{kP(k)}{\sum_s sP(s)} \cdot \frac{k\beta\Theta}{\mu + k\beta\Theta}\,. \tag{3.65}$$

Besides the trivial solution $\rho_k = 0$, $\rho_k > 0$ will be a solution of the system as long as

$$\frac{\beta}{\mu}\frac{\langle k^2\rangle}{\langle k\rangle} > 1\,, \tag{3.66}$$

which can be identified as the basic reproduction number of the system. Better still, we can put all parameters relating to the disease in one side and all the ones coming from the network in the other so that

$$R_0 = \frac{\beta}{\mu} > \frac{\langle k \rangle}{\langle k^2 \rangle} \,. \tag{3.67}$$

Hence, the epidemic threshold is not 1 anymore. Instead, it is a function of the connectivity of the network. But remember that the pourpose of the model was to study disease propagation on computer networks which, according to Albert et al., were not only scale-free but also had an exponent of $\gamma = 2.45$ [253]. Thus, in the internet $\langle k^2 \rangle \to \infty$, implying that equation (3.67) is actually

$$R_0 > \frac{\langle k \rangle}{\langle k^2 \rangle} \to 0 \,. \tag{3.68}$$

In other words, the epidemic threshold fades out. Moreover, two months later Liljeros et al. showed that the network of human sexual contacts was also scale free [255].

This result is the answer to all the questions that have arisen throughout this section. First, it explains why so many computer viruses were able to persist without growing exponentially. Indeed, if the epidemic threshold had been 1 then they all had to be really close to 1. However, as it tends to 0, they can be anywhere between 0 and 1, be able to infect a macroscopic fraction of the population and at the same time doing it slowly. It is also worth highlighting that from very different approaches Anderson and May obtained essentially the same result (3.57). The reason is simply that in the mean-field approximation we have neglected the connections and only considered groups of nodes with k neighbors, which is equivalent to Anderson and May groups of individuals making i new sexual partners per unit time. Even more, if we add the fact that the network of sexual contacts is scale free, we finally understand the reason behind the *virtual absence of a threshold* they were talking about in the 1980s. Furthermore, this result refutes the hypothesis of the core-noncore gonorrhea model. In order to have two distinct groups, the sexual contact network should have a binomial degree distribution, but it does not. Nevertheless, the idea was not completely wrong as the role of the core is played by the nodes with large degree, the *hubs* of the network.

Furthermore, note that to obtain expression (3.67) the only property of the network that we have used is that it is uncorrelated. Whence, it can be used to study any network that we desire. In particular, in the case of random graphs the degree distribution is Poisson, implying that $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$. Thus, for those networks the epidemic threshold is simplified to

$$R_0 = \frac{\beta}{\mu} > \frac{1}{\langle k \rangle + 1} \,, \tag{3.69}$$

which is the result obtained by the earliest studies of gonorrhea propagation (3.53) (except for the $+1$, but its role will be elucidated in a moment). In other words, we can say that the problem in those models is that they implicitly assumed a random contact network when they should have used a scale free network instead. In figure 3.9 we show a comparison of the final size of the epidemic as a function of R_0 between ER networks, SF networks and the homogeneous mixing approach.

Note that most of the ideas had been around for years, but did not get that much attention. We could argue that what really made the difference in the case of the work by Pastor-Satorras and Vespignani was the use of data. Indeed, many theoretical approaches can be proposed, but without experimental data it is not possible to really gauge their importance. Thus, once again, this highlights how incorporating more data into already existing models can make the difference.

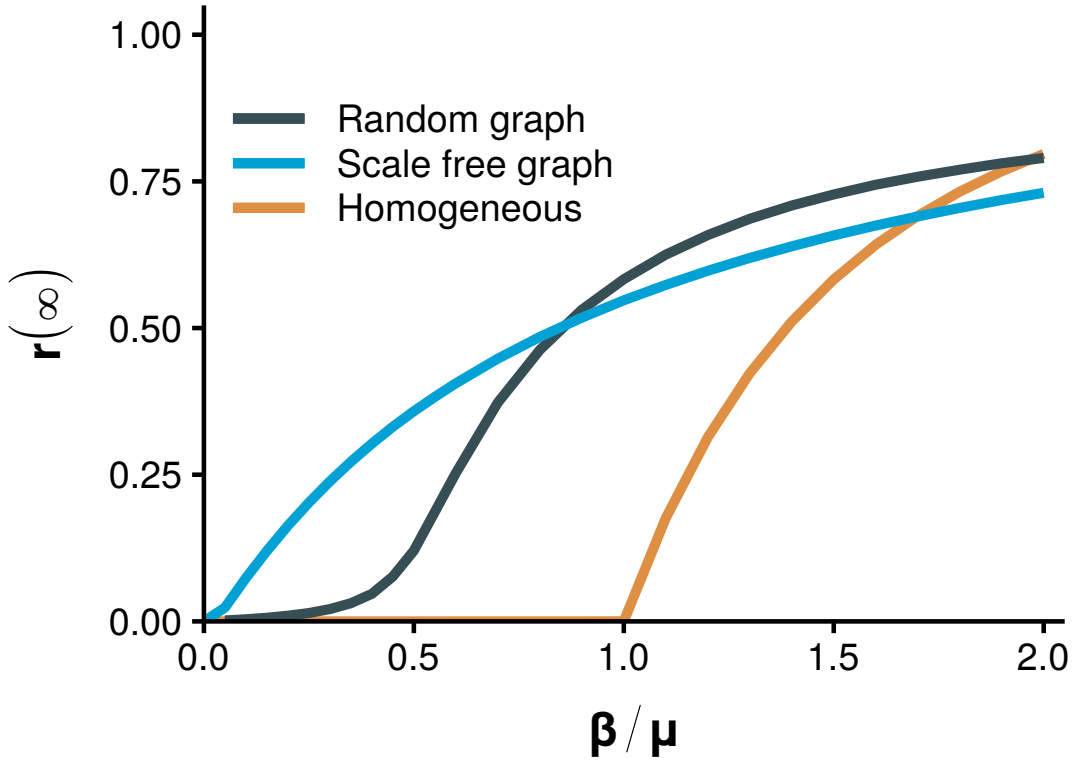

Figure 3.9: Epidemic threshold and topology. Total fraction of recovered individuals in equilibrium conditions as a function of β/μ. In the homogeneous mixing approach the epidemic threshold is 1. When the SIR model is implemented in a random network with $\langle k\rangle = 3$ the epidemic threshold is 1/3 (3.71). Conversely, in a SF network with $\langle k\rangle = 3$ and $\langle k^2\rangle = 113$ the epidemic threshold is 0.03. Note that the size of the network is $N = 10^4$, with a maximum degree of $k_{\max} = \sqrt{N}$ to avoid correlations (see chapter 2). Hence, the threshold does not vanish completely, as it is supposed to be 0 only the limit of $N \to \infty$.

To conclude, we should address why the equation obtained by Anderson and May for the SIR model did not have a -1 factor (3.57), the one from Andersson also for the SIR model did (3.59) and the one from Pastor-Satorras and Vespignani for the SIS model did not (3.67). The first observation is that in the SIS model on a network, it is possible that if i infects j, then i recovers and j infects i. This cannot happen in the SIR model, so for a node of degree k only $k-1$ links can transmit the disease,

$$\Theta = \sum_k \frac{(k-1)P(k)\rho_k}{\sum_s sP(s)}\,, \tag{3.70}$$

leading to the threshold [256]

$$R_0^{\text{SIR}} > \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}\,. \tag{3.71}$$

Which explains the discrepancy between (3.59) and (3.67). Lastly, note that in the case of Anderson and May they did not consider a fixed structure, such as a network, but rather that at each time step every individual would seek i new sexual partners. Thus, the correction accounting where the disease came from does not apply to their model.

3.3.2 The generating function approach

There are multiple techniques that can be used to solve the system (3.61) and obtain the value of the epidemic threshold (see [159, 256] for a review). In this section we will

describe an approach introduced by Newman in 2002 inspired by percolation [257], based on the use of generating functions, that is specially suited to analyze directed networks. This is the methodology that we will use in section 3.3.3 to study disease propagation in directed multiplex networks.

Although for the moment we are only interested in undirected networks, we will introduce the methodology considering that the networks might have both directed and undirected links [258], as this will be the approach used in section 3.3.3. Hence, suppose that the probability that a random node of the network has j incoming links, l outgoing links and m undirected links is p_{jlm}. The generating function of the distribution will be

$$G(x,y,z) = \sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} p_{jlm}x^j y^l z^m \,. \tag{3.72}$$

As long as p_{jlm} is normalized this function has the property $G(1,1,1) = 1$ and

$$\langle k_d\rangle = \frac{\mathrm{d}G(1,1,1)}{\mathrm{d}x} \equiv G^{(1,0,0)}(1,1,1)\,, \tag{3.73}$$

where $\langle k_d\rangle$ is the average number of incoming links in the network. As for any incoming link there has to be also an outgoing link,

$$\langle k_d\rangle = \frac{\mathrm{d}G(1,1,1)}{\mathrm{d}y} \equiv G^{(0,1,0)}(1,1,1) = G^{(1,0,0)}(1,1,1)\,. \tag{3.74}$$

Lastly, for the undirected links

$$\langle k_u\rangle = \frac{\mathrm{d}G(1,1,1)}{\mathrm{d}z} \equiv G^{(0,0,1)}(1,1,1)\,. \tag{3.75}$$

A related quantity that will be needed for the derivation is the generating function of the *excess degree* distribution, which is the degree distribution of nodes reached by following a randomly chosen link, without considering the one we came along. Note that if we choose a node at random, its degree will depend on p_k, however if we follow a link the probability will be higher the more links the node has, kp_k. Thus, the generating function obtained by following a directed link is

$$H_d(x,y,z) = \frac{\sum_{jlm} jx^{j-1}y^l z^m}{\sum_{jkm} jp_{jlm}} = \frac{1}{\langle k_d\rangle}G^{(1,0,0)}(x,y,z)\,, \tag{3.76}$$

similarly if we follow a directed link in the reverse direction,

$$H_r(x,y,z) = \frac{\sum_{jlm} lx^j y^{l-1} z^m}{\sum_{jkm} lp_{jlm}} = \frac{1}{\langle k_d\rangle}G^{(0,1,0)}(x,y,z)\,, \tag{3.77}$$

and lastly if we follow an undirected link,

$$H_u(x,y,z) = \frac{\sum_{jlm} mx^j y^l z^{m-1}}{\sum_{jkm} mp_{jlm}} = \frac{1}{\langle k_u\rangle}G^{(0,0,1)}(x,y,z)\,. \tag{3.78}$$

The next step is to take into account that the disease will not be transmitted through all the links. Indeed, we define the probability of a link "being infected" in the sense that node i transmits the disease to j using that link as T (regardless of it being directed or undirected). Hence, the probability of a node having exactly a of the j links emerging from

it infected is given by the binomial distribution $\binom{j}{a}T^a(1-T)^{j-a}$. Under these assumptions, the generating function is modified so that

$$\begin{aligned} G(x,y,z;T) &= \sum_{jlm} p_{jlm} \left[\sum_{a=0}^{j} \binom{j}{a} (Tx)^a (1-T)^{j-a} \sum_{b=0}^{l} \binom{l}{b} (Ty)^b (1-T)^{l-b} \right. \\ &\qquad \left. \sum_{c=0}^{m} \binom{m}{c} (Tz)^c (1-T)^{m-c} \right] \\ &= \sum_{jlm} p_{jlm} (1-T+Tx)^j (1-T+Ty)^l (1-T+Tz)^m \\ &= G(1+(x-1)T, 1+(y-1)T, 1+(z-1)T)\,. \end{aligned} \tag{3.79}$$

Analogously, the generating functions for the distribution of infected links of a node reached by following randomly chosen links are:

$$\begin{aligned} H_f(x,y,z;T) &= H_f(1+(x-1)T, 1+(y-1)T, 1+(z-1)T) \\ H_r(x,y,z;T) &= H_r(1+(x-1)T, 1+(y-1)T, 1+(z-1)T) \\ H_u(x,y,z;T) &= H_u(1+(x-1)T, 1+(y-1)T, 1+(z-1)T)\,. \end{aligned} \tag{3.80}$$

The fundamental quantity that we want to obtain is the number s of nodes contained in an outbreak that begins at a randomly selected node. Let $g(w,T)$ be the generating function for the probability that a randomly chosen node belongs to a group of infected nodes of a given size:

$$g(w;T) = \sum_s P_s(T) w^s\,. \tag{3.81}$$

To solve it, we also need to evaluate the probability that a randomly chosen link leads to a node belonging to a group of infected nodes of given size. The generating function of the distribution reads

$$h_d(w;T) = \sum_t P_t(T) w^t\,. \tag{3.82}$$

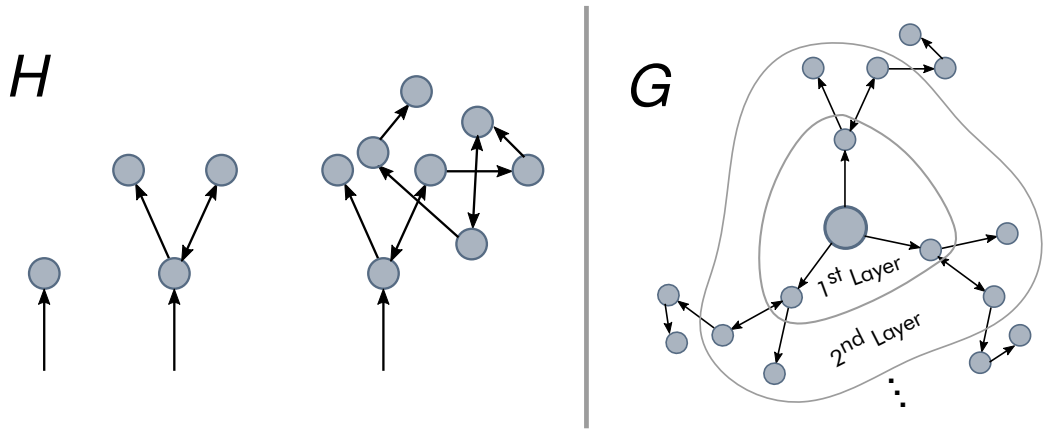

Figure 3.10: Scheme of the generating function approach. Left: The generating function of the excess degree, H, gives the distribution of links (directed and undirected) of a node reached by following a random link. Right: As the infection starts in a node, the generating function of the node's degree, G, has to be used. Hence, $G(H(x))$ gives the distribution of links in the first layer, $G(H(H(x))$ in the second layer, etc.

This expression satisfies a condition of the form

$$h_d(w;T) = wH_d(1, h_d(w;T), h_u(w;T))\,. \tag{3.83}$$

Similarly, in the case of undirected links,

$$h_u(w;T) = wH_u(1, h_d(w;T), h_u(w;T))\,. \tag{3.84}$$

With the expressions (3.80) and these two last equations, we have completely defined the distribution of t. It follows (see figure 3.10) that if the disease starts at a randomly chosen node the distribution is

$$g(w;T) = wG(1, h_d(w;T), h_u(w;T))\,, \tag{3.85}$$

yielding an average size of outbreaks of

$$\langle s\rangle = \sum_s sP_s(T) = \left.\frac{\mathrm{d}g(w;T)}{\mathrm{d}w}\right|_{w=1}\,. \tag{3.86}$$

Performing the derivatives and setting $w = 1$ we obtain

$$\begin{aligned} g' &= 1 + G^{(0,1,0)}h'_d + G^{(0,0,1)}h'_u \\ h'_d &= 1 + H_d^{(0,1,0)}h'_d + H_d^{(0,0,1)}h'_u \\ h'_u &= 1 + H_u^{(0,1,0)}h'_d + H_u^{(0,0,1)}h'_u\,, \end{aligned} \tag{3.87}$$

where we have dropped the arguments of the functions for readability. Inserting these equations in (3.86) we obtain

$$\begin{aligned} \langle s\rangle = 1 &+ \frac{G^{(0,1,0)}\left(1 - H_d^{(0,1,0)} + H_u^{(0,1,0)}\right)}{\left(1 - H_d^{(0,1,0)}\right)\left(1 - H_u^{(0,0,1)}\right) - H_u^{(0,1,0)}H_d^{(0,0,1)}} \\ &+ \frac{G^{(0,0,1)}\left(1 - H_d^{(0,0,1)} + H_u^{(0,0,1)}\right)}{\left(1 - H_d^{(0,1,0)}\right)\left(1 - H_u^{(0,0,1)}\right) - H_u^{(0,1,0)}H_d^{(0,0,1)}}\,. \end{aligned} \tag{3.88}$$

Note that this expression diverges if

$$\left(1 - H_d^{(0,1,0)}\right)\left(1 - H_u^{(0,0,1)}\right) - H_u^{(0,1,0)}H_d^{(0,0,1)} = 0\,. \tag{3.89}$$

In other words, equation (3.89) sets the condition for the epidemic threshold. The last step is to note that

$$G^{(1,0,0)}(1,1,1;T) = TG^{(1,0,0)}(1,1,1)\,, \tag{3.90}$$

and similarly for the rest of equations. Hence, equation (3.89) reads

$$\left(1 - TH_d^{(0,1,0)}\right)\left(1 - TH_u^{(0,0,1)}\right) - T^2H_u^{(0,1,0)}H_d^{(0,0,1)} = 0\,, \tag{3.91}$$

where now the arguments of the functions are $(1,1,1)$ instead of $(1,1,1;T)$.

In the particular case of undirected networks this expression further simplifies to

$$1 - TH_u^{(0,0,1)} = 0 \Rightarrow T = \frac{1}{H_u^{(0,0,1)}}\,. \tag{3.92}$$

To rewrite this expression in a more familiar format we can calculate the explicit dependency of $H_u^{(0,0,1)}$ as a function of the network topology:

$$\begin{aligned} H_u^{(0,0,1)}(1,1,1) &= \frac{1}{\langle k \rangle} \frac{\mathrm{d}}{\mathrm{d}z} G^{(0,0,1)}(1,1,1) = \frac{1}{\langle k \rangle} \frac{\mathrm{d}^2}{\mathrm{d}z^2} \sum_m p_m z^m \Big|_{z=1} \\ &= \frac{1}{\langle k \rangle} \sum_m m(m-1) p_m = \frac{1}{\langle k \rangle} \left(\langle k^2 \rangle - \langle k \rangle \right) . \end{aligned} \tag{3.93}$$

Therefore, the epidemic threshold is

$$T = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} . \tag{3.94}$$

3.3.3 Directionality reduces the epidemic threshold in directed multiplex networks

And the "cosmological principles" were, I fear, dogmas that should not have been proposed. (Karl Popper)

As we saw in chapter 2, network science constitutes a whole field of research on its own. Therefore, any advance in the understanding of networks in general might also have its applications in the study of disease spreading on networks. In particular, we can investigate the dynamics of diseases on the multilayer networks we introduced in section 2.1.3 [259]. One option can be to have the same network pattern in all layers but different dynamics on each of them, such as modeling the spreading of two interacting diseases in the same population [260] or the interplay between information and disease spreading that we discussed in the introduction [261]. On the other hand, we can have the same dynamics in all layers but diverse interaction patterns in each of them, in a similar fashion as our model of section 3.2.3.

In this work we will focus on the latter, i.e., the same dynamics in the whole system but different networks in the layers. Even more, we will consider that the networks can have directed links, something that is usually disregarded in epidemic models (note that adding direction to links implies that more data is necessary than just knowing that there is a relationship between two agents). Some relevant examples of the importance of the directionality in this context are the case of meerkats in which transmission varies between groomers and groomees [262] and even in the transmission of HIV that we have briefly discussed, as male-to-female transmission is 2.3 times greater than female-to-male transmission [120]. Similarly, when addressing the problem of diseases that can be transmitted among different species, it is important to account for the fact that they might be able to spread from one type of host to the other, but not the other way around. For instance, the bubonic plague can be endemic in rodent populations and spread to humans under certain conditions. If it evolves to the pneumonic form, it may then spread from human to human [263]. Analogously, Andes virus spreads within rodent populations, but it can be transmitted to humans and then spread via person-to-person contacts [264].

Recall that in multilayer networks there are two types of links: intralayer (those contained within layers) and interlayer (those connecting nodes set in different layers). Our objective is to understand how the epidemic threshold is influenced by the directionality of both intralayer and interlayer links. In particular, we will consider multiplex networks composed by two layers with either homogeneous or heterogeneous degree distributions in the layers (i.e., ER or SF networks). Besides, we will analyze several combinations of directionality: (i) Directed layer - Undirected interlinks - Directed layer (DUD); (ii) Directed layer - Directed interlinks - Directed layer (DDD); and Undirected layer - Directed interlinks - Undirected layer (UDU). For the sake of comparison, we will also include the

standard scenario, namely, (iv) Undirected layer - Undirected interlinks - Undirected layer (UUU). We will implement a susceptible-infected-susceptible (SIS) model on these networks and study the evolution of the epidemic threshold as a function of the directionality and the coupling strength between layers. In addition, we will derive analytically the epidemic threshold using generating functions (see 3.3.2) to obtain theoretical insights on the underlying mechanisms driving the dynamics of these systems.

First, we implement stochastic SIS dynamics on the two layer multiplex networks. Note that as there are two types of links, we can associate a different spreading probability to each of them: the interlayer spreading probability, γ, and the intralayer spreading probability, β [265]. Accordingly, a node can transmit the disease with probability β to those susceptible neighbors contained in the same layer and with probability γ to those set in the other layer. As a consequence, the epidemic threshold will depend on both parameters. Thus, henceforth we will define the epidemic threshold as β_c and explore its value as a function of γ (note that previously we defined the epidemic threshold as the ratio β/μ, but in this case we will keep fixed the value of μ for simplicity).

In the simulations, all the nodes are initially susceptible. The spreading starts when one node is set to the infectious state. Then, at each time step, each infected node spreads the disease through each of its links with probability β if the link is contained in a layer and with probability γ if the link connects nodes in different layers. Besides, each infected node recovers with probability μ at each time step. The simulation runs until a stationary state for the number of infected individuals is reached.

To determine the epidemic threshold we fix the value of γ and run the simulation over multiple values of β, repeating 10^3 times the simulation for each of those values. The minimum value of β at which, on average, the number of infected individuals in the steady state is greater than one determines the value of the epidemic threshold. This procedure is then repeated for several values of γ to obtain the dependency of β_c with the spreading across layers. Lastly, this dependency is evaluated for 100 realizations of each network considered in the study and their $\beta_c(\gamma)$ curves are averaged.

For the cases in which the interlinks are directed, we need to add another parameter to the model. If all the links were to point in the same direction, the epidemic threshold would be trivially the one of the source layer and thus the multiplex structure would play no role. For this reason, for each directed link connecting layers u and v we set the directionality to be $u \to v$ with probability p and $u \leftarrow v$ with probability $(1-p)$. Consequently, in networks with directed interlinks the epidemic threshold will be given as a function of this probability p.

The results, figure 3.11, signal that the consequences of changing the directionality of some links is completely different for SF and ER networks. In particular, in 3.11A, we can see that for networks with $\langle k \rangle = 6$ the epidemic threshold is very similar in both UUU and DUD configurations. This effect is again seen for denser networks, $\langle k \rangle = 12$, implying that it is the directionality of the interlinks, and not the one of the links contained within layers, the main driver of the epidemic in these networks. On the other hand, in figure 3.11B we can see that this behavior is not replicated in SF networks. Certainly, there is a large difference between the curves of the UUU and DUD configurations, implying that the directionality of intralinks is much more important in this type of networks. A similar pattern is observed in figures 3.11C and 3.11D, in which the interlinks are directed. Moreover, in all the cases considered the epidemic threshold is always lower for those configurations with undirected links within the layers, compared to those in which those links are directed, given the same interlink directionality.

To get further insights into the mechanisms driving this behavior we proceed to compute analytically the epidemic threshold. We introduced the generating function of the network, equation (3.72), saying that it accounts for the probability of having j incoming links, l

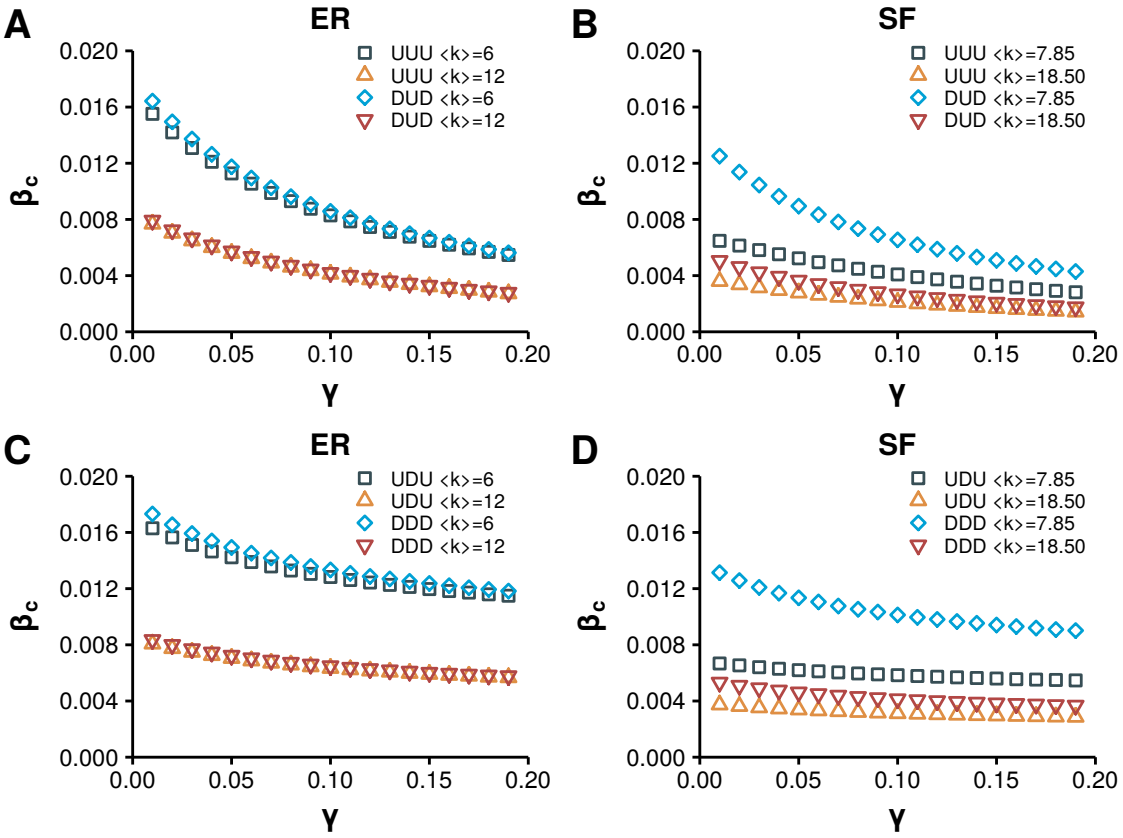

Figure 3.11: Epidemic threshold in directed multilayer networks according to SIS simulations. Several configurations of networks are considered: A) ER networks with undirected interlinks; B) SF networks with undirected interlinks; C) ER networks with directed interlinks; D) SF networks with directed interlinks. In all cases $\mu = 0.1$ the number of nodes is $N = 2 \cdot 10^4$ and for each directionality configuration there are two sets of networks with different average degree as shown in the legend. In the networks with directed interlinks $p = 0.5$.

outgoing links and m undirected links. However, in this case the directionality of links within the layers is always the same (we do not mix directed and undirected links). Hence, we can use j as an indicator for directed links when we have directed intralinks, or we can regard it as the number of undirected links otherwise. This frees m to be used for the interlinks. In other words, the generating function will now be $G(x, z)$ if the network has the shape UXU and $G(x, y, z)$ if it is DXD, with z representing the links connecting different layers.

Analogously, the definition of the generating function for the excess distribution (3.76) does not change. The first difference is encountered when we want to obtain the probability of a link being infected. In the previous case, we set said probability equal to T in all links, but now we have β for links within layers and γ for links across layers. Thus, we keep T as the probability of a link within layers being infected, and denote the probability of the other set of links being infected as T_{uv}. With these definitions equation (3.79) now reads

$$G(x, y, z; T, T_{uv}) = G(1 + (x-1)T, 1 + (y-1)T, 1 + (z-1)T_{uv}) \,. \tag{3.95}$$

Next, we introduced the generating function used to calculate the probability that a randomly chosen link belongs to the group of infected nodes. We distinguished h_d and h_u if the links were directed or undirected respectively. In this case, as the directionality is the same, what we need to define is h_1 if the link is in layer 1, h_2 if it is in layer 2, h_{12} if it

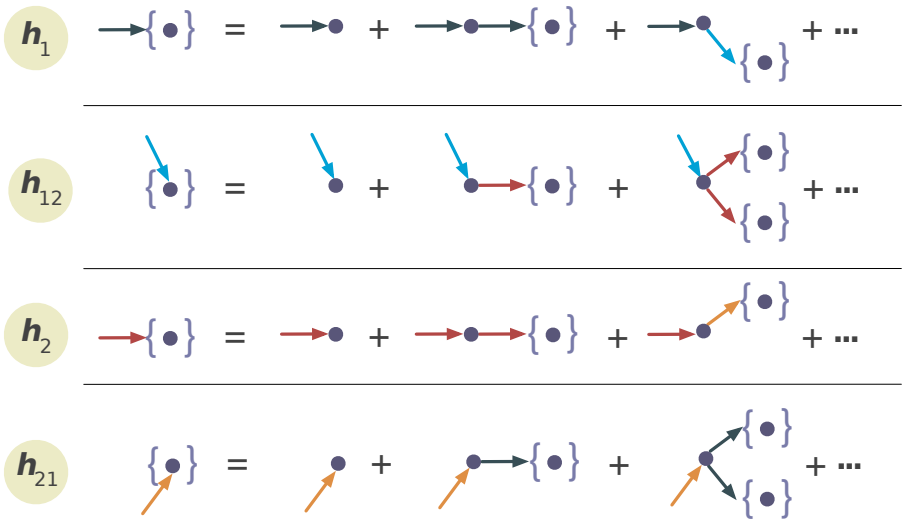

Figure 3.12: Scheme of the generating function on multilayer networks. Recursive relation of generating functions for the size distribution of outbreaks by following a link in layer 1, h_1, from 1 to 2, h_{12}, in layer 2, h_2, and from 2 to 1, h_{21}.

is a link going from layer 1 to layer 2 and h_{21} if it is going from layer 2 to layer 1. The recursive relations in this case (see figure 3.12) read

$$\begin{aligned} h_1(w;T,T_{uv}) &= wH_1(1,h_1(w;T,T_{uv}),h_{12}(w;T,T_{uv});T,T_{uv}) \\ h_2(w;T,T_{uv}) &= wH_2(1,h_2(w;T,T_{uv}),h_{21}(w;T,T_{uv});T,T_{uv}) \\ h_{12}(w;T,T_{uv}) &= wH_{12}(1,h_2(w;T,T_{uv}),h_{21}(w;T,T_{uv});T,T_{uv}) \\ h_{21}(w;T,T_{uv}) &= wH_{21}(1,h_1(w;T,T_{uv}),h_{12}(w;T,T_{uv});T,T_{uv})\,. \end{aligned} \tag{3.96}$$

Then, the generating function for the distribution of the size of an outbreak starting in a randomly chosen node in layer 1 is

$$g(w;T,T_{uv}) = wG(1,h_1(w;T,T_{uv}),h_{12}(w;T,T_{uv});T,T_{uv})\,. \tag{3.97}$$

Leading to the expression of the average size of an outbreak:

$$\begin{aligned} \langle s\rangle = \sum_s sP_s(T) &= \left.\frac{\mathrm{d}g(w;T,T_{uv})}{\mathrm{d}w}\right|_{w=1} \\ &= 1 + G^{(0,1,0)}h_1' + G^{(0,0,1)}h_{12}'\,. \end{aligned} \tag{3.98}$$

As in the previous case, this equation diverges when the denominator is equal to 0. Hence, after some algebra, the condition that establishes the epidemic threshold reads

$$\begin{aligned} 0 = &\left[\left(1-H_1^{(0,1,0)}\right)\left(1-H_{12}^{(0,0,1)}H_{21}^{(0,0,1)}\right) - H_1^{(0,0,1)}H_{12}^{(0,0,1)}H_{21}^{(0,1,0)}\right] \\ &\cdot\left[\left(1-H_2^{(0,1,0)}\right)\left(1-H_{12}^{(0,0,1)}H_{21}^{(0,0,1)}\right) - H_2^{(0,0,1)}H_{21}^{(0,0,1)}H_{12}^{(0,1,0)}\right] \\ &- H_1^{(0,0,1)}H_2^{(0,0,1)}H_{12}^{(0,1,0)}H_{21}^{(0,1,0)}\,. \end{aligned} \tag{3.99}$$

Note that this expression works for all the configurations we are considering in this work, given that we choose the proper values of H_x. For instance, for the DUD configuration we have

$$\begin{aligned} H_1^{(0,1,0)} &= H_2^{(0,1,0)} = T\langle k\rangle \\ H_1^{(0,0,1)} &= H_2^{(0,0,1)} = T\langle k\rangle \\ H_{12}^{(0,1,0)} &= H_{21}^{(0,1,0)} = T_{uv} \\ H_{12}^{(0,0,1)} &= H_{21}^{(0,0,1)} = T_{uv} \end{aligned} \tag{3.100}$$

yielding

$$T_c = \frac{1 - T_{uv}}{\langle k \rangle} \,. \tag{ER-DUD}$$

Similarly, we can obtain the epidemic threshold for the rest of the configurations:

$$T_c = \frac{1 - T_{uv}}{\langle k \rangle + 1 - T_{uv}} \,, \tag{ER-UUU}$$

$$T_c = \frac{2}{\langle k \rangle (2 + m + \sqrt{m(m+8)})} \,, \tag{ER-DDD}$$

where $m = p(1-p)T_{uv}^2$,

$$T_c = \frac{2(1 + \langle k \rangle) + m' - \sqrt{m'(4 + 8\langle k \rangle + m')}}{2((1 + \langle k \rangle)^2 - m'\langle k \rangle)} \,, \tag{ER-UDU}$$

with $m' = \langle k \rangle p(1-p)T_{uv}^2$. These results were simplified thanks to the property $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$ of Poisson distributions. For the case of SF, on the other hand, we cannot do this simplification and thus some expressions will depend on both moments of the distribution:

$$T_c = \frac{1 - T_{uv}}{\langle k \rangle} \,, \tag{DUD-SF}$$

$$T_c = \frac{\langle k \rangle (1 - T_{uv})}{\langle k^2 \rangle (1 - T_{uv}) + \langle k \rangle^2 T_{uv}} \,, \tag{UUU-SF}$$

$$T_c = \frac{2}{\langle k \rangle (2 + m + \sqrt{m(m+8)})} \,, \tag{DDD-SF}$$

where $m = p(1-p)T_{uv}^2$, and lastly

$$T_c = \frac{2\langle k^2 \rangle \langle k \rangle + \langle k \rangle^2 \left(\langle k \rangle m - \sqrt{m(4\langle k^2 \rangle + \langle k \rangle^2 (4+m))} \right)}{2(\langle k^2 \rangle^2 - \langle k \rangle^4 m)} \,. \tag{UDU-SF}$$

These expressions closely match the results obtained in the simulations, figure 3.13. Again, we can observe that the value of the epidemic threshold of the DUD configuration in SF networks tends to the value of the UUU configuration for large values of the spreading probability across layers, mimicking the behavior of ER networks. Hence, in general, we can conclude that the directionality (or lack of) of the interlinks is the main driver of the epidemic spreading process. The exception is the limit of small spreading from layer to layer as in this scenario the directionality of interlinks makes SF networks much more resilient. Altogether, the conclusion is that directionality reduces the impact of disease spreading in multilayer systems.

It is worth point out that these results are not only relevant for the situations we have described in this chapter so far. One particularly interesting and open challenge is to quantify the effects that the interplay between different social networks could have on spreading dynamics. The theoretical framework developed here is particularly suited to study this and similar challenges related to the spreading of information in social networks. On one hand, because social links are not always reciprocal [266], specially in online systems in which a user is not necessarily followed by her followings. Besides, disease-like models have been widely used to study information dissemination [259, 267]. For this reason, we have analyzed the dependence of the epidemic threshold with the inter-spreading rate in a real social network composed by two layers, figure 3.14A. The first layer of the multilayer system is made up by the directed set of interactions in a subset of users of the now defunct FriendFeed platform, whereas the second layer is defined by the directed set of interactions

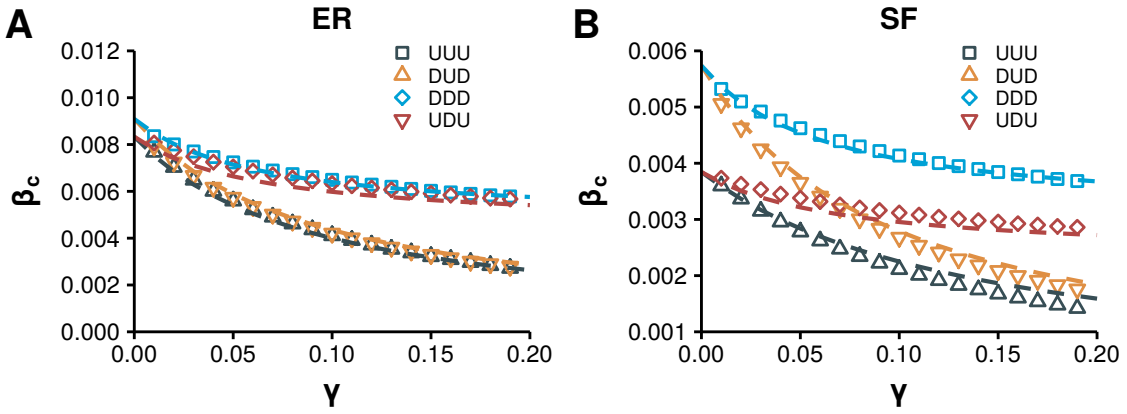

Figure 3.13: Epidemic threshold in directed multilayer networks, simulations vs. theory. A) Comparison between the results of the stochastic simulations (points) and the theoretical predictions (lines) for the ER set of networks. B) As A) but for SF networks.

of those same users in Twitter. Even though this multiplex network corresponds to a DUD configuration, we have also explored the other configurations that we have studied. Note that in contrast to the synthetic networks used so far, in this network the layers have different average degree. In particular, the FriendFeed layer has 4,768 nodes and 29,501 directed links, resulting in an average out-degree of 6.19 while the Twitter layer is composed by 4,786 nodes and 40,168 directed links, with an average out-degree of 8.42. Nevertheless, their degree distributions are both heavy tailed, resembling the power law structure of SF networks, although the maximum degree in the FriendFeed network is much larger than in the Twitter network [268].

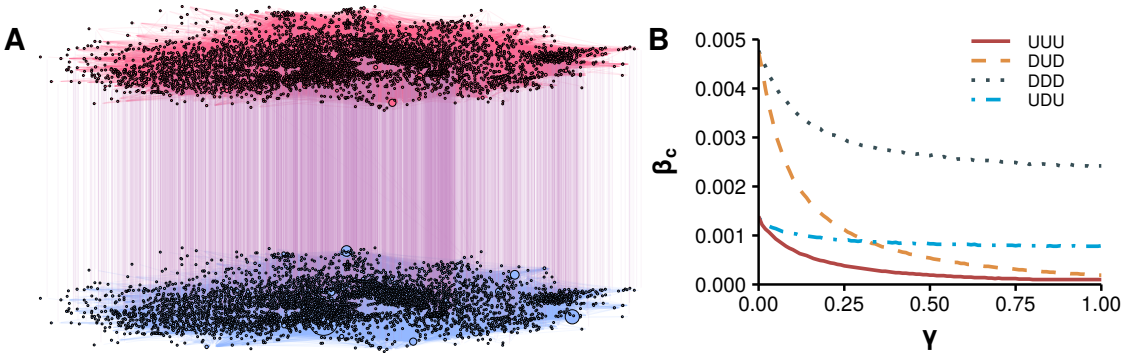

Figure 3.14: Epidemic threshold in a multilayer social system. Epidemic threshold obtained from simulations in a multiplex network composed by users of two different social platforms: FriendFeed and Twitter. The original network, panel A, has directed intralinks and undirected interlinks, corresponding to a DUD configuration. Nevertheless, to explore the effects of directionality in a real scenario, the four discussed configurations are considered in panel B. For those configurations with directed interlinks we set $p = 0.5$.

The results, figure 3.14B, confirm the findings of synthetic networks. In particular, configurations with some directionality are always more resilient against the spreading. Consequently, information travels much more easily in undirected systems than in directed systems. This is particularly worrisome given that even though Twitter can be modeled as a directed network, social networks such as Facebook and Whatsapp should be modeled using undirected configurations and, recently, these two platforms were identified as one of the main sources of misinformation spreading [269].

In summary, we have seen the importance that networks have in shaping disease dynamics. Hence, as more data becomes available, our network models should be improved in order to better account for the real mechanisms behind such a dynamics. To this end, in this section we have developed a framework that allows studying disease-like processes in multilayer networks with, possibly, directed links. This represents an important step towards the characterization of diffusion processes in interdependent systems. Our results show that directionality has a positive impact on the system's resistance to disease propagation. Furthermore, we have seen that the way in which interconnected social networks are coupled can determine their ability to spread information. Hence, the insights obtained in this work can be applied to a plethora of systems and show that more emphasis should be put in studying the role of interlinks and directionality in diffusion processes.

3.4 Age and network structures

The problems that we have studied in this chapter have shown us that, when we consider no longer that humans are particles, the dynamics of an epidemic can change dramatically. Note that the mass action approximation was just a handy tool to overcome either the scarcity of data in the past or the analytical intractability of some formulations. However, nowadays we have both enough data and computational power to introduce many more details in the dynamics.

It is now time to return to where we left at the end of chapter 2. In section 2.6.3 we introduced a mathematical framework that allowed us to create networks in which both the degree distribution and the age contact patterns of the population could be taken into account. With all the information we have gathered about disease dynamics, we can finally analyze the implications of this choice.

In the following, we will consider four different scenarios depending on the data that one may have at her disposal:

1. Homogeneous mixing with $\langle C \rangle$: suppose that the only data available is the average number of contacts and individual may have. In this case, we would be in the same situation as in the studies of gonorrhea that we presented in section 3.3. According to equation (3.53) the epidemic threshold in this situation is

$$\frac{\beta}{\mu} = \frac{1}{\langle C \rangle} \,. \tag{3.101}$$

2. Homogeneous mixing with age contact patterns: if data about the mixing patterns of the population is available, we can improve the model by creating multiple homogeneous mixing groups, one for each age bracket, as we discussed in section 3.1.1. Note that this formulation is similar to the one that Anderson and May introduced for studying interacting groups with different activity patterns, equation (3.54). Hence, the epidemic threshold, according to equation (3.57), should be

$$\frac{\beta}{\mu} = \frac{\langle C \rangle}{\langle C^2 \rangle} \,. \tag{3.102}$$

3. Network information: if we have only information about the network structure, the epidemic threshold is given by (3.71),

$$\frac{\beta}{\mu} = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \,. \tag{3.103}$$

4. Network and age information: if we are able to obtain information about both the network structure and the age mixing patterns of the population, we can build the network of interactions using the techniques introduced in section 2.6.3.

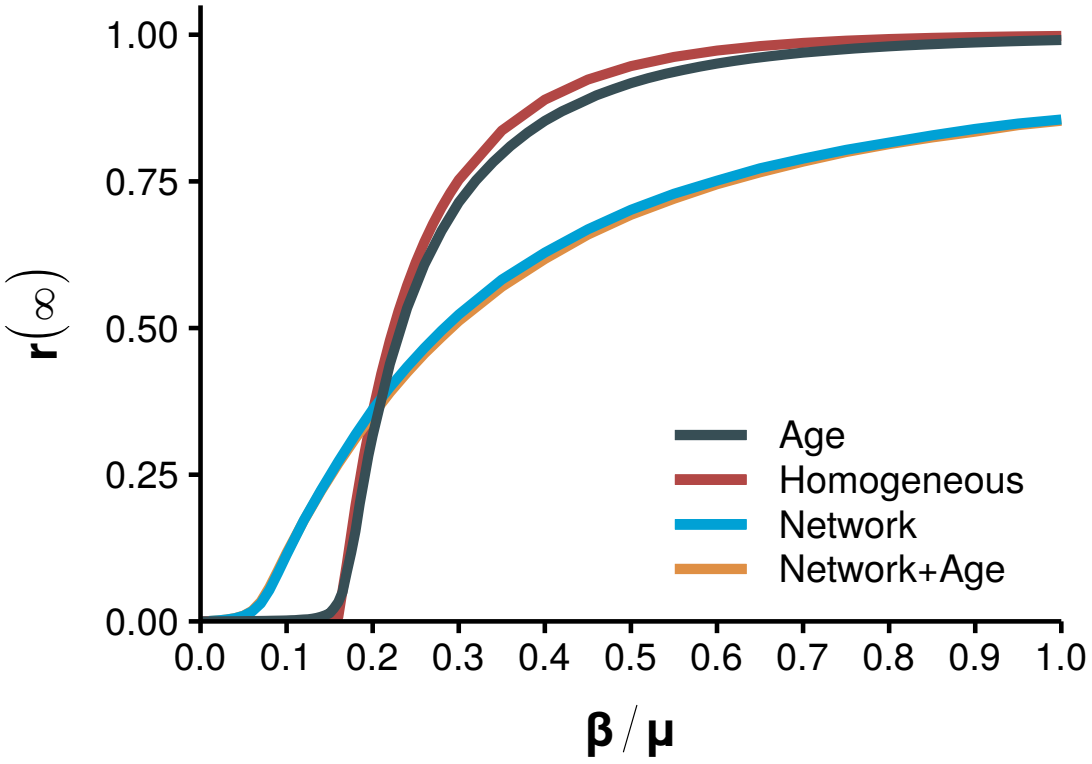

Figure 3.15: Phase diagram for different amounts of data. We compare the total number of recovered individuals as a function of the ratio β/μ for four models, each one determined by the amount of data used. In the homogeneous scenario the only information is $\langle C \rangle = 6.18$ which yields an epidemic threshold of approximately 0.16. Next, we extract the information about the age mixing patterns of Belgium in 2005 from the polymod study [143] and weight the matrix so that the average number of contacts is also 6.18. This, in turn, produces $\langle C^2 \rangle = 40.37$ yielding an epidemic threshold slightly over 0.15. To model the network structure we have assumed that the degree distribution follows a power law with $\langle k \rangle = 6.18$ and $\langle k^2 \rangle = 102.27$, resulting in a threshold of 0.06. Lastly, we have combined this network distribution with the data from Belgium to build the age contact network.

To properly compare the four situations we will set in all of them the same average number of contacts, which we denote by $\langle C \rangle$ when there is no network information and by $\langle k \rangle$ when there is. Next, we perform numerical simulations of these scenarios, using the SIR model introduced so far (with the adequate modifications depending on the amount of data available), to compare the evolution of the epidemic size as a function of the ratio β/μ, figure 3.15.

We can clearly see the effect that the heterogeneity of the network introduces in the system. For the homogeneous scenario and the homogeneous with information about the mixing patterns, the epidemic threshold is almost the same. However, when we introduce a network with a power law degree distribution, the heterogeneity in the contacts is much larger, yielding a smaller epidemic threshold. Thus, it seems that from this point of view it is more important to be able to collect data about the contact structure of the population than from the age mixing patterns.

On the other hand, note that there are many diseases which can affect differently an individual depending on her age. This effect was particularly important, for instance, in the 2009 H1N1 pandemic [170]. Furthermore, age is one of the factors used to classify people

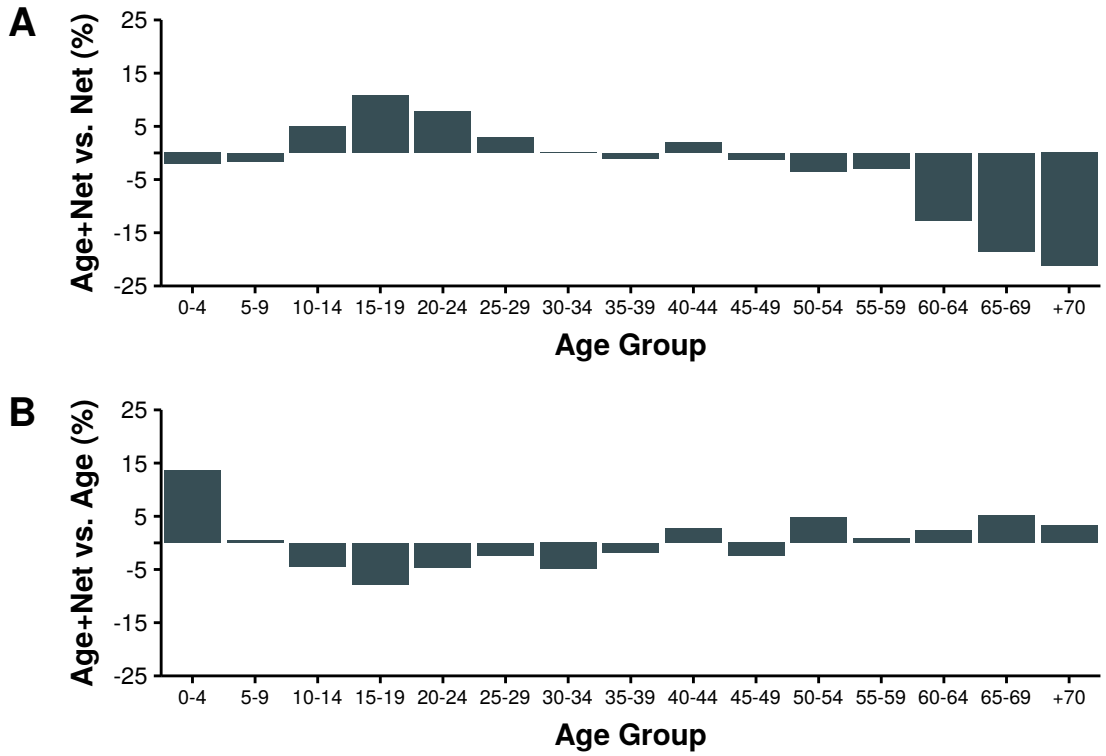

Figure 3.16: Comparison of attack rate per age group. In A we compare the number of recovered individuals in each age bracket in the model with age and network structure against the model with information of the network structure, so that positive values imply that the attack was larger in the model with more information, and vice versa. In B the comparison is done between the model with all the information and the homogeneous with age mixing patterns. Note that to compare similar situations a value of $\beta = 0.21$ has been chosen, as it is the value in which the three dynamics intersect in figure 3.15.

into risk groups, which in turn are the main targets of vaccination campaigns. Hence, even if, according to figure 3.15, knowledge of the age structure does not seem too important, once we look closer into the dynamics of the system the situation changes.

To illustrate this, in figure 3.16, we compare the attack rate of each group in the different models. In particular, in panel A we show the relative change on the amount of recovered individuals per age group between the model with all the information and the model that only considers the network structure. As we can see, the complete model has higher attack rates in teenagers, while having much lower attack rates among the elderly. Thus, even if the total attack rate in both situations is the same, if the decision to administer a vaccine is only based on the network structure we would be making a big mistake.

Conversely, adding the network information to the age mixing matrices has a smaller effect, indicating that from this point of view it is more important to collect data about the age mixing patterns of the population than from the network structure. Note that the network structure in these models is not extracted from data. Hence, the differences that we can see in panel B do not have any specific interpretation rather than that the power law distribution that we chose is responsible for those differences.

To summarize, it is clear that the more data we have, the better, as long as we use it properly and fully understand it. Yet, if for some reason we need to choose which data we need to collect, it is important to know the final application, as it is not straightforward

to say which information is more valuable. In particular, we have seen that the network structure has dramatic effects on both the epidemic threshold and the incidence of the epidemic. But, on the other hand, having information about the mixing patterns of the population might be more valuable in some situations, such as determining risk groups for vaccination.

4
Diving into the anthill

Mathematical regularities arise in the human world as soon as one shifts the attention from the individual to the collective [270]. In human societies there are transitions from disorder to order, like the spontaneous formation of a common language or the emergence of consensus about a specific issue. There are further examples of scaling, self-organization and universality. These macroscopic phenomena naturally call for a statistical approach to social behavior in which the basic constituents of the system are not particules but humans [271]. In fact, in 1842, Auguste Comte, credited as the father of sociology, already divised this possibility: "Now that the human mind has grasped celestial and terrestial physics, mechanical and chemical, organic physics, both vegetable and animal, there remains one science, to fill up the series of sciences of observation - social physics. This is what men have now most need of [...]" [19].

Despite the early observations in the XIX century about the possible existence of social physics, this area of research is still in its infancy. To understand why this is so, it might be enlightening to look at how research progressed in other areas of physics. For instance, in the XVI century Tycho Brahe recorded the position of celestial objects with unprecedented accuracy and quantity. After his death, his assistant Johannes Kepler analyzed his data and extracted the three basic laws describing planetary movement that bear his name. These, in turn, inspired Newton in the formulation, by the end of the XVII century, of the laws of motion and universal gravitation. It could be argued, then, that despite the great advances in sociology of the last century, we have just arrived to the first step of the process. That is, we are finally gathering data with unprecedented accuracy and quantity. To demonstrate it, and fully comprehend the paradigm shift that this represents, we can take the example of rumors.

During the second World War, the attention of social scientists was forcibly drawn to the subject of rumors. Not only it became aparent that wartime rumors impaired public morale and confidence, but it could also be used as a weapon of enemy propaganda. Initially, research was focused on understanding the psychology of rumors, interpreting them as something belonging to the individual. For instance, in 1944 Knapp defined a rumor as "a proposition for belief of topical reference disseminated without official verification". He even proposed that to control rumors the people had to be well informed, have confidence in their leaders and even that authorities should prevent idleness, monotony and personal disorganization as rumors - he said - do not thrive among purposeful, industrious and busy people [272]. Soon after, in 1948, Allport and Postman slightly modified Knapp's definition and said that a rumor was " a proposition for belief, *passed along from person to person, usually by word of mouth*, without secure standards of evidence being present"[1]

To study this spreading process, they performed several experiments. However, they were already aware of the limitations that in-lab experiments had in the particular context

[1]We have emphasized the crucial change in the definition. Rumor was no longer just something, it was something that *spread from person to person*. The obvious similarities of this definition with disease dynamics led Daley and Kendall to propose in 1964 that the spread of a rumor in a closed community should resemble the spread of an epidemic. Furthermore, they adapted the SIR model presented in chapter 3 to this context using ignorants, spreaders and stiflers [273].

of rumors. For instance, they had to oversimplify rumors in order to track them. Further, the intrinsic motivation of spreading a rumor is lost if you are inside a lab and a scientist is telling you to do it, being the willingness to spread substituted by the willingness to cooperate with the experimenter. They also noted that outside the laboratory the narrator tends to add color to her story, but inside the laboratory the teller feels that her reputation is at stake and does her best to transmit it in the most precise way. Moreover, they usually worked with groups of six or seven individuals.

In contrast, in 2018 Vosoughi et al. were able to investigate the diffusion of more than 126,000 true and false news stories within a population of 3 million people using Twitter data [274]. They found that false news spread farther, faster, deeper and more broadly than the truth, something that clearly could not have been studied 50 years before. Furthermore, they also investigated the role that bots in charge of systematically spreading false news could play in the spreading, but found that even though they accelerated a bit the spreading, they did not affect their total reach. Yet, few months latter, Shao et al. analyzed a much broader set of news, with almost 400,000 thousand articles and found evidence that bots do play a key role in spreading low-credibility content [275]. This contradiction is a sign that new data sources not only provide information that was not accessible before, with unprecedented accuracy and quantity, but that they also represent a subject worth of study on their own.

In this context, in section 4.1 we will study the dynamics of a Spanish online discussion board, Forocoches, with the objective of disentangling how its microscopic properties lead to our macroscopic observations. This will be based on the work

- A. Aleta, J. O'Brien, J. Gleeson, and Y. Moreno, Dynamics of discussion threads in online boards: the case of Forocoches, *In preparation*, 2019.

As McFarland et al. noted, not only these new platforms might be interesting on their own, but also give raise to new social phenomena that could not take place without digital intermediation. For example, some years ago people were simply technically unable to share photos on the scale and frequency they do today. These technologically enabled social transactions are a specific category of behaviors, some which may affect, in turn, offline social dynamics and structures. Hence, data generated on digitally-mediated platforms represent new categories of social action, not different from other phenomena of sociological interest [66].

Along these lines, we will conclude this section analyzing crowd dynamics in a digital setting. In particular, we will analyze the dynamics that emerged in an event that took place on February 2014, in which nearly a million players joined together and played a crowd controlled game, i.e., a game in which the character of the videogame was controlled simultaneously by all players. Clearly, this type of event was completely unattainable, at least with such magnitude, without the Internet. Yet, we will see that patterns that are common in the offline world had their reflection on this event. This section will be based on the work

- A. Aleta and Y. Moreno, The dynamics of collective social behavior in a crowd controlled game, *EPJ Data Sci.*, vol. 8, pp. 1–16, Jun 2019.

Besides, as we shall see, not only the two systems that we will explore are interesting on their own, but they will also allow us to discuss the dynamics that emerges when humans come together in groups. Since the late XIX century the concept of group has received a lot of attention from psychologists and sociologists as it was observed that a group is not just the addition of the individuals that compose it. The appearance of the Internet, rather than breaking the boundaries that lead naturally to groups, has allowed the formation of new and larger groups - as some sort of virtual ant colonies.

4.1 Online discussion boards

A discussion board, or Internet forum, is an online discussion site where people can hold conversations in the form of posted messages [278]. These platforms are hierarchically organized in a tree-like structure. Each forum can contain a set of sub-forums dedicated to specific topics. Then, inside each sub-forum users can begin a new conversation by opening a *thread*. In turn, other users can participate in the conversation by sending *posts* to the thread.

In the last decade, social networks have revolutionized the way we interact with each other. Yet, Internet forums precede modern online social networks by several decades. The precursors of forums date from the late 1970s, although the first proper Internet forum as we know them today was the World-Wide Web Interactive Talk created in 1994 [279]. As Rheingold noted in 1993, in these platforms virtual communities were created, exceeding the limits of the offline world [280]. He stated that the main characteristics of these communities were the fact that they belonged to the *cyberspace*, that they were based on public discussion and that personal relationships could be developed among the participants. Of these aspects, probably the most characteristic one is the fact that there are no physical boundaries in these communities, allowing people from all over the world to come together. This already raised the interest of several researchers during the late 1990s and early 2000s [281], although we find particularly interesting the thoughts of the jurist Cass R. Sunstein [282]. In 1999, he published a work on group polarization and stated that:

"Many people have expressed concern about processes of social influence on the Internet. The general problem is said to be one of fragmentation, with certain people hearing more and louder versions of their own preexisting commitments, thus reducing the benefits that come from exposure to competing views and unnoticed problems. But an understanding of group polarization heightens these concerns and raises new ones. A 'plausible hypothesis is that the Internet-like setting is most likely to create a strong tendency toward group polarization when the members of the group feel some sense of group identity'. If certain people are deliberating with many like-minded others, views will not be reinforced but instead shifted to more extreme points. This cannot be said to be bad by itself - perhaps the increased extremism is good - but it is certainly troublesome if diverse social groups are led, through predictable mechanisms, toward increasingly opposing and ever more extreme views. It is likely that processes of this general sort have threatened both peace and stability in some nations; while dire consequences are unlikely in the United States, both fragmentation and violence are predictable results. As we have seen, group polarization is intensified if people are speaking anonymously and if attention is drawn, though one or another means, to group membership. Many Internet discussion groups have precisely this feature. It is therefore plausible to speculate that the Internet may be serving, for many, as a breeding group for extremism."

These words predicted, for instance, the problem of echo chambers - people only viewing information in social networks coming from those who think like them [283] -, the role of the Internet in the arab spring [284] or the appearance of extremist groups - such as incels [285] - roughly 20 years ahead and when the Internet was, in comparison with today, still in its infancy. Admittedly, his views were probably based on the large amount of research performed on group polarization that was carried out during the XX century. Psychologists, sociologists, economists, politicians... the fact that groups are not just the addition of individuals had already attracted the interest of scientists coming from very diverse fields.

The previous examples show that these discussion platforms are worth being studied on their own. But bear in mind that these systems also provide tons of valuable data about how people interact that can, in turn, be used to test hypothesis about social behavior

that were put forward in other contexts. For instance, in 2005 Berger and Heath proposed the concept of idea *habitats* [286]. They argued that ideas have a set of environmental cues that prime people to think about them and to believe it may be relevant to pass along. Although their definition of habitat is quite broad (for instance, the current season is one of the cues building the habitat), we can clearly see that human groups in general, and online groups in particular, can be examples of habitats. Moreover, they said that to really test their ideas they would need a "perfect but unobtainable database" such as a "searchable database of all conversations". Even though discussion platforms do not posses all the information, as discussions might be influenced by external factors, it might be possible to find examples of conversations that only make sense within a particular online system. In such a case, the system would surely represent a database of all conversations. In fact, in section 4.1.1 we will see one example along these lines.

This data can also help us understand how culture disseminates and evolves. In 1985, Sperber proposed that culture could be studied under the lenses of epidemiology - as something that propagates. Yet, he doubted that mathematical models were ever going to be needed to model cultural transmission [287]. Few years later, in 1997, Axelrod presented his seminal work on cultural dissemination. With a very simple mathematical model, he demonstrated that social influence, contrary to the expectations, could naturally lead to cultural polarization rather than homogenization [288]. The accelerated rate at which online platforms evolve, in comparison to their offline counterparts, can be used to test these assumptions in the light of data. Furthermore, the boundary between the online and offline culture is getting thinner now that all cultural expressions and personal experiences are shared across the internet. Thus, this data can be used to study the evolution of the new culture that is being formed, the culture of real virtuality in Castells terms [72].

To conclude this introduction, we can give yet another example of the opportunities that having such large amounts of data represent. In 2010 Onnela and Reed-Tsochas studied the popularity of all Facebook applications that were available when the data was collected [289]. This, they claimed, removed the sampling bias that was usually present in the studies of social influence, in which only successful products were actually taken into account. By doing so, they discovered that the popularity of those applications was governed by two very different regimes: an individual one in which social influence plays no role and a collective one in which installations were clearly influenced by the behavior of others. They proposed that this type of studies could be extrapolated to other online systems. For instance, they gave the example of the (back then) online book retailer Amazon and the online DVD rental service Netflix, which allowed their users to rate their products. This would lead to an endogenously generated social influence, at a rate unprecedented in the offline world, with important economic consequences. Actually, the fact that consumers were influenced by opinions found in the Internet was something that already attracted the attention of researchers in the early 2000s in, precisely, the context of Internet forums [290].

It should be clear the wide range of possibilities that analyzing discussion boards provide. Yet, the following sections will have much more modest goals. Our objective is to understand the dynamics of the board which, in turn, should help us to study much more complex phenomena such as social influence in the future.

4.1.1 Description of Forocoches

+++ Divide By Cucumber Error. Please Reinstall Universe And Reboot +++
("Hogfather", Terry Pratchett)

Forocoches is a Spanish discussion board created in 2003 to talk about cars[2]. Back in those days it was common to have forums of very diverse topics, unlike modern social

[2]To put this date into perspective, Facebook was created in 2004, although its Spanish version was not

networks in which all the information is gathered in the same place. This fact can easily be constated by looking at the name of the subsections that compose the board, table 4.1. However, the discussions in the forum evolved throughout the years with more and more people gathering in the *General* subsection. Nowadays, this subsection contains over 80% of all the posts in the board and the discussions cover many topics that have nothing to do with cars, as it can be seen in figure 4.1.

General Area	**Technical Area & Info**
General	Mechanics
Electronics/Informatics	Car-Audio
Employment	Insurances
Travels	Traffic/Radars
Meetings	Tuning
InverForo	
ForoCoches Area	**Gaming Area**
ForoCoches	Cars games
Competition	Online games
Electric cars	**Commercial Area**
Classic cars	PIVE 8 Plan
Minivans	Buy & Sell - Professional
4x4 and SUVs	**Buy & Sell Area**
Modeling	Buy & Sell - Engine
Trucks/Vans/Buses	Buy & Sell - Audio/Tuning
Motorbikes	Buy & Sell - Electronics
	Buy & Sell - General

Table 4.1: Subsections of the board (translated from Spanish). Most of the terms are related to cars, but the distribution of posts across the board is very heterogeneous with 80% of all the messages posted in the *General* subsection.

A remarkable aspect of the forum is that since 2009 people can not register freely as in most social networks. Instead, to be able to create an account an invitation from a previous member is needed, and they were quite limited for a few years. Currently, there are some commercial campaigns that grant invitations making it slightly easier to obtain one, but in any case it is a much more closed community than common social networks in which anyone can create a new account. Note also that despite this fact the board has grown continuously since its creation, figure 4.2A.

Before going any further we should briefly describe the functioning of the board. As in other discussion boards, the information is organized in a tree-like structure. Each section is composed by a set of subsections. In each subsection, a new discussion can be started by opening a thread. Then, users can send posts to continue said discussion. From now own, we will restrict ourselves to the study of the General subsection, as it is the one with a broader set of topics and it is also the most active one as stated previously.

In the General subsection all threads that have received a new post within the last

released until 2008 [291]. Similarly, Twitter was created in 2006 and its Spanish version was released in late 2009 [292].

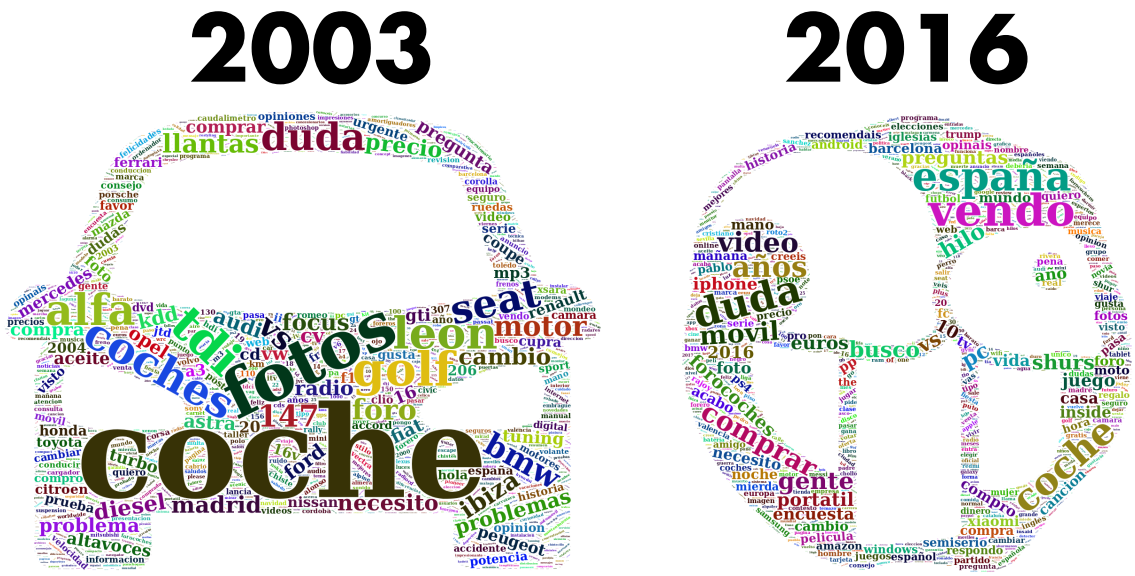

Figure 4.1: Topic evolution in Forocoches. Worclouds of the words used as thread titles. In 2003 the most common words were related to cars. Some of them refer to particular car models: alfa (alfa romeo), golf (Volkswagen golf), leon (seat leon), etc. Others represent car parts, technologies or accessories: tdi (turbocharged direct injection), cv (horsepower), llantas (rims), aceite (oil), cd, dvd, mp3... On the other hand, in 2016 the most common words refer to a broader set of topics. There are terms related to politics (pp, psoe, podemos and ciudadanos which were the main political parties in Spain in that year), technology (amazon, xiaomi, pc, iphone...), games (ps4, juego, pokemon...) to name a few.

24 hours are visible, although they are organized in a set of pages containing 40 threads each (very much like Google results). The threads appear in reverse chronological order, that is, the first thread is the one which received a new post most recently. Note that this is completely different from other social platforms in which the information is organized according to the liking of the user or related to her followers/friends. Thus, this should remove the problem of echo chambers that we mentioned previously, as people are shown all the information that is in the board regardless of whether it is suited to their likes or not. Although, admittedly, there could still be a bias due to the forum only containing a certain type of information, at least it is much easier to analyze, since it is not necessary to have precise data about the behavior of each single user.

Inside each thread, posts are organized in chronological order, being the first post the one that initiated the conversation and the last one the most recent one. Posts can contain text, images or videos. Besides, it is possible to cite a previous post in the thread (or in another thread). This does not modify the ordering of the posts, nor adds any points or likes to it. Indeed, unlike other platforms there are not any measures of popularity of posts, such as retweets or favorites. It should be noted that each thread can only contain up to 2,000 posts. Once the limit is reached the thread gets automatically closed and if users want to continue with the conversation they need to start a new thread. Nevertheless, the great majority of threads never reach that limit. This fact is shown in figure 4.2B, where the distribution of the threads sizes is plotted.

Posts can only be sent by people that registered an account in the forum. An account has a nickname associated, as well as possibly a profile picture and some more information about the user. Unlike social networks, it is not possible to automatically track the activity of other users by following them or being friends (although it is always possible to go to

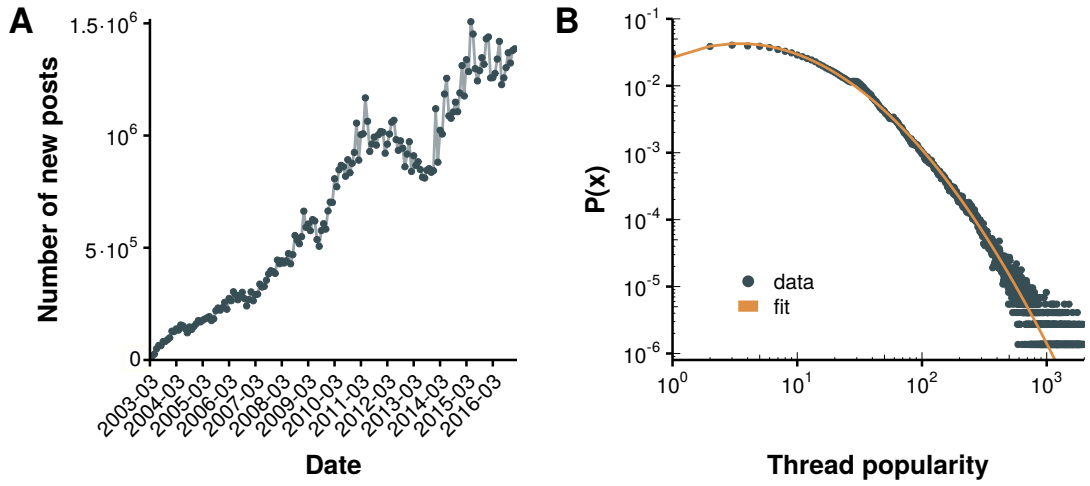

Figure 4.2: Statistics of Forocoches. A) Number of new posts per month as a function of time. The activity in the forum has increased continuously since its creation in 2003. B) Distribution of thread popularity measured as number of posts per thread. The distribution can be fitted by a lognormal distribution (which is commonly found in online social media [293, 294]) with parameters $\mu = 2.79$ and $\sigma = 1.25$.

their profile and check their latest posts). Thus, this system does not posses an explicit social network and thus the interactions between individuals should be based more on the topic than on social factors. Yet, we should emphasize that even if there is not an explicit underlying network like the ones we can find in social networks, it would be possible to construct networks that provide insights about the characteristics of the system. For instance, it would be possible to consider that users are nodes and that two users should be linked if they participate in the same thread. Further, these links could be weighted by the number of times this event occurs. Then, it would be possible to study how the information flows in the system or whether there are some underlying structures that might be hidden, such as groups of users that tend to always discuss about the same ideas together.

At this point, we should give some more details about the size of the board. Figure 4.2A shows that as of 2016 the forum received more than 1.5 million posts per month. According to the most recent statistics provided by the board, that number is now over 4 million. There are over 5 million threads, 340 million posts and 800 thousand users registered [295]. Although these numbers pale in comparison with the large social networks that are spread all over the world, note that in this case 90% of the traffic comes from Spain. This has some interesting consequences. On the one hand, it is much smaller than other social networks, making it easier to analyze but, at the same time, it is large enough to convey robust statistics. Moreover, the fact that the traffic comes mostly from Spain also facilitates the study of Spanish events without the sampling biases that arise when one uses the geo-location of users to determine where they come from in social media such as Twitter [296, 297].

Interestingly, it is possible to find remarkable similarities between Forocoches and other Internet platforms like Twitter. For instance, in figure 4.3, we show the daily activity patterns in both systems. In the case of Forocoches the data refers to the year 2015, while in Twitter it represents the tweets sent by people who had their geo-location activated and sent the tweets from within the United Kingdom during a week of October in 2015. As expected, both systems reflect the offline activity patterns of the population, with lower activities during the night. Yet, even though both systems exhibit a pattern that we could call double peaked, one at lunch time and another one at the beginning of the night,

there are clear differences that might be related to the sociological characteristics of both countries. Again, this highlights that it is possible to extract much more information from these datasets that what it might seem at first glance.

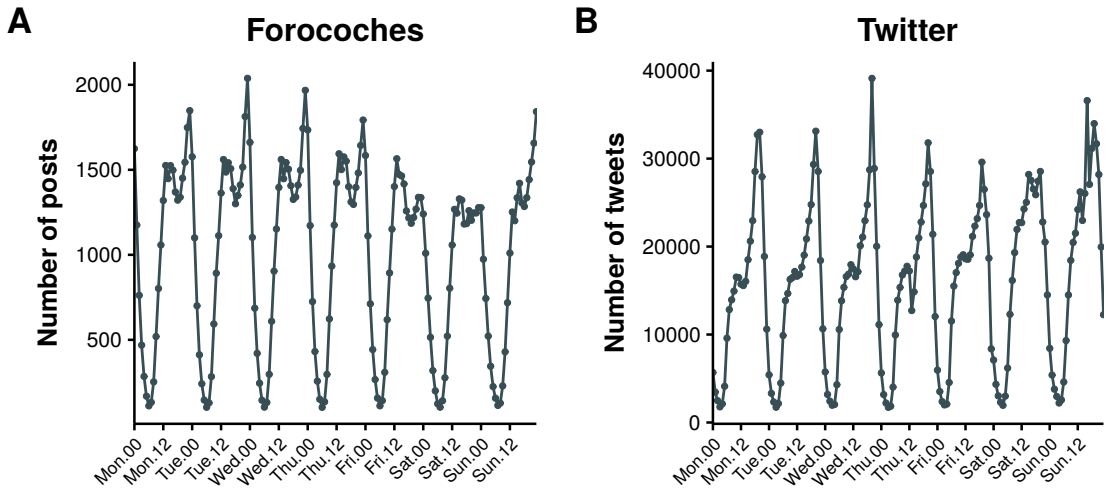

Figure 4.3: Daily activity of users in online social networks. A) Average number of posts sent as a function of time during 2015 in Forococches. B) Average number of tweets sent as a function of time in October 2015 by users who had their geo-location activated and sent them from within the United Kingdom.

Another example of the possibilities that the study of these systems bring is shown in figure 4.4. The emergence of new social contexts has enabled slang to abound on the web [298]. Although it is not possible to track the whole evolution of terms that are used all over the Internet, it is possible to find words that only have meaning in a certain context. In this particular case, we show two examples of terms that have meaning only within the board. The interesting thing about them is not their meaning but the fact that it is possible to track their whole evolution, something that obviously cannot be done in the offline world [299]. This information can then be used to study the dynamics of the cultural evolution of language [300]. In other words, we have the *database of all conversations* that Berger and Heath needed to test their hypothesis of cultural habitats.

Our goals are, however, much more modest for this part of the thesis. Our objective is to understand the mechanics behind the macroscopic behavior of the forum, which in turn should help us in the future to study more specific characteristics of the system such as the ones described so far. The starting point will be the following observation. Threads that have been inactive for over 24 hours are not removed like in other boards. Even though they are not present anymore in the list that can be directly accesed from the front page, they can still be accessed either by having their link or by finding them using Google or the own search engine of the forum. Nevertheless, it has been observed that, in Google, over 90% of the users do not go beyond the first page of results [301]. Hence, it seems reasonable to assume that users will tend to focus on the 40 threads that are on the front page. Thus, given that the more recently a thread has received a post, the most likely it is to be found in the first positions, we hypothesize that the dynamics of the forum should follow some sort of self-exciting process. In particular, we will focus on non-homogeneous Poisson processes, which have yielded satisfactory results when used to study other online social platforms, such as Twitter [302] and Reddit [303] (see [304] for a recent review on other applications of these processes).

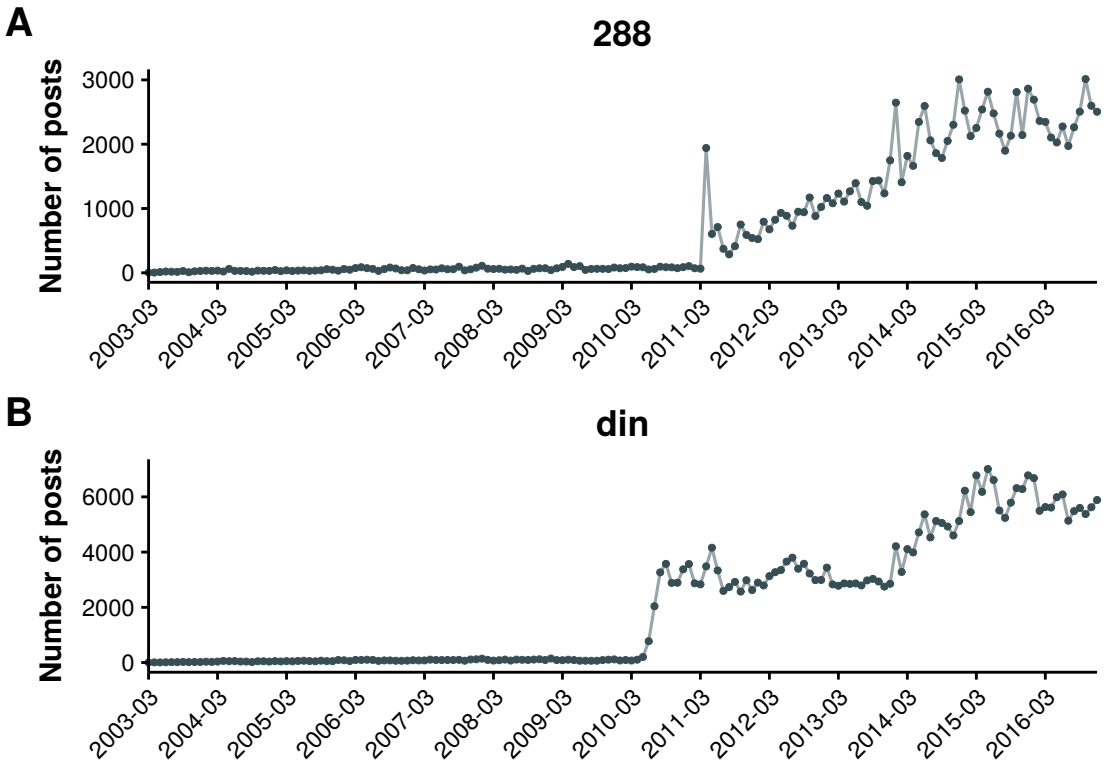

Figure 4.4: Evolution of slang in Forocoches. Usage of two memes in posts as a function of time. A) The term "288" originated when a user started a thread on the 8th of April of 2011 with the title "$48 \div 2(9+3) =$????", prompting people to give their answer. A debate on whether the division or the multiplication had to be performed first arose, with 288 being the solution if the division is performed first. After that thread the term became a meme that is used as a joke to answer questions related to numbers. B) The term "din" originated in a thread started on the 30th of May of 2010. The first person to answer the thread (after the person that created it) wrote "DIN del POST" (din of the post) probably due to a mistake (the letter d is next to f, which would be used to write *fin*, end). From that point on, the term gained popularity as a way of saying that someone posted an argument that answered the question being discussed.

4.1.2 Introduction to inhomogeneous Poisson processes

In general, point processes are used to describe the random distribution of points in a given mathematical space. In our case, this mathematical space will be the positive real line, so that events will be distributed across time. Moreover, we are not interested in the specific distribution of each event but rather on their cumulative count, as our objective is to elucidate the mechanisms leading to thread growth. In this case, point processes can be described as counting processes [305].

A *counting process* is a stochastic process defined by the number of events that have been observed (arrived) until time t, $N(t)$ with $t \geq 0$. Thus, $N(t) \in \mathbb{N}_0$, $N(0) = 0$ and it is a right-continuous step function with increments of size $+1$. Further, we denote by $\mathcal{H}_u$ with $u \geq 0$ the *history* of the arrivals up to time u. It is completely equivalent to refer to this process as a point process defined by a sequence of ordered random variables $T = \{t_1, t_2, \ldots\}$.

These processes are characterized by the conditional intensity function, which reflects

the expected rate of arrivals conditioned on $\mathcal{H}_t$:

$$\lambda(t|\mathcal{H}_t) = \lim_{h\to 0} \frac{P\{N(t,t+h] > 0|\mathcal{H}_t\}}{h}\,. \tag{4.1}$$

The most common example of these processes is the homogeneous Poisson process, in which the conditional intensity is constant. Using equation (4.1) this can be properly defined as

$$\begin{aligned} P\{N(t,t+h] = 1|\mathcal{H}_t\} &= \lambda h + o(h) \\ P\{N(t,t+h] > 1|\mathcal{H}_t\} &= o(h) \\ P\{N(t,t+h] = 0|\mathcal{H}_t\} &= 1 - \lambda h + o(h) \\ \Rightarrow \lambda(t|\mathcal{H}_t) &= \lambda\,, \end{aligned} \tag{4.2}$$

with $\lambda > 0$. An interesting consequence of this definition is that the distance between two consecutive points in time is an exponential random variable with parameter λ. This, in turn, implies that the distribution is memoryless, i.e., the waiting time (or interarrival time) until the next event does not depend on how much time has elapsed.

Conversely, a Poisson process is said to be inhomogeneous when the conditional intensity depends on time:

$$\begin{aligned} P\{N(t,t+h] = 1|\mathcal{H}_t\} &= \lambda(t) h + o(h) \\ P\{N(t,t+h] > 1|\mathcal{H}_t\} &= o(h) \\ P\{N(t,t+h] = 0|\mathcal{H}_t\} &= 1 - \lambda(t) h + o(h) \\ \Rightarrow \lambda(t|\mathcal{H}_t) &= \lambda(t)\,. \end{aligned} \tag{4.3}$$

In this section, we are interested in a specific type of inhomogeneous Poisson processes known as self-exciting or Hawkes processes, as introduced by Alan G. Hawkes in 1971 [306]. In these processes the conditional intensity not only depends on time, but also on the whole history of the event. Hence, it is given by

$$\lambda(t) = \lambda_0(t) + \int_0^t \phi(t-s)\mathrm{d}N_s\,. \tag{4.4}$$

The first term of this equation is the *background intensity* of the process while $\phi(t-s)$ is the *excitation function*. This way, the conditional intensity depends on all previous events in a way that is determined by the excitation function. Henceforth, we may refer to the function $\phi(t-s)$ as the *kernel* of the process.

Although the function $\phi(t-s)$ can take almost any form, to gain some intuition about these processes a convenient choice is the exponential function. In fact, that was the function that Hawkes used to illustrate his paper. Hence, if $\phi(t-s) = \alpha\exp(-\beta(t-s))$, we can rewrite equation (4.4) as

$$\lambda(t) = \lambda_0(t) + \int_0^\infty \alpha e^{-\beta(t-s)}\mathrm{d}N_s = \lambda_0(t) + \sum_{t_i<t} \alpha e^{-\beta(t-t_i)}\,, \tag{4.5}$$

where the constant α can be interpreted as the instantaneous excitation of the system when a new event arrives and β as the rate at which said arrival's influence decays.

In figure 4.5 we show an example of the intensity obtained using an exponential kernel. As it can be seen, every time a new event arrives, the intensity is incremented by a factor α leading to new, clustered, arrivals. Then the intensity decays at rate β until it reaches the value of the background intensity. It is worth remarking that events in Hawkes processes tend to be clustered, i.e., the interarrival time is not independent as in homogeneous processes.

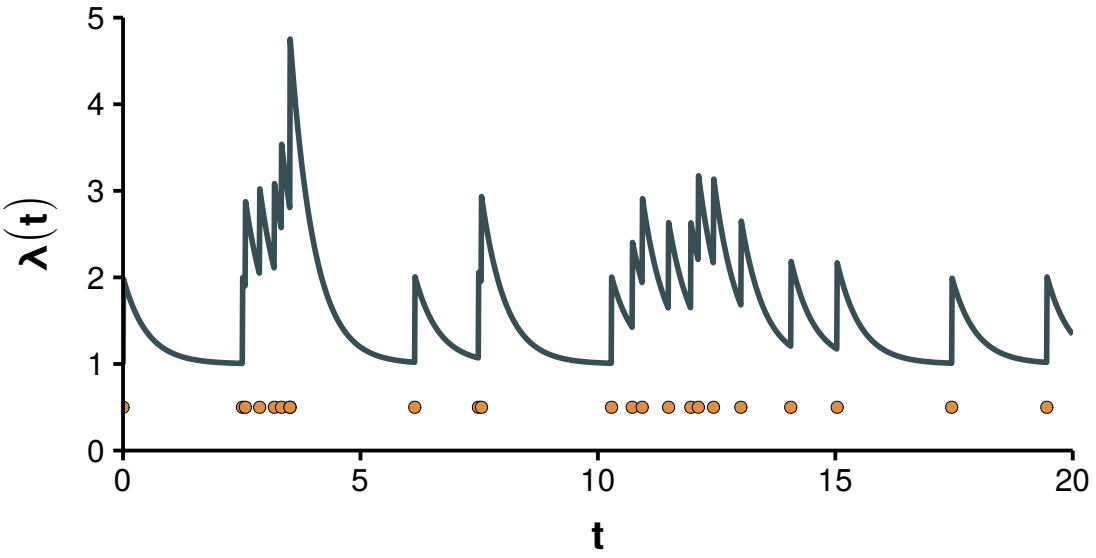

Figure 4.5: Conditional intensity function of a self-exciting process. Simulation of a Hawkes process with exponential kernel, $\lambda_0 = 1$, $\alpha = 1$ and $\beta = 2$. The curve shows the value of the conditional intensity, while dots mark the moments at which a new event arrived.

This figure can also be used to introduce a different interpretation of the process. Suppose that the stream of immigrants arriving to a country forms a homogeneous Poisson process with rate λ_0. Then, each individual can produce zero or more children independently of one another but following a simple inhomogeneous Poisson process (without excitation). The global arrival of new people to the country would then follow a Hawkes process. In the terminology of the forum, we could say that new posts arrive to the thread at a rate $\lambda_0(t)$, which might depend on time because the activity of the users changes during the day (as we saw in figure 4.3), and that each of those posts sprout themselves a sequence of new posts until the thread disappears from the front page (its intensity gets back to the background intensity).

In branching terminology, this immigration-birth representation describes the Galton-Watson process that we briefly discussed in the introduction, albeit with a modified time dimension [307]. In this context, it is possible to define the *branching ratio* of the process as

$$n = \int_0^\infty \phi(t)\mathrm{d}t = \int_0^\infty \alpha e^{-\beta s}\mathrm{d}s = \frac{\alpha}{\beta}\,, \tag{4.6}$$

which is the average number of offspring generated by each point event [308]. Both the definition of this parameter and its shape should ring some bells. Indeed, this expression is equivalent to the definition of the basic reproduction number that we saw in section 3.2. In fact, the SIR model can be studied as a Hawkes process [309]. Actually, the study of point processes has partially its origin in the demographic problems studied by mathematicians during the beginning of XX century such as Lotka, who was also the one that introduced the concept of the basic reproductive number in demography as discussed in section 3.2 [305].

A particularly successful application of Hawkes processes was introduced by Ogata in 1988 in the context of earthquakes [310]. Specifically, he used Hawkes processes to describe the occurrence of major earthquakes and the aftershocks that follow them, although he chose a different kernel. He proposed that the intensity should decay following a power law

so that

$$\lambda(t) = \lambda_0(t) + \sum_{i<t_i} \frac{\alpha}{(t - t_i + c)^{1+\beta}} \,. \tag{4.7}$$

Interestingly, he named his model for seismology the Epidemic-Type Aftershock Sequence model (ETAS).

The contribution of Ogata was not simply the introduction of the model to seismology. What really made his work outstanding was that, in a time where most researchers on point processes were mainly focused on their theoretical properties, he established a road map for how to apply point process models to real data using a formal likelihood-based inference framework [311]. The next section will be devoted to this issue.

4.1.3 Fitting Hawkes processes

If our intuition is correct, the arrival of posts to threads in Forocoches should be well described by a self-exciting process. In order to test this hypothesis we need two ingredients. First, we have to estimate the parameters that would yield the observed time sequence of a given thread. Then, we need to measure the quality of the model.

To estimate the set of parameters describing a thread we will use maximum likelihood estimation [312]. Suppose that $\{t_1, t_2, \ldots, t_n\}$ is a realization over time $[0, T]$ from a point process with conditional intensity function $\lambda(t)$. The likelihood of the process as a function of the set of parameters θ can be expressed as

$$\mathcal{L}(\theta) = \left[\prod_{i=1}^{n} \lambda(t_i|\theta)\right] \exp\left(-\int_0^T \lambda(u|\theta)\mathrm{d}u\right) , \tag{4.8}$$

and the log-likelihood is thus given by

$$l(\theta) = \ln \mathcal{L}(\theta) = \sum_{i=1}^{n} \ln[\lambda(t_i|\theta)] - \int_0^T \lambda(u|\theta)\mathrm{d}u \,. \tag{4.9}$$

For simplicity, we will assume that the background intensity is either zero or constant, so that $\lambda_0(t) \equiv \lambda_0$. Hence, in the particular case of an exponential kernel, equation (4.5), the log-likelihood reads

$$l = -\lambda_0 t + \frac{\alpha}{\beta} \sum_{i=1}^{n} \left[e^{-\beta(t_n - t_i)} - 1\right] + \sum_{i=1}^{n} \ln[\lambda_0 + \alpha A(i)] \,, \tag{4.10}$$

where $A(i) = e^{-\beta(t_i - t_{i-1})}(1 + A(i-1))$ with $A(0) = 0$. As there is no closed form solution, it is necessary to numerically obtain the maximum of this function. Fortunately, this recursive relation greatly reduces the computational complexity of the problem. For this reason exponential kernels or power law kernels with exponential cut-off are the preferred choice in the analysis of high frequency trading [313]. Nevertheless, to speed up the computation, it is convenient to also calculate the derivatives of the log-likelihood:

$$\begin{aligned}
\frac{\partial l}{\partial \lambda_0} &= -t_n + \sum_{i=1}^{n} \frac{1}{\lambda_0 + \alpha A(i)} \\
\frac{\partial l}{\partial \alpha} &= \sum_{i=1}^{n} \frac{A(i)}{\lambda_0 + \alpha A(i)} + \frac{1}{\beta} \sum_{i=1}^{n} \left[e^{-\beta(t_n - t_i)} - 1\right] \\
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{n} \frac{\alpha A'(i)}{\lambda_0 + \alpha A(i)} - \frac{\alpha}{\beta^2} \sum_{i=1}^{n} \left[e^{-\beta(t_n - t_i)} - 1\right] + \frac{\alpha}{\beta} \sum_{i=1}^{n} \left[-(t_n - t_i)e^{-\beta(t_n - t_i)}\right] ,
\end{aligned} \tag{4.11}$$

where $A'(i) = e^{-\beta(t_i - t_{i-1})}\left[-(t_i - t_{i-1})(1 + A(i-1)) + A'(i-1)\right]$ and $A'(0) = 0$.

Similarly, the log-likelihood for the power law kernel defined in equation (4.7) can be expressed as

$$l = -\lambda_0 t - \frac{\alpha}{\beta}\sum_{i=1}^{n}\left(\frac{1}{c^\beta} - \frac{1}{(t_n - t_i + c)^\beta}\right) + \sum_{i=1}^{n}\ln\left[\lambda_0 + \sum_{j=1}^{i}\frac{\alpha}{(t_i - t_j + c)^{1+\beta}}\right]. \quad (4.12)$$

In this case the computation of the kernel for long time sequences is more costly. The gradient for this expression reads

$$\begin{aligned}
\frac{\partial l}{\partial \lambda_0} &= -t_n + \sum_{i=1}^{n}\frac{1}{\lambda_0 + \alpha A(i)} \\
\frac{\partial l}{\partial \alpha} &= \sum_{i=1}^{n}\frac{A(i)}{\lambda_0 + \alpha A(i)} - \frac{1}{\beta}\sum_{i=1}^{n}\left(\frac{1}{c^\beta} - \frac{1}{(t_n - t_i + c)^\beta}\right) \\
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{n}\frac{-\alpha L A(i)}{\lambda_0 + \alpha A(i)} + \frac{\alpha}{\beta^2}\sum_{i=1}^{n}\left(\frac{1}{c^\beta} - \frac{1}{(t_n - t_i + c)^\beta}\right) + \frac{\alpha}{\beta}\left(\frac{\ln(c)}{c^\beta} - \frac{\ln(t_n - t_i + c)}{(t_n - t_i + c)^\beta}\right) \\
\frac{\partial l}{\partial c} &= -\sum_{i=1}^{n}\frac{\alpha(1+\beta)A'(i)}{\lambda_0 + \alpha A(i)} + \alpha\sum_{i=1}^{n}\left(\frac{1}{c^{\beta+1}} - \frac{1}{(t_n - t_i + c)^{\beta+1}}\right)
\end{aligned} \quad (4.13)$$

with $A(i) = \sum_{j=1}^{i}(t_i - t_j + c)^{-1-\beta}$, $LA(i) = \sum_{j=1}^{i}\ln(t_i - t_j + c)(t_i - t_j + c)^{-1-\beta}$ and $A'(i) = \sum_{j=1}^{i}(t_i - t_j + c)^{-2-\beta}$.

With these expressions we can easily estimate the set of parameters that would fit each thread in our dataset. To asses the quality of the fit, a common approach is to use tools as the Akaike information criterion (AIC) [314]. However, as already noticed by Ogata, AIC and related methods can provide information about which is the best model of all the ones being considered, but it does not say anything about whether there is a better model outside that set. Fortunately, there is a better option.

Suppose that the point process data $\{t_i\}$ are generated by the conditional intensity $\lambda(t)$. We define the *compensator* of the counting process as

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s\,, \quad (4.14)$$

which in the case of the exponential kernel is equal to

$$\Lambda(t_k) = \lambda_0 t_k - \frac{\alpha}{\beta}\sum_{i=1}^{k-1}\left[e^{-\beta(t-t_i)} - 1\right], \quad (4.15)$$

and for the power law kernel is

$$\Lambda(t_k) = \lambda_0 t_k + \frac{\alpha}{\beta}\sum_{i=1}^{k-1}\left(\frac{1}{c^\beta} - \frac{1}{(t_k - t_i + c)^\beta}\right). \quad (4.16)$$

With this definition we can now enunciate the random time change theorem [315]. If $\{t_1, t_2, \ldots, t_k\}$ is a realization over time $[0, T]$ from a point process with conditional intensity function $\lambda(t)$, then the transformed points $\{t_1^*, t_2^*, \ldots, t_k^*\}$ given by $t_i^* = \Lambda(t_i)$ form a Poisson process with unit rate.

Therefore, if the estimated conditional intensity $\lambda(t|\theta)$ is a good approximation to the true $\lambda(t)$, then the transformed points t_i^* should behave according to a Poisson process with unit rate. To test if the series forms a Poisson process we will check two of their main properties:

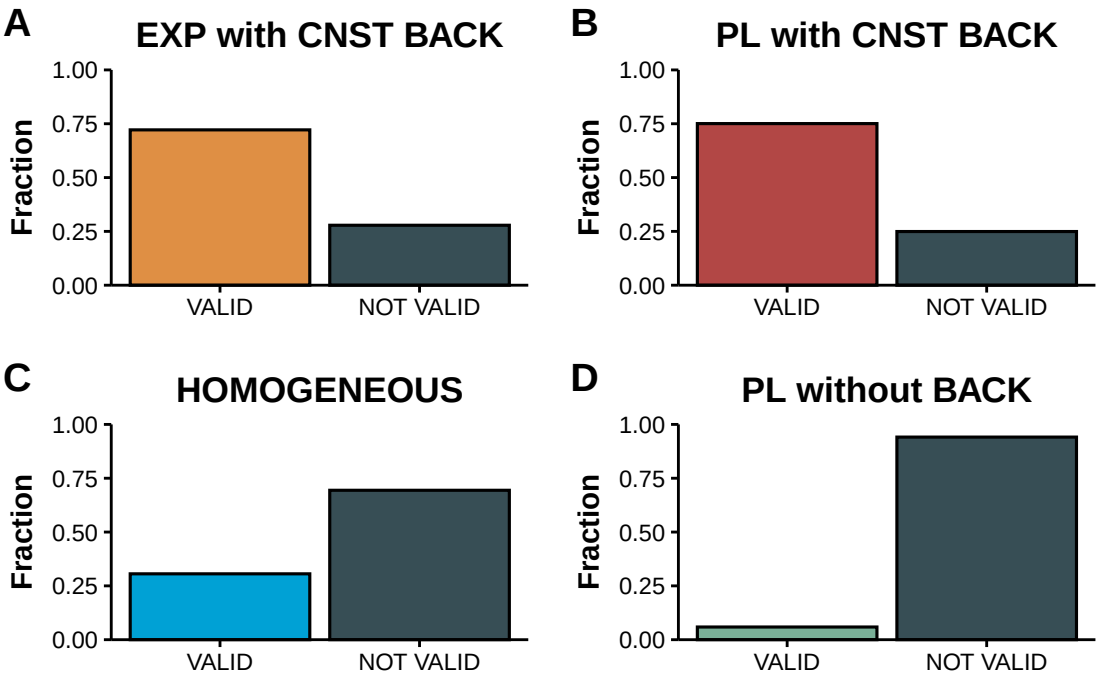

Figure 4.6: Fitting Hawkes processes to Forocoches threads. Each panels shows the fraction of threads that successfully pass all the tests described in section 4.1.3 with different kernel choices. A) Exponential kernel with constant background intensity. B) Power law kernel with constant background intensity. C) Homogeneous Poisson process. D) Power law kernel without background intensity.

- Independence: the interarrival times of the transformed points, $\tau_i^* = t_i^* - t_{i-1}^*$ should be independent. This can be tested using the Ljung-Box test. The null hypothesis of this test is that the data presents no auto-correlations (in other words, they are independent). If the p-value is higher than 0.05 then the hypothesis cannot be discarded and thus the data might be independent.

- Unit rate: if the values of $\{\tau_i^*\}$ are extracted from an exponential with unit rate then the quantity $x_k = 1 - e^{-\tau_k^*}$ is uniformly distributed in the interval $[0, 1]$. We can test this hypothesis using the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests.

Only if the estimated $\lambda(t|\theta)$ passes all these tests we will accept that it is correctly describing the evolution of a thread as a Hawkes process with the kernel under consideration. With these tools we are finally ready to asses if the dynamics of the board can be captured by these processes or not.

4.1.4 The dynamics of the board

We consider all threads that started between 01-01-2011 and 01-01-2012 with 10 or more posts, which represent nearly 230,000 different conversations. To each thread we fit: a homogeneous Poisson process; a Hawkes process with exponential kernel and constant background intensity; with power law kernel and constant background intensity; and with power law kernel without background intensity.

In figure 4.6 we show the fraction of threads that successfully pass all the tests for each kernel choice. For the moment we are not asking which model is better, only which one can fit the largest amount of threads. As we can see, both the exponential kernel and the power law kernel with constant background are able to model 75% of the threads. In

contrast, an homogeneous Poisson model can only explain 25% of the threads and a power law kernel without background intensity only a tiny fraction of roughly 5%.

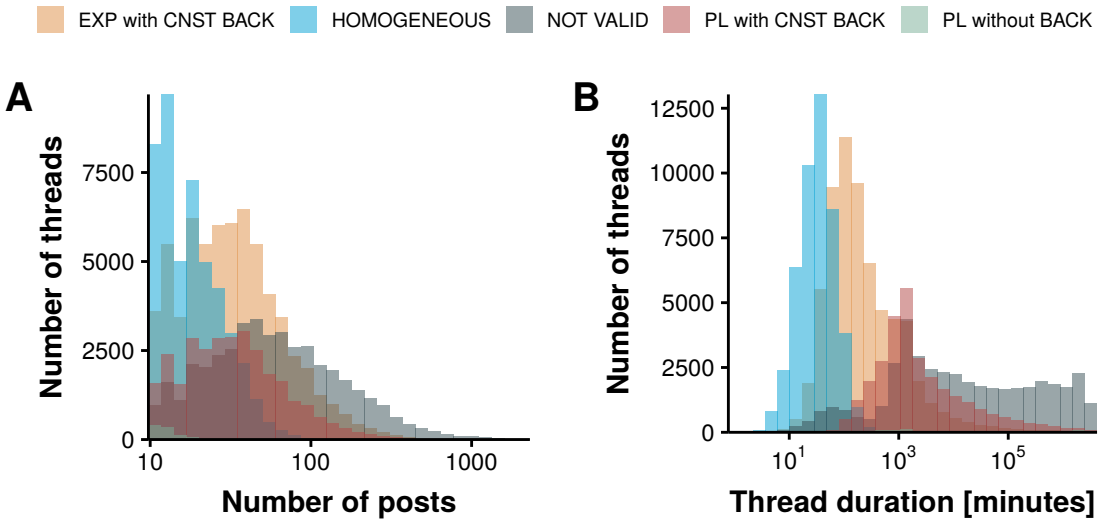

Figure 4.7: Best model as a function of external factors. For each thread that is successfully described by any of the processes that we are considering, we select the model that better fits the data using BIC. In panel A we show the distribution of those threads as a function of their popularity, i.e., their number of posts. In panel B we show the distribution as a function of the time length of the thread instead, i.e., the difference in minutes between the last and first posts.

These results partially confirm our hypothesis, as the dynamics of most threads can be well described with Hawkes processes. However, to fully understand the mechanisms underlying this system we need to address the question of what is the difference between those threads that are correctly described and those that are not. In order to do so, we first determine which is the best model for each thread. We choose to asses this using the Bayesian information criterion (BIC) as it penalizes more those models with several parameters than AIC. This is quite important given that each choice of the kernel yields a different amount of parameters.

In figure 4.7A we plot the distribution of thread size (total number of posts) distinguishing which model is better fitted to each thread. The results are quite interesting. First, the power law kernel without background intensity can only fit a tiny fraction of very short threads, signaling that the background activity of the forum is very important. Then, we find that the threads that can be fitted by a homogeneous model tend to be also rather small. In order to be able to explain larger threads, either the exponential or the power law kernels with background intensity are needed. Lastly, the longest threads cannot be described using these models.

These observations seem to point out in a similar direction as in the Facebook work that we discussed in 4.1. Indeed, in that setting, it was observed that there was a transition between a regime in which popularity was completely independent from the collective action of users and a regime in which social influence was important. In a similar fashion, we find that small threads can be studied as homogeneous Poisson processes, i.e., the arrival of new posts is independent of the ones that are already there. Conversely, once social influence comes into play, threads can reach a larger amount of popularity.

The only thing left is to disentangle why the most popular threads cannot be captured by these models. To do so, in figure 4.7B we show the distribution of thread duration,

measured as the time elapsed between the very first post and the last one, as a function of which model better fits the thread. In this case we can see that those threads better fitted by a homogeneous Poisson model are those that last only for a few minutes. Once their length is over a few hours, the exponential kernel is needed. For even longer threads, a slower decay rate is needed, hence the power law fits better. Lastly, threads that are exceptionally long cannot be fitted by any of these models. This is, however, not surprising.

Recall that in figure 4.3 we saw that the daily patterns of activity highly depend on the time of the day. Hence, it is to be expected that when a thread last for over a few hours, the effects that this activity can have in the background intensity start to be noticeable. Yet, we have considered that the background intensity is constant, something that clearly goes against this observation. Hence, to be able to explain the behavior of longer threads, a background intensity that is somehow proportional to the activity of the forum would be needed.

In conclusion, we have seen that data from discussion boards conveys a large array of opportunities for research. We have focused on disentangling the underlying dynamics of the system, for which we have proposed that a self-exciting process would be adequate. The results presented in this section signal that this hypothesis is correct, showing that there are two regimes in the forum: one in which activity is essentially random and one in which social influence plays a key role. However, in order to be able to completely characterize all types of threads, more complex models, such as background intensities that depend on the hour of the day, would be needed.

4.2 The dynamics of a crowd controlled game

> *The intelligence of that creature known as a crowd is the square root of the number of people in it.* ("Jingo", Terry Pratchett)

Collective phenomena have been the subject of intense research in psychology and sociology since the XIX century. There are several ways in which humans gather to perform collective actions, although observations suggest that most of them require some sort of diminution of self-identity [316]. One of the first attempts to address this subject was Le Bon's theory on the psychology of crowds in which he argued that when people are part of a crowd they lose their individual consciousness and become more primitive and emotional thanks to the anonymity provided by the group [317]. In the following decades, theories of crowd behavior such as the convergence theory, the emergent norm theory or the social identity theory emerged. These theories shifted away from Le Bon's original ideas, introducing rationality, collective norms and social identities as building blocks of the crowd [318, 319].

The classical view of crowds as an irrational horde led researchers to focus on the study of crowds as something inherently violent, and thus, to seek for a better understanding and prediction of violence eruption, or at least, to develop some strategies to handle them [320]. However, the information era has created a new kind of crowd, as it is no longer necessary to be in the same place to communicate and take part of collective actions. Indeed, open source and "wiki" initiatives, as well as crowdsourcing and crowdworking, are some examples of how crowds can collaborate online in order to achieve a particular objective [321, 322]. Although this offers a plethora of opportunities, caution has to be taken because, as research on the psychology of crowds has shown, the group is not just the simple addition of individuals [323]. For example, it has been observed that the group performance can be less efficient than the sum of the individual performances if they had acted separately [324]. What are the conditions for this to happen and whether the group is more than the individuals composing it are two current challenges of utmost importance if, for instance, one wants to use online crowds as a working force.

To be able to unlock the potential of collective intelligence, a deeper understanding of the functioning of these systems is needed [325]. Examples of scenarios that can benefit from further insights into crowd behavior include new ways to reach group decisions, such as voting, consensus making or opinion averaging, as well as finding the best strategies to motivate the crowd to perform some task [326]. Regarding the latter, as arbitrary tasks usually are not intrinsically enjoyable, to be able to systematically execute crowdsourcing jobs, some sort of financial compensation is used [327]. This, however, implies dealing with new challenges, since many experiments have demonstrated that financial incentives might undermine the intrinsic motivation of workers or encourage them to only seek for the results that are being measured, either by focusing only on them or by free-riding [328, 329, 330]. A relevant case is given by platforms such as Amazon's Mechanical Turk, that allow organizations to pay workers that perform micro-tasks for them, and that have already given rise to interesting questions about the future of crowd work [331]. In particular, its validity to be used for crowdsourcing behavioral research has been recently called into question [332].

Notwithstanding the previous observations, it is possible to find tasks that are intrinsically enjoyable by the crowd due to their motivational nature, which is ultimately independent of the reward [330]. This is one of the basis of online citizen science. In these projects, volunteers contribute to analyze and interpret large datasets which are later used to solve scientific problems [333]. To increase the motivation of the volunteers, some of these projects are shaped as computer games [334]. Examples range from the study of protein folding [335] to annotating people within social networks [336] or identifying the presence of cropland [337].

It is thus clear that to harness the full potential of crowds in the new era, we need a deeper understanding of the mechanisms that drive and govern the dynamics of these complex systems. To this aim, here we study an event that took place in February 2014 known as Twitch Plays Pokémon (TPP). During this event, players were allowed to control simultaneously the same character of a Pokémon game without any kind of central authority. This constituted an unprecedented event because in crowd games, each user usually has its own avatar and it is the common action of all of them what produces a given result [338]. Due to its novelty, in the following years it sprouted similar crowd controlled events such as The Button in 2015 [339] or Reddit r/place in 2017 [340, 341]. Similarly to those which came after it, TPP was a completely crowd controlled process in which thousands of users played simultaneously for 17 days, with more than a million different players [342]. TPP is specially interesting because it represents an out of the lab social experiment that became extremely successful based only on its intrinsic enjoyment and, given that it was run without any scientific purpose in mind, it represents a natural, unbiased (i.e., not artificially driven) opportunity to study the evolution and organization of crowds. Furthermore, the whole event was recorded in video, the messages sent in the chat window were collected and both are available online[3] [343]. Hence, in contrast to the offline crowd events that were studied during the last century, in this case we possess a huge amount of information of both the outcome of the event but, even more important, the evolution of the crowd during the process.

[3]The chat logs can have either seconds (YYYY-MM-DD HH:MM:SS) or minute (YYYY-MM-DD HH:MM) resolution. The game started on February 12, 2014 at 23:16:01 UTC, but the first log recorded corresponds to February 14, 2014 at 08:16:19 GMT+1. Besides, the log data between February 21, 2014 at 04:25:54 GMT+1 and 07:59:22 GMT+1 is missing. We extracted the position of the tug of war meter that will be described in section 4.2.3 as well as the game mode active at each time from the videos using optical character recognition techniques.

4.2.1 Description of the event

On February 12, 2014, an anonymous developer started to broadcast a game of Pokémon Red on the streaming platform Twitch. Pokémon Red was the first installment of the Pokémon series, which is the most successful role playing game (RPG) franchise of all time [344]. The purpose of the game was to capture and train creatures known as Pokémons in order to win increasingly difficult battles based on classical turn-based combats. However, as Pokémon Go showed in the summer of 2016, the power of the Pokémon franchise goes beyond the classical RPG games and is still able to attract millions of players [345].

On the other hand, Twitch is an online service for watching and streaming digital video broadcast. Its content is mainly related to video games: from e-sports competitions to professional players games or simply popular individuals who tend to gather large audiences to watch them play, commonly known as streamers. Due to the live nature of the streaming and the presence of a chat window where viewers can interact among each other and with the streamer, in these type of platforms the relationship between the media creator and the consumer is much more direct than in traditional media [346]. Back in February 2014, Twitch was the 4th largest source of peak internet traffic in the US [347] and nowadays, with over 100 million unique users, it has become the home of the largest gaming community in history [348].

The element that distinguished this stream from the rest was that the streamer did not play the game. Instead, he set up a bot in the chat window that accepted some predefined commands and forwarded them to the input system of the video game. Thus, anyone could join the stream and control the character by just writing one of those actions in the chat. Although all actions were sent to the video game sequentially, it could only perform one at a time. As a consequence, all commands that arrived while the character was performing a given action (which takes less than a second) did not have any effect. Thus, it was a completely crowd controlled game without any central authority or coordination system in place. This was not a multiplayer game, this was something different, something new [349].

Due to its novelty, during the first day the game was mainly unknown with only a few tens of viewers/players and as a consequence little is known about the game events of that day [350]. However, on the second day it started to gain viewers and quickly went viral, see figure 4.8. Indeed, it ramped up from 25,000 new players on day 1 (note that the time was recorded starting from day 0 and thus day 1 in game time actually refers to the second day on real time) to almost 75,000 on day 2 and an already stable base of nearly 10,000 continuous players. Even though there was a clear decay on the number of new users after day 5, the event was able to retain a large user base for over two weeks. This huge number of users imposed a challenge on the technical capabilities of the system, which translated in a delay of between 20 and 30 seconds between the stream and the chat window. That is, users had to send their commands based on where the player was up to 30 seconds ago.

Although simple in comparison to modern video games, Pokémon Red is a complex game which can not be progressed effectively at random. In fact, a single player needs, on average, 26 hours to finish the game [351]. Nevertheless, only 7 commands are needed to complete the game. There are 4 movement commands (*up*, *right*, *down* and *left*), 2 actions commands (*a* and *b*, *accept* and *back/cancel*) and 1 system button (*start* which opens the game's menu). As a consequence the gameplay is simple. The character is moved around the map using the four movement commands. If you encounter a wild Pokémon you will have to fight it with the possibility of capturing it. Then, you will have to face the Pokémons of trainers controlled by the machine in order to obtain the 8 medals needed to finish the game. The combats are all turn-based so that time is not an important factor. In each turn of a combat the player has to decide which action to take for which the movement buttons along with *a* and *b* are used. Once the 8 medals have been collected there is a final encounter after which the game is finished. This gameplay, however, was

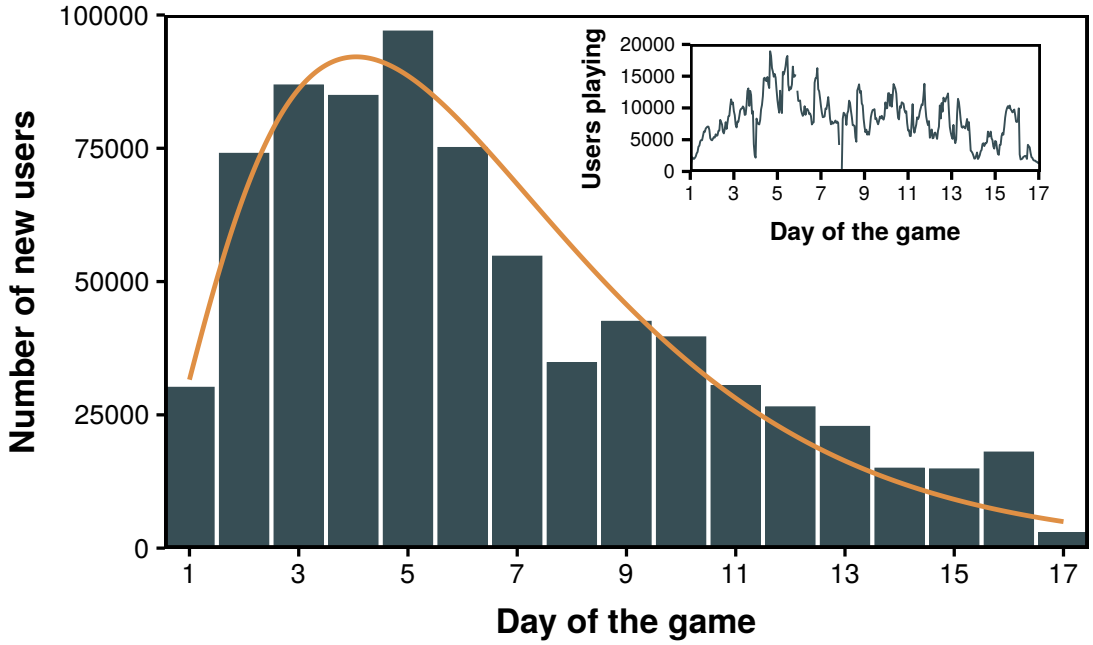

Figure 4.8: Popularity of the stream. Number of new users that arrived each day. The histogram is fitted to a gamma distribution with parameters $\alpha = 2.66$ and $\beta = 0.41$. Note that this reflects those users who inputted at least one command, not the number of viewers. In the inset we show the total number of users who sent at least 1 message each hour, regardless on whether they were new players or not.

much more complex during TPP due to the huge number of players sending commands at the same time and the lag present in the system.

A remarkable aspect of the event is that actions that would usually go unnoticed, such as selecting an object or nicknaming a Pokémon, yielded unexpected outcomes due to the messy nature of the gameplay. The community embraced these outcomes and created a whole narrative around them in the form of jokes, fan art and even a religion-like movement based on the judeo-christian tradition [352] both in the chat window and in related media such as Reddit. Although these characteristics of the game are outside of the scope of this thesis, it is another example of the new possibilities that digital systems bring in relation to the study of naming conventions and narrative consensus [353]. As we saw in section 4.1.1, language can evolve in digital platforms, with users developing new words that do not have any meaning outside the habitat where they were created. Not only it is a sign of the sociological richness of these systems, but also they might provide clues about the origin and evolution of slang in the offline world.

Returning to the discussion about the gameplay, even if it was at a slower peace, progress was made. Probably the first thing that comes to ones mind when thinking on how progressing was possible is the famous experiment by Francis Galton in which he asked a crowd to guess the weight of an ox. He found that the average of all estimates of the crowd was just 0.8% higher than the real weight [354]. Indeed, if lots of users were playing, the extreme answers should cancel each other and the character would tend to move towards the most common command sent by the crowd. Note, however, that as they were not voting, actions deviating from the mean could also be performed by pure chance. In general, this did not have great effects but as we will see in section 4.2.2 there were certain parts of the game where this was extremely relevant.

It is worth stressing that, to form a classical wise crowd, some important elements are needed, such as independence [355]. That is, the answer of each individual should not be

influenced by other people in the crowd. In our case, this was not true, as the character was continuously moving. Indeed, the big difference of this crowd event to others is that opinions had effect in real time, and hence, people could see the tendency of the crowd and change its behavior accordingly. Theoretical [356] and empirical studies [357] have shown that a minority of informed individuals can lead a naïve group of animals or humans to a given target in the absence of direct communication. Even in the case of conflict in the group, the time taken to reach the target is not increased significantly [357] which would explain why it only took the crowd 10 times more to finish the game than the average person. Although this amount may seem high, as we shall see later, the crowd got stuck in some parts of the game for over a day, increasing the time to finish. However, if those parts were excluded, the game progress can be considered to be remarkably fast, despite the messy nature of the gameplay.

As a matter of fact, the movement of the character on the map can be probably better described as a swarm rather than as a crowd. Classical collective intelligence, such as the opinions of a crowd obtained via polls or surveys, has the particularity stated previously of independence but, in addition, asynchrony. It has been shown that when there is no independence, that is, when users can influence each other, as long as the process is asynchronous the groups decisions will be distorted by social biasing effects [358]. Conversely, when the process is synchronous, mimicking natural swarms, these problems can be corrected [359]. Indeed, by allowing users to participate in decision making processes in real time with feedback about what the rest is doing, in some sort of human swarm, it is possible to explore more efficiently the decision space and reach more accurate predictions than with simple majority voting [360]. Admittedly, the interaction in the online world is so different that maybe the term crowd cannot be straightforwardly applied to online gatherings. In fact, it has recently been suggested that online crowds might be better described as swarms as something in-between crowds and networks [361].

Even though the characteristics described so far already make this event very interesting from the research point of view, on the sixth day the rules were slightly changed, which made the dynamics even richer. After the swarm had been stuck in a movement based puzzle for almost 24 hours, the developer took down the stream to change the code. Fifteen minutes later the stream was back online but this time commands were not executed right away. Instead, they were added up and every 10 seconds the most voted command was executed. In addition, it was possible to use compound commands made of up to 9 simple commands such as $a2$ or $aleftright$ which would correspond to executing a twice or a, $left$ and $right$ respectively. Thus, the swarm became a crowd with a majority rule to decide which action to take. As it waited 10 seconds between each command, progress was slow and, twenty minutes after, that time was reduced to 5 seconds. However, the crowd did not like this system and started to protest by sending $start9$ which would open and close the menu repeatedly impeding any movement. This riot, as it was called, lasted for 8 minutes (figure 4.9), moment when the developer removed the voting system. However, two hours later the system was modified again. Two new commands were added: *democracy* and *anarchy*, which controlled some sort of tug of war voting system over which rules to use. If the fraction of people voting for democracy went over a given threshold, the game would start to tally up votes about which action to take next. If not, the game would be played using the old rules. This system split the community into "democrats" and "anarchists" who would fight for taking control of the game. Therefore, the system would change between a crowd-like movement and a swarm-like movement purely based on its own group interactions. We will analyze this situation in section 4.2.3.

4.2.2 The ledge

On the third day of the game, the character arrived to the area depicted in figure 4.10 (note that the democracy/anarchy system we just described had not been introduced yet).

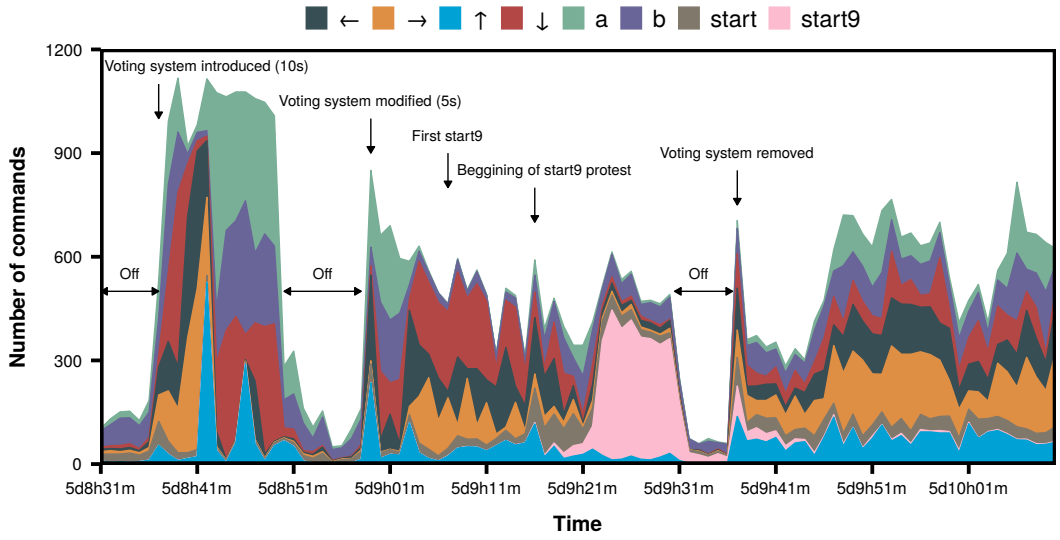

Figure 4.9: Introduction of the voting system. Command distribution after the first introduction of the voting system. Once the system was back online votes would tally up over a period of 10 seconds. After 15 minutes the system was brought down to reduce this time to 5 seconds. This, however, did not please the crowd and it started to protest. The first *start*9 was sent at 5d9h8m but went almost unnoticed. Few minutes after, it was sent again but this time it got the attention of the crowd. In barely 3 minutes it went from 4 *start*9 per minute to over 300, which stalled the game for over 8 minutes. The developer brought down the system again and removed the voting system, introducing the anarchy/democracy system a few hours later.

Each node of the graph represents a tile of the game. The character starts on the light blue node on the left part of the network and has to exit through the right part, an event that we will define as getting to one of the light blue nodes on the right. The path is simple for an average player but it represented a challenge for the crowd due to the presence of the yellow L-nodes. These nodes represent ledges which can only be traversed going downwards, effectively working as a filter that allows flux only downwards. Thus, one good step will not cancel a bad step, as the character would be trapped down the ledge and will have to find a different path to go up again. For this reason, this particular region is highly vulnerable to actions deviating from the norm, either caused by mistake or performed intentionally by griefers, i.e., individuals whose only purpose is to annoy other players and who do so by using the mechanisms provided by the game itself [362, 363] (note that in social contexts these individuals are usually called trolls [364]). Indeed, there are paths (see red nodes in figure 4.10) where only the command *right* is needed and which are next to a ledge so that the command *down*, which is not needed at all, would force the crowd to go back and start the path again. Additionally, the existence of the lag described in section 4.2.1 made this task even more difficult.

In figure 4.11A we show the time evolution of the amount of messages containing each command (the values have been normalized to the total number of commands sent each minute) since the beginning of this part until they finally exited. First, we notice that it took the crowd over 15 hours to finish an area that can be completed by an optimal walk in less than 2 minutes. Then, we can clearly see a pattern from 2d18h30m to the first time they were able to reach the nodes located right after the red ones, approximately 3d01h10m: when the number of *rights* is high the number of *lefts* is low. This is a signature of the character trying to go through the red nodes by going right, falling down the ledge,

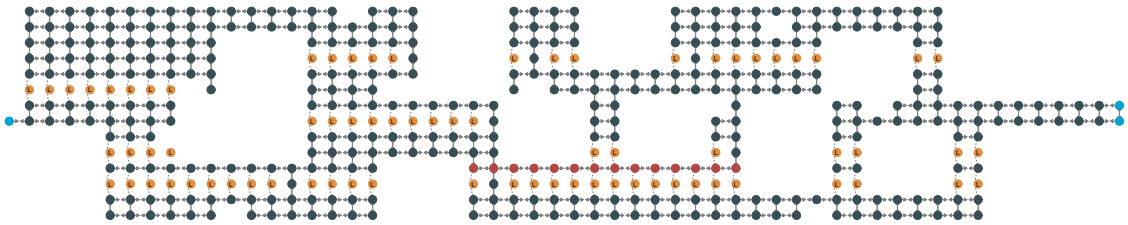

Figure 4.10: Network representation of the ledge area. It is possible to go from each node to the ones surrounding it using the commands *up*, *right*, *down* and *left*. The only exception are the yellow nodes labeled L which correspond to ledges. If the character tries to step on one of those nodes it will be automatically sent to the node right below it, characteristic that is represented by the curved links connecting nodes above and below ledges. Light blue nodes mark the entrance and exit of the area and red nodes highlight the most difficult part of the path. Note that as the original map was composed by discrete squared tiles this network representation is not an approximation but the exact shape of the area.

and going left to start over. Once they finally reached the nodes after the red path (first arrival) they had to fight a trainer controlled by the game, combat which they lost and as a consequence the character was transported outside of the area and they had to enter and start again from the beginning. Again, we can see a similar left-right pattern until they got over that red path for the second time, which in this case was definitive.

The ledge is a great case study of the behavior of the crowd because the mechanics needed to complete it is very simple (just moving from one point to another), which facilitates the analysis. But, at the same time, it took the players much longer to finish this area than what is expected for a single player. To address all these features, we propose a model aimed at mimicking the behavior of the crowd. Specifically, we consider a n-th order Markov Chain so that the probability of going from state x_m to x_{m+1} depends only on the state x_{m-n}, thus accounting for the effect of the lag of the dynamics. Furthermore, the probabilities of going from one state to another will be set according to the behavior of the players in the crowd.

To define these probabilities, we first classify the players in groups according to the total number of commands they sent in this period: G1, users with 1 or 2 commands (46% of the users); G2, 3 or 4 commands (18%); G3, between 5 and 7 commands (13%); G4, between 8 and 14 commands (12%); G5, between 15 and 25 commands (6%); and G6, more than 25 commands (5%). These groups were defined so that the total number of messages sent by the first three is close to 50,000 and 100,000 for the other three (if we had selected the same value for all of them, either we would have lost resolution in the small ones or we would have obtained too many groups for the most active players). Interestingly, the time series of the inputs of each of these groups are very similar. Actually, if we remove the labels of the 42 time series and cluster them using the euclidean distance, we obtain 7 clusters, one for each command. Even more, the time series of each of the commands are clustered together, figure 4.11B. In other words, the behavior of users with medium and large activities are not only similar to each other, but they are also equivalent to the ones coming from the aggregation of the users who only sent 1 or 2 commands.

In this context we could argue that users with few messages tend to act intuitively as they soon lose interest. According to the social heuristics hypothesis [365], fast decisions tend to increase cooperation, which in this case would mean trying to get out of the area as fast as possible. Similarly, experiments have shown that people with prosocial predispositions tend to act that way when they have to make decisions quickly [366]. Thus, users that send few commands might tend to send the ones that get the character closer

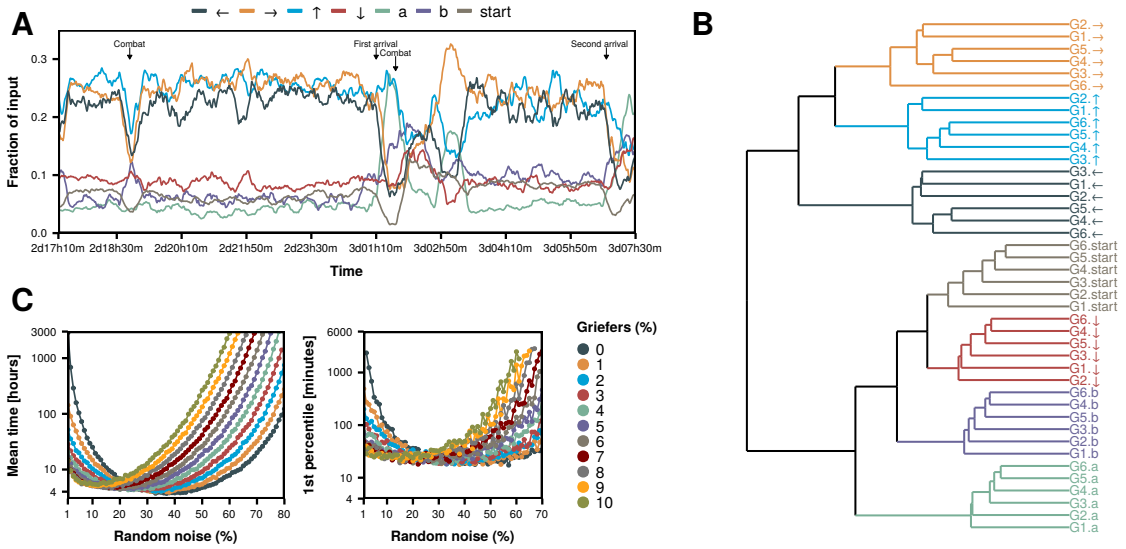

Figure 4.11: Study of the ledge event. A) Time evolution of the fraction of commands sent each minute. Note that a single player should be able to finish this area in a few minutes, but the crowd needed 15 hours. The time series has been smoothed using moving averages. B) Hierarchical clustering of the time series of each group of users. C) Left: Mean time needed to exit the area according to our model as a function of the fraction of griefers and noise in the system. Right: 1st percentile of the time needed to exit the area, note that the y axis is given in minutes instead of hours.

to the exit, which would explain why without being aware of it, they behave as those users that tried to progress for longer. However, coordination might not be so desirable in this occasion. The problem with players conforming with the majoritarian direction or mimicking each other is that they will be subject to herding effects [367, 368] which in this particular setting can be catastrophic due to the lag present in the system. Indeed, if we set the probabilities in our model so that the next state in the transition is always the one that gets you closer to the exit but with 25 seconds of delay (that is, the probability of going from state x_m to x_{m+1} is the probability of going from x_{m-n} to the state which follows the optimal path), the system gets stuck in a loop and is never able to reach the exit.

Nevertheless, the chat analysis shows that players were not perfectly coordinated. Thus, to make our model more realistic we consider that each time step there are 100 users with different behaviors introducing commands. In particular, we consider variable quantities of noisy users who play completely at random, griefers who only press down to annoy the rest of the crowd and the herd who always sends the optimal command to get to the exit. The results, figure 4.11C, show that the addition of noise to the herd breaks the loops and allows the swarm to get to the exit. In particular, for the case with no griefers we find that with 1 percent of users adding noise to the input the mean time needed to finish this part is almost 3,000 hours. However, as we increase the noise, time is quickly reduced with an optimal noise level of around 40% of the swarm. Conversely, the introduction of griefers in the model, as expected, increases the time needed to finish this part in most cases. Interestingly though, for low values of the noise, the addition of griefers can actually be beneficial for the swarm, allowing the completion of this area in times compatible to the observed ones. Indeed, by breaking the herding effect, griefers are unintentionally helping the swarm to reach their goal.

Whether the individuals categorized as noise were producing it unintentionally or doing

it on purpose to disentangle the crowd (an unknown fraction of users were aware of the effects of the lag and they tried to disentangle the system [369]) is something we can not analyze because, unfortunately, the resolution of the chat log in this area is in minutes and not in seconds. We can, however, approximate the fraction of griefers in the system thanks to the special characteristics of this area. Indeed, as most of the time the command *down* is not needed −on the contrary, it would destroy all progress−, we can categorize those players with an abnormal number of *downs* as griefers. To do so, we take the users that belong to $G6$ (the most active ones) and compare the fraction of their inputs that corresponds to *down* between each other. We find that 7% have a behavior that could be categorized as outlier (the fraction of their input corresponding to *down* is higher than 1.5 times the inter quartile range). More restrictively, for 1% of the players, the command *down* represents more than half of their inputs. Both these values are compatible with the observed time according to our model, even more if we take into account that the model is more restrictive as we consider that griefers continuously press down (not only near the red nodes). Thus, we conclude that users deviating from the norm, regardless of being griefers, noise or even very smart individuals, were the ones that made finishing this part possible.

4.2.3 The politics of the crowd

As we already mentioned, on the sixth day of the game the input system was modified. This resulted in the *start*9 riot that led to the introduction of the *anarchy/democracy* system. From this time on, if the fraction of users sending *democracy*, out of the total amount of players sending the commands *anarchy* or *democracy*, went over 0.75 (later modified to 0.80) the game would enter into democracy mode and commands would be tallied up for 5 seconds. Then, the meter needed to go below 0.25 (later modified to 0.50) to enter into anarchy mode again. Note that these thresholds were set by the creator of the experiment.

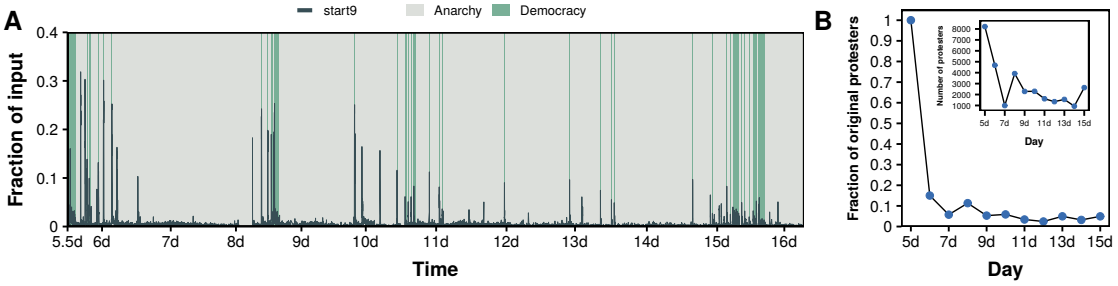

Figure 4.12: Overview of *start*9 protests throughout the game. A) Fraction of input corresponding to the *start*9 command. B) Fraction of users who where in the original *start*9 riot (inset, total number of protesters each day). There were *start*9 protests 10 days after the first one even though less than 10% of the protesters had been part of the first one.

The introduction of the voting system was mainly motivated by a puzzle where the crowd had been stuck for over 20 hours with no progress. Nonetheless, even in democracy mode, progress was complex as it was necessary to retain control of the game mode plus taking into account lag when deciding which action to take. Actually, the tug of war system was introduced at the middle of day 5, yet the puzzle was not fully completed until the beginning of day 6, over 40 hours after the crowd had originally arrived to the puzzle. One of the reasons why it took so long to finish it even after the introduction of the voting system is that it was very difficult to enter into democracy mode. Democracy was only "allowed" by the crowd when they were right in front of the puzzle and they would go into anarchy mode quickly after finishing it. Similarly, the rest of the game was mainly played under anarchy mode. Interestingly, though, we find that there were more "democrats" in

the crowd (players who only voted for democracy) than "anarchists" (players who only voted for anarchy). Out of nearly 400,000 players who participated in the tug of war throughout the game, 54% were democrats, 28% anarchists and 18% voted at least once for both of them. Therefore, the introduction of this new system did not only split the crowd into two polarized groups with, as we shall see, their own norms and behaviors, but also created non trivial dynamics between them.

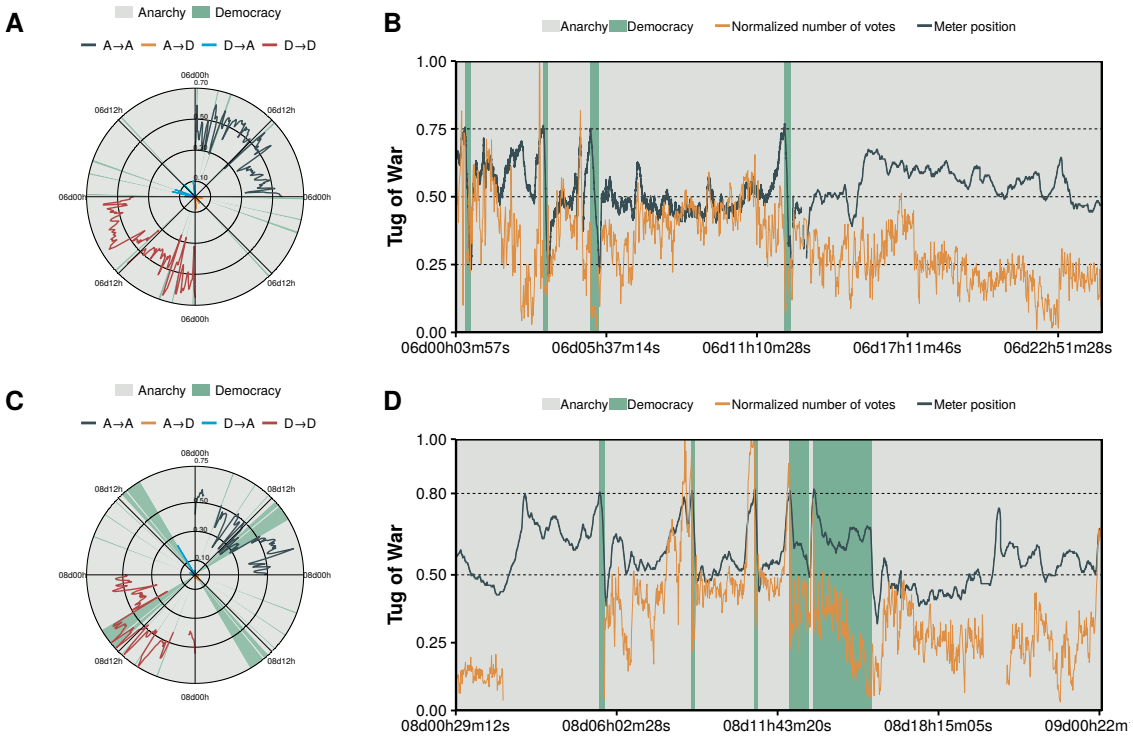

Figure 4.13: Politics of the crowd. Days 6 (top) and 8 (bottom). In every plot the gray color denotes when the game was played under anarchy rules and the green color when it was played under democracy rules. The polar plots represent the evolution of the fraction of votes corresponding to anarchy/democracy while distinguishing if the user previously voted for anarchy or democracy: first quadrant, votes for anarchy coming from users who previously voted for anarchy ($A \to A$); second quadrant, votes for democracy coming from anarchy ($A \to D$); third quadrant, votes for democracy coming from democracy ($D \to D$); fourth quadrant, votes for anarchy coming from democracy ($D \to A$). In the other plots we show the evolution of the total number of votes for anarchy or democracy as a function of time normalized by its maximum value (orange) as well as the position of the tug of war meter (blue). When the meter goes above 0.75 the system enters into democracy mode (green) until it reaches 0.25 (these thresholds were latter changed to 0.80 and 0.50 respectively) when it enters into anarchy mode (gray) again. The gap in the orange curve in panel D is due to the lack of data in that period.

The first question that arises is what might have motivated players to join into one group or the other. From a broad perspective, it has been proposed that one of the key ingredients behind video game enjoyment is the continuous perception of one's causal effects on the environment, also known as effectance [370], thanks to their immediate response to player inputs. In contrast, a reduction of control, defined as being able to influence the dynamics according to one's goals, does not automatically lower enjoyment [371]. This might explain why some people preferred anarchy. Under its rules, players saw that the game was continuously responding to inputs, even if they were not exactly the

ones they sent. On the other hand, with democracy, control was higher at the expense of effectance, as the game would only advance once every 5 seconds. The fact that some people might have preferred anarchy while others democracy is not surprising as it is well known that different people might enjoy different aspects of a game [372]. In the classical player classification proposed by Bartle [373] for the context of MUDs (multi-user dungeon, which later evolved into what we now today as MMORPGs - massively multiplayer online role-playing games) he already distinguished four types of players: achievers, who focus on finishing the game (who in our context could be related to democrats); explorers, who focused on interacting with the world (anarchists); socializers, who focused on interacting with other players (those players who focused on making fan art and developing narratives); and killers, whose main motivation was to kill other players (griefers). Similarly, it has been seen in the context of FPSs (first person shooters) that player-death events, i.e., loosing a battle, can be pleasurable for some players (anarchists) while not for others (democrats) [374].

However, when addressing the subject of video games entertainment, it is always assumed that the player has complete control over the character, regardless of whether it is a single player game or a competitive/cooperative game. TPP differs from those cases in the fact that everyone controlled the same character. As a consequence, enjoyment is no longer related to what a player, as a single individual, has done but rather to what they, as a group, have achieved. From the social identity approach perspective this can be described as a shift from the personal identity to the group identity. This shift would increase conformity to the norms associated to each group but as the groups were unstructured their norms would be inferred from the actions taken by the rest of the group [318]. New group members would then perform the actions they saw appropriate for them as members of the group, even if they might be seen as antinormative from an outside perspective [375]. This key component of the theory is clearly constated in the behavior of the anarchists. Indeed, every time the game entered in democracy mode, anarchists started to send *start*9 as a form of protest, hijacking the democracy. Interestingly, this kept happening even though most of the players who were in the original protest did not play anymore (see figure 4.12). Thus, newcomers adopted the identity of the group even if they had not participated in its conception. Even more, stalling the game might have been regarded as antisocial behavior from the own anarchists point of view when they were playing under anarchy rules, but when the game entered into democracy mode it suddenly turned into an acceptable behavior, something that is predicted by the theory.

To further explore the dynamics of these two groups, we next compare two different days: day 6 and day 8. Day 6 was the second day after the introduction of the anarchy/democracy dynamics and there were not any extremely difficult puzzles or similar areas where democracy might have been needed. On the other hand, day 8 was the day when the crowd arrived to the safari zone, which certainly needed democracy mode since the available number of steps in this area is limited (i.e., once the number of steps taken inside the area exceeds 500, the player is teleported to the entrance of the zone). We must note that, contrary to what we observed in section 4.2.2, in this case commands coming from low activity users are not equivalent to the ones coming from high activity users. In particular, low activity users tend to vote much more for democracy (see figure 4.14). As a consequence, although if we only take out the users with just 1 vote the position of the meter is unaffected, if we remove users with less than 10 votes the differences start to be noticeable. As such, it would not be adequate to remove low activity users in general from the analysis. Our results are summarized in figure 4.13.

One of the most characteristic features of groups is their polarization [282, 376]. The problem in the case we are studying is that as players were leaving the game while others were constantly coming in, it is not straightforward to measure polarization. The fact that

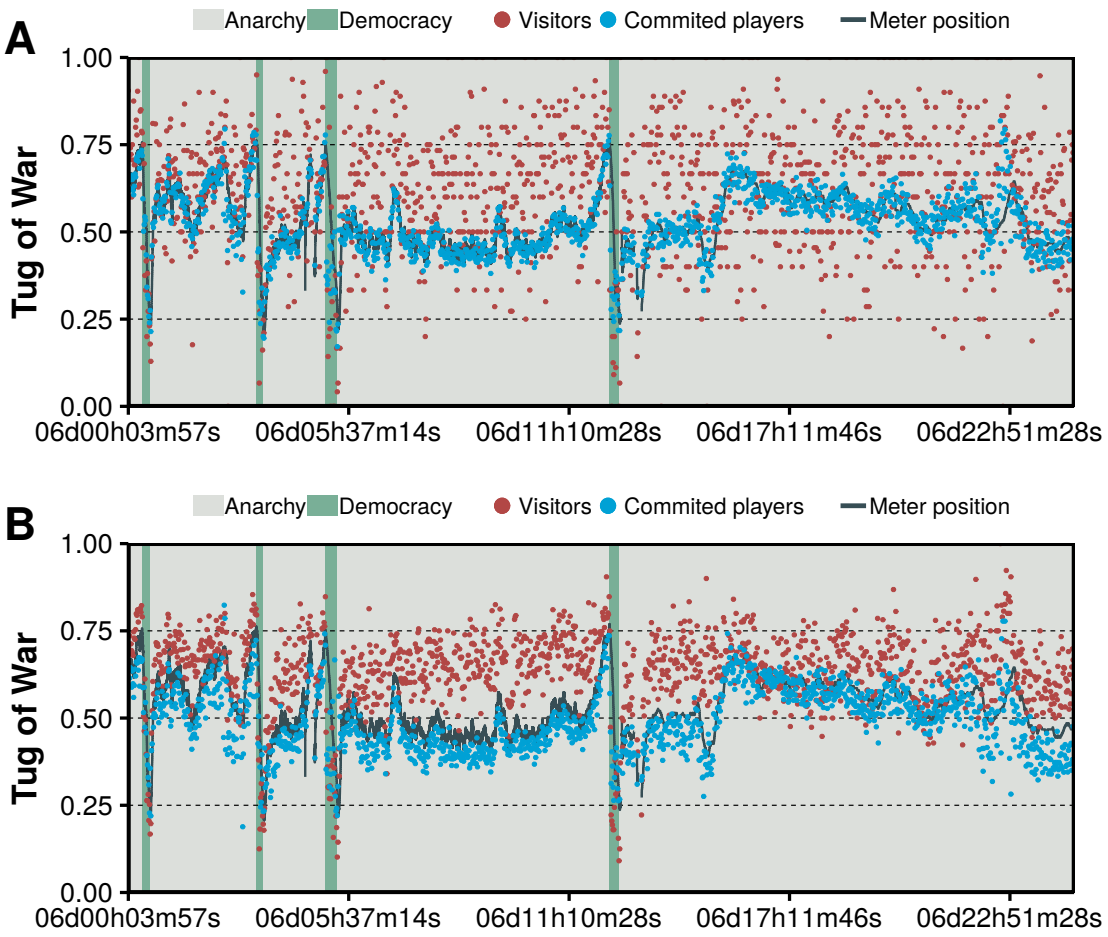

Figure 4.14: Tug of war commitment. Hypothetical meter position of the political tug of war if only votes from commited players - those who sent at least 2 votes (top) or 10 votes (bottom) throughout the whole game - are taken into account (blue) and if only votes from visitors - only one vote (top) or less than 10 (bottom) - are taken into account (red). In contrast to the ledge event, the behavior of users who sent few commands clearly differs from the ones with several commands. Visitors had a clear tendency towards democracy, while committed players preferred anarchy.

the number of votes for democracy could increase at a given moment did not mean that anarchists changed their opinion, it could be that new users were voting for democracy or simply that players who voted for anarchy stopped voting. Then, to properly measure polarization we consider 4 possible states for each user. They are defined by both the current vote of the player and the immediately previous one (note that we have removed players who only voted once, but this does not affect the measure of the position of the meter, see figure 4.14A): $A \to A$, first anarchy then anarchy; $A \to D$, first anarchy then democracy; $D \to D$, first democracy then democracy; $D \to A$, first democracy then anarchy. As we can see in figures 4.13A and 4.13C the communities are very polarized, with very few individuals changing their votes. The fraction of users changing from anarchy to democracy is always lower than 5%, which indicates that anarchists form a very closed group. Similarly, the fraction of users changing from democracy to anarchy is also very low, although there are clear bursts when the crowd exits democracy mode. This reflects that those who changed their vote from anarchy to democracy do so to achieve a particular goal, such as going through a mace, and once they achieve the target they instantly lose

interest in democracy.

With such degree of polarization the next question is how was it possible for the crowd to change from one mode to the other. To do so, we shift our attention to the number of votes. In figure 4.13B we can see that every time the meter gets above the democracy threshold it is preceded by an increase in the total number of votes. Then, once under democracy mode, the total number of votes decays very fast. Finally, there is another increment before entering again into anarchy mode. Thus, it seems that every time democrats were able to enter into their mode they stopped voting and started playing. This let anarchists regain control even though they were less users, leading to a sharp decay of the tug of war meter. Once they exited democracy mode, democrats started to vote again to try to set the game back into democracy mode. In figure 4.13D we can see initially a similar behavior in the short periods when democracy was installed. However, there is a wider area were the crowd accepted the democracy, this marks the safari zone mentioned previously. Interestingly, we can see how democrats learned how to keep their mode active. Initially there was the same drop on users voting and on the position of the meter seen in the other attempts. This forced democrats to keep voting instead of playing, which allowed them to retain control for longer. Few minutes later the number of votes decays again but in this case the position of the meter is barely modified probably due to anarchists finally accepting that they needed democracy mode to finish this part. Even though they might have implicitly accepted democracy, it is worth noting that the transitions $A \rightarrow D$ are minimum (figure 4.13C). Finally once the mission for which the democracy mode was needed finished, there is a sharp increment in the fraction of transitions $D \rightarrow A$.

4.2.4 The challenges of digital crowds

In this section we have analyzed a crowd based event where nearly 1 million users played a game with the exact same character. Remarkably, the event was not only highly successful in terms of participants but also in length, lasting for over two weeks. As we discussed in the introduction of section 4.2, motivating a crowd to complete a project is not an easy task. Yet, this event is an example that this can happen even in the absence of any material reward, signaling once again that online crowds have their own rules which might depart from what has been studied in the offline world.

Although the overall success of the event is probably due to a mixture of many factors, there is one that we can extract from the chat logs which is quite interesting. The game was disordered, progress was slower than if played individually, and often really bad actions were taken (such as mistakenly releasing some of the strongest Pokémons) which might have led to frustration. Indeed, by looking at the stretchable words sent by the users [377] it is possible to measure the frustration the players felt during the event, figure 4.15. Although usually frustration has a negative connotation, in the context of games it has been observed that frustration and stress can be pleasurable as they motivate players to overcome new challenges [378]. Actually, there is a whole game genre known as “masocore” (a portmanteau of masochism and hardcore) which consists of games with extremely challenging gameplay built with the only purpose of generating frustration on the players [379]. Similarly, there are games which might be simpler but that have really difficult controls and strange physics, such as QWOP, Surgeon Simulator or Octodad, which are also built with the sole aim of generating frustration [380]. Thus, the mistakes performed by the crowd might have not been something dissatisfactory but completely the opposite, they might have been the reason why this event was so successful.

One of the particularly frustrating areas was the ledge, a part of the game that can be completed in a few minutes but that took over 15 hours to complete. We have seen that in this area the behavior of low and high activity users is quite similar, even though they might have been unaware of it. Besides, we have built a model to explain how the crowd was able to finally exit this part and shown how a minority - either in the form of griefers,

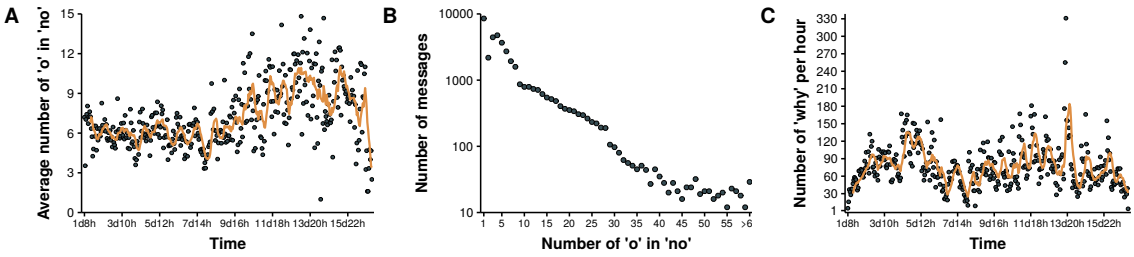

Figure 4.15: Measures of frustration. A) Players expressed their frustration by adding more times the letter *o* when they wanted to say *no*. Even though frustration was present throughout the event, it was incremented after the events of what is known as *Bloody Sunday*. B) Distribution of the number of *o*. Interestingly, the relationship is not linear as the word *noo* tends to appear less than *nooo* or *noooo*, which indicates that when players were frustrated they overexpressed it. C) Number of messages containing the word *why* per hour. This indicates that many players did not understand the actions of the crowd, which probably made them feel frustrated.

smart users or simply naïve individuals - can lead the crowd to a successful outcome, even in the lack of consensus. Note also that the fact that they only needed roughly 1/3 of the time to traverse the area on their second attempt compared to their first one might be a signal of the crowd learning how to break the herding effect. Unfortunately, with just two observations we cannot test this hypothesis. It would be interesting, though, to design experiments inspired by this event with the purpose of measuring if the crowd is able to learn and, if it does so, how long does it take, what would happen if a fraction of the crowd is substituted by new players, etc.

To conclude this section, we have also analyzed the effects that the introduction of a voting system had in the crowd. We have seen how the crowd was split into two groups and we have been able to explain the behavior of these groups using the social identity approach. We saw how norms could last within groups longer than their own members, as predicted by the theory. Note that this theory was introduced during the 1980s, way before Internet was as widespread as it is today, and still it can be applied, at least in this case, to online groups. Hence, despite the many differences that exist between the online and offline worlds, maybe they are not that far apart after all.

5 Conclusions

Do you know the saying "The whole is greater than the sum of the parts?" It is an insane statement. It is a nonsense. But now I believe that it is true.

("Thief of Time", Terry Pratchett)

We began this thesis showing that usually science advances in small steps, rather than in big leaps. This observation is even truer for complex systems, where a unified theory does not exist yet (if it ever does) and all we have is a collection of short stories.

In chapter 2 we focused on studying one of the most important tools used in complex systems, networks. In particular, we addressed the problem of how to create adequate null models for a network as a function of the availability of data. Recall that the most naïve approximation is to use random graphs, as something to compare the real network against. However, this procedure has two main drawbacks. First, it is possible that the microscopic structure of the network is such that it yields higher order structures not present in random graphs. Although this is clearly valuable information, it might fool us into thinking that the system under consideration evolved to specifically create those structures. Instead, it is possible that they are just a direct consequence of lower order properties, which should then be the focus of our research.

The other main issue is that, as networks can be used in a huge amount of systems with very diverse characteristics, comparing a real network with a completely random graph might lead us to think that the network is stranger than it actually is. For instance, it has been observed that in friendship networks the number of triangles (i.e., if A and B are friends and B and C also, then A and C are friends too) is much larger than in random networks. This is indeed an important property of these networks. Yet, if we were given a new friendship network, we might not be interested in measuring if the number of triangles is larger than expected at random, because we already know that it will probably be. Instead, it might be more enlightening to check if that number is higher than in other friendship networks, or in a null model that reflects the common characteristics of these networks. This example can be extended to a lot of different systems. As a consequence, there is almost one null model for each application that we can think of.

When we wanted to study the anomalies present in the betweenness centrality of transportation networks, we could have created a null model that specifically took into account the characteristics of these systems, such as the population living in each municipality, their size, etc. Instead, we chose to follow the framework inspired by Jaynes of seeing statistical physics as a problem of information, which yielded the exponential random graph model. This more general framework allowed us to determine that the observed anomalies were just a consequence of the weight distribution. Hence, rather than wondering why these networks exhibit those anomalies, the focus should be on studying the mechanisms leading to those weight distributions. Furthermore, this study also highlighted the importance of having the proper amount of data.

We concluded the chapter applying the formalism to create multilayer contact networks using data of both the contact distribution of individuals and their age mixing patterns. We showed how said information can be extracted from real datasets and introduced into

the model to generate realistic contact networks. Note that in this case we did not want to create a null model for comparison processes, but rather to build networks in an unbiased way given the available data. Thus, another advantage of the exponential random graph model is that it unifies several problems into one unique framework.

In chapter 3, we focused on the mathematical study of epidemic spreading. We followed the historical development of the field, from the simplest approximation to highly detailed models. Moreover, at each step, we observed the influence of including more data in the models. In the first case, we saw that the challenge of incorporating data is not only restricted to the problem of obtaining it, but that it is also really important to be aware of its characteristics. In particular, we saw that the age contact matrices cannot be naïvely applied to any population, as they already encode implicitly some information about it. Thus, if the population changes, the matrices also have to change.

Next, we created a highly realistic numerical model for the spreading of influenza-like diseases and showed that common theoretical assumptions might not be good enough to capture the complexity of the process. In particular, we observed that the definition of one of the most important quantities in epidemiology, the basic reproduction number, does not successfully capture the real dynamics of the epidemic. The reason was that the mathematical definition relies on several assumptions that are invalidated by data.

On the other hand, the third study was focused on extending the theoretical models of disease spreading in multilayer networks to the case in which the direction of the links is known. We showed that populations in which the underlying network possesses some directionality are more resilient against an epidemic that those that are completely undirected. Admittedly, thanks to online social platforms, it is much easier to obtain this information for social systems. Nevertheless, the basic formulation can be easily adapted to analyze the spreading of information. Hence, our results also imply that the role that platforms in which the communication is undirected can play a very different role from the ones that are directed.

Lastly, we analyzed the consequences of using different epidemic models as a function of data availability, with particular emphasis on the networks that we created at the end of chapter 2. We saw that the more data we have, the better, but that for some applications the simplest models with less data can also provide valuable information. In particular, even though knowledge of the underlying network is crucial to determine the epidemic threshold, information on the age structure of the population is essential for the correct definition of risk groups.

To conclude, in chapter 4, we analyzed two examples of collective social behavior using data extracted from very different sources. First, we showed that Hawkes processes can be effectively used for the analysis of online boards such as Forocoches. Furthermore, we were able to distinguish two different types of activity, one that was independent from the rest of the users and another one in which the social component of the process was indispensable. We finished the chapter studying an online crowd event, Twitch Plays Pokémon. We saw that despite its unique characteristics, some properties of the online crowd were similar to the ones that offline crowds exhibit signaling, once again, that modern societies are intertwined with the online world.

To sum up, we have overviewed a tiny fraction of the field of complex systems, with special emphasis on the role that new data can have in problems ranging from the most theoretical work to highly realistic computer simulations. We hope that this collection of short stories will show the huge diversity of problems that are still open in the field of complex systems and, at the same time, shed some light on them.

5.1 Future work

There are multiple ways in which the results presented in this thesis can be extended, either to deepen our knowledge of particular systems or to increase our global understanding of complex systems. Some of them have already started, while others are currently just projects.

In chapter 2, we focused on studying the exponential random graph model. Despite its many advantages, it is important to bear in mind that it also has its drawbacks. For instance, the computational resources needed for numerically obtaining the parameters of the model can be quite large, depending on the size and characteristics of the networks. Furthermore, there are currently many researchers, coming from very different fields, who want to use networks but lack the technical background needed to understand this model. For this reason, one of the next steps will be to summarize all the possible null models found in the literature, systematically studying their advantages and disadvantages. The objective of this work will be to provide a reference to those researchers working on complex systems who might not be used to study networks and, hence, are not aware of the pitfalls that simple techniques can have.

Regarding chapter 3, note that the main driver of the four studies was data: (1) how to handle data; (2) theory vs data-driven simulations; (3) improving theories in light of new data; and (4) combining data. Thus, in future works we will continue to explore new data sources, sometimes using them to improve theoretical approaches, at other times to create more realistic simulations in which many different types of data can be combined. For instance, we are currently studying a new dataset which contains information about the daily routines of workers in a hospital, with the objective of devising effective strategies for the reduction of the spreading of health care associated infections.

Lastly, in chapter 4, we saw two examples of online collective behavior. For the case of Forocoches, there is still a lot of work to do. Regarding its dynamics, we can add non-constant background intensities to better characterize the behavior of threads, or explore further the relation between the success of a thread and its content (or the users that participate in it). Furthermore, the data itself can also be used to study the creation and evolution of memes as we hinted, or as a complement for the analysis of events that are currently mainly studied using data from Twitter. On the other hand, the work of Twitch Plays Pokémon can be considered, for the moment, closed, although it has sprouted some ideas about mimicking the rules of the game in simpler settings, in order to be able to perform controlled experiments on the behavior and organization of crowds. Nevertheless, this chapter has also shown that there are many new research opportunities in online systems, sometimes with connections to the "classical" offline world, but at other times with completely different characteristics. We will be vigilant, and as new data appears and new phenomena are uncovered, we will explore them.

6 Bibliography

[1] J. W. Gibbs, *Elementary principles in statistical mechanics*. Scribner's sons, 1902.

[2] S. Carnot, *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*. Bachelier, 1824.

[3] S. Carnot, *Reflections on the motive power of heat*. John Wiley & Sons, 1897.

[4] B. Clapeyron, *Scientific Memoirs*, ch. Memoir on the Motive Power of Heat. Richard and John E. Taylor, 1837.

[5] R. Clausius, *The Mechanical Theory of Heat with its Applications to the Steam Engine and to the Physical Properties of Bodies*, ch. On several convenient forms of the fundamental equations of the mechanical theory of heat. Taylor and Francis, 1867.

[6] R. Clausius, *The Mechanical Theory of Heat with its Applications to the Steam Engine and to the Physical Properties of Bodies*, ch. On a modified form of the second fundamental theorem in the mechanical theory of heat. Taylor and Francis, 1867.

[7] S. G. Brush, *The Kind Of Motion We Call Heat. Volume 1*. North Holland, 1986.

[8] L. Campbell and W. Garnett, *The Life of James Clerk Maxwell*. Macmillan and Co., 1882.

[9] E. Garber, Aspects of the Introduction of Probability into Physics, *Centaurus*, vol. 17, pp. 11–40, Mar 1973.

[10] R. Clausius, On the mean length of the paths described by the separate molecules of gaseous bodies on the occurrence of molecular motion: together with some other remarks upon the mechanical theory of heat, *Philos. Mag.*, vol. 17, pp. 81–91, Feb 1859.

[11] B. Mahon, *The man who changed everything*. Wiley, 2003.

[12] J. C. Maxwell, On the Stability of the Motion of Saturn's Rings; an Essay which obtained the Adams' Prize for the Year 1856, in the University of Cambridge, *Mon. Notices Royal Astron. Soc*, vol. 19, pp. 297–304, Jun 1859.

[13] J. C. Maxwell, Illustrations of the dynamical theory of gases.—Part I. On the motions and collisions of perfectly elastic spheres, *Philos. Mag.*, vol. 19, pp. 19–32, Jan 1860.

[14] M. J. Klein, The Development of Boltzmann's Statistical Ideas, in *The Boltzmann Equation* (E. G. D. Cohen and W. Thirring, eds.), pp. 53–106, Springer Vienna, 1973.

[15] C. Cercignani, *Chance in Physics: Foundations and Perspectives*, ch. The Rise of Statistical Mechanics, pp. 25–38. Springer Berlin Heidelberg, 2001.

[16] S. G. Brush, *The Kind Of Motion We Call Heat. Volume 2*. North Holland, 1986.

[17] J. L. Lebowitz and O. Penrose, Modern ergodic theory, *Phys. Today*, vol. 26, p. 23, Dec 1973.

[18] G. Eknoyan, Adolphe Quetelet (1796–1874)—the average man and indices of obesity, *Nephrol. Dial. Transplant.*, vol. 23, pp. 47–51, Sep 2007.

[19] A. Comte and H. Martineau, *The Positive Philosophy of Auguste Comte*. Cambridge University Press, 2009.

[20] G. Jahoda, Quetelet and the emergence of the behavioral sciences, *SpringerPlus*, vol. 4, pp. 1–10, Dec 2015.

[21] H. W. Watson and F. Galton, On the Probability of the Extinction of Families, *J. Royal Anthropol. Inst.*, vol. 4, pp. 138–144, 1875.

[22] D. G. Kendall, Branching Processes Since 1873, *J. London Math. Soc.*, vol. s1-41, pp. 385–406, Jan 1966.

[23] F. Galton, *Inquiries into Human Faculty and Its Development*. Macmillan, 1883.

[24] S. S. Schweber, Physics, Community and the Crisis in Physical Theory, *Phys. Today*, vol. 46, pp. 34–40, Nov 1993.

[25] A. N. Shiryayev, *Selected Works of A. N. Kolmogorov*. Springer, Dordrecht, 1992.

[26] T. E. Harris, *The Theory of Branching Processes*. Springer-Verlag Berlin Heidelberg, 1963.

[27] N. Metropolis, The Beginning of the Monte Carlo Method, *Los Alamos Science*, vol. 15, 1983.

[28] N. Bacaër, *A Short History of Mathematical Population Dynamics*. Springer, London, 2011.

[29] S. R. Broadbent and J. M. Hammersley, Percolation processes: I. Crystals and mazes, *Math. Proc. Cambridge Philos. Soc.*, vol. 53, pp. 629–641, Jul 1957.

[30] C. E. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, vol. 27, pp. 379–423, Jul 1948.

[31] E. T. Jaynes, Information Theory and Statistical Mechanics, *Phys. Rev.*, vol. 106, pp. 620–630, May 1957.

[32] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Principles of maximum entropy and maximum caliber in statistical physics, *Rev. Mod. Phys.*, vol. 85, pp. 1115–1141, Jul 2013.

[33] H. B. Callen, *Thermodynamics and an Introducion to Termostatistics*. John Wiley & Sons, 1985.

[34] P. W. Bridgman, *The Nature of Thermodynamics*. Harvard University Press, 1943.

[35] A. Ben-Naim, *A Farewell to Entropy*. World Scientific Publishing Company, 2008.

[36] P. W. Anderson, More Is Different, *Science*, vol. 177, pp. 393–396, Aug 1972.

[37] L. Pietronero, Complexity ideas from condensed matter and statistical physics, *Europhys. News*, vol. 39, pp. 26–29, Nov 2008.

[38] P. W. Anderson, *More and Different*. World Scientific Publishing Company, Sep 2011.

[39] M. Mezard and A. Montanari, *Information, Physics, and Computation*. New York, NY, USA: Oxford University Press, Inc., 2009.

[40] P. Castiglione, M. Falcioni, A. Lesne, and A. Vulpiani, *Chaos and Coarse Graining in Statistical Mechanics*. Cambridge University Press, Aug 2008.

[41] E. N. Lorenz, Deterministic Nonperiodic Flow, *J. Atmospheric Sci.*, vol. 20, pp. 130–141, Mar 1963.

[42] E. N. Lorenz, Section of Planetary Sciences: The Predictability of Hydrodynamics Flow, *Trans. N. Y. Acad. Sci.*, vol. 25, pp. 409–432, Feb 1963.

[43] R. M. May, Simple mathematical models with very complicated dynamics, *Nature*, vol. 261, pp. 459–467, Jun 1976.

[44] M. Schroeder, *Fractals, Chaos, Power Laws*. W. H. Freeman and Company, 1991.

[45] P. Bak, C. Tang, and K. Wiesenfeld, Self-organized criticality: An explanation of the 1/f noise, *Phys. Rev. Lett.*, vol. 59, pp. 381–384, Jul 1987.

[46] N. W. Watkins, G. Pruessner, S. C. Chapman, N. B. Crosby, and H. J. Jensen, 25 Years of Self-organized Criticality: Concepts and Controversies, *Space Sci. Rev.*, vol. 198, pp. 3–44, Jan 2016.

[47] C. S. Holling, Understanding the Complexity of Economic, Ecological, and Social Systems, *Ecosystems*, vol. 4, pp. 390–405, Aug 2001.

[48] H. A. Simon, The Architecture of Complexity, *Proc. Am. Philos. Soc*, vol. 106, pp. 467–482, Dec 1962.

[49] M. E. J. Newman, Resource Letter CS–1: Complex Systems, *Am. J. Phys*, vol. 79, p. 800, Jul 2011.

[50] G. Parisi, Complex systems: a physicist's viewpoint, *Physica A*, vol. 263, pp. 557–564, Feb 1999.

[51] J. L. Moreno, *Who shall survive?: A new approach to the problem of human interrelations*. Nervous and Mental Disease Publishing Co, 1934.

[52] M. Zhang, Social Network Analysis: History, Concepts, and Research, in *Handbook of Social Network Technologies and Applications*, pp. 3–21, Springer US, 2010.

[53] S. Milgram, The Small-World Problem, *Psychol. Today*, vol. 1, May 1967.

[54] M. Gluckman, *The judicial process among the barotse of northern rhodesia*. Manchester University Press, 1955.

[55] J. Park and M. E. J. Newman, Statistical mechanics of networks, *Phys. Rev. E*, vol. 70, p. 066117, Dec 2004.

[56] P. W. Holland and S. Leinhardt, An Exponential Family of Probability Distributions for Directed Graphs, *J. Am. Stat. Assoc.*, vol. 76, pp. 33–50, Mar 1981.

[57] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature*, vol. 393, pp. 440–442, Jun 1998.

[58] D. J. Watts, *Small Worlds – The Dynamics of Networks between Order and Randomness*. Princeton University Press, 2003.

[59] S. H. Strogatz, Exploring complex networks, *Nature*, vol. 410, pp. 268–276, Mar 2001.

[60] H. E. Stanley, V. Afanasyev, L. A. N. Amaral, S. V. Buldyrev, A. L. Goldberger, S. Havlin, H. Leschhorn, P. Maass, R. N. Mantegna, C.-K. Peng, P. A. Prince, M. A. Salinger, M. H. R. Stanley, and G. M. Viswanathan, Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics, *Physica A*, vol. 224, pp. 302–321, Feb 1996.

[61] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli, The statistical physics of real-world networks, *Nat. Rev. Phys.*, vol. 1, pp. 58–71, Jan 2019.

[62] G. Caldarelli, S. Wolf, and Y. Moreno, Physics of humans, physics for society, *Nat. Phys.*, vol. 14, p. 870, Sep 2018.

[63] J. A. P. Heesterbeek and M. G. Roberts, How mathematical epidemiology became a field of biology: a commentary on Anderson and May (1981) 'The population dynamics of microparasites and their invertebrate hosts', *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 370, p. 20140307, Apr 2015.

[64] D. Lazer, R. Kennedy, G. King, and A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis, *Science*, vol. 343, pp. 1203–1205, Mar 2014.

[65] D. Lazer and J. Radford, Data ex Machina: Introduction to Big Data, *Annu. Rev. Sociol.*, vol. 43, pp. 19–39, Jul 2017.

[66] D. A. McFarland, K. Lewis, and A. Goldberg, Sociology in the Era of Big Data: The Ascent of Forensic Social Science, *Am. Soc.*, vol. 47, pp. 12–35, Mar 2016.

[67] S. Halford and M. Savage, Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research, *Sociology*, vol. 51, pp. 1132–1148, Jun 2017.

[68] J. Henrich, S. J. Heine, and A. Norenzayan, The weirdest people in the world?, *Behav. Brain Sci.*, vol. 33, pp. 61–83, Jun 2010.

[69] A. Martín-Martín, E. Orduna-Malea, J. M. Ayllón, and E. D. López-Cózar, Back to the past: on the shoulders of an academic search engine giant, *Scientometrics*, vol. 107, pp. 1477–1487, Mar 2016.

[70] A. Acharya, A. Verstak, H. Suzuki, S. Henderson, M. Iakhiaev, C. C. Y. Lin, and N. Shetty, Rise of the Rest: The Growing Impact of Non-Elite Journals, 2014.

[71] N. Marres, *Digital Sociology: The Reinvention of Social Research*. Wiley, 2017.

[72] M. Castells, *The Rise of the Network Society 2nd Ed.* John Wiley & Sons, 2010.

[73] E. O. Wilson, *Consilience: the unity of knowledge*. Vintage Books, 1999.

[74] M. Newman, *Networks*. Oxford University Press, Oct 2018.

[75] V. Latora, V. Nicosia, and G. Russo, *Complex Networks*. Cambridge University Press, Feb 2019.

[76] Albert-László, *Network Science*. Cambridge University Press, 2016.

[77] A. Aleta and Y. Moreno, Multilayer Networks in a Nutshell, *Annu. Rev. Condens. Matter Phys.*, vol. 10, pp. 45–62, Mar 2019.

[78] L. G. A. Alves, A. Aleta, F. A. Rodrigues, Y. Moreno, and L. A. Nunes Amaral, Centrality anomalies in complex networks as a result of model over-simplification, *Sent for publication*, 2019.

[79] A. Aleta, G. Ferraz de Arruda, and Y. Moreno, Generating data-driven age contact networks, *In preparation*, 2019.

[80] P. Erdős and A. Rényi, On random graphs, *Publ. Math.*, vol. 6, p. 2, 1959.

[81] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, Multilayer networks, *J. Complex Netw.*, vol. 2, pp. 203–271, Jul 2014.

[82] A. Aleta, S. Meloni, and Y. Moreno, A Multilayer perspective for the analysis of urban transportation systems, *Sci. Rep.*, vol. 7, p. 44359, Mar 2017.

[83] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[84] C. Orsini, M. M. Dankulov, P. Colomer-de Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov, Quantifying randomness in real networks, *Nat. Commun.*, vol. 6, p. 8627, Oct 2015.

[85] C. Payrató-Borràs, L. Hernández, and Y. Moreno, Breaking the Spell of Nestedness: The Entropic Origin of Nestedness in Mutualistic Systems, *Phys. Rev. X*, vol. 9, Aug 2019.

[86] J. Fründ, K. S. McCann, and N. M. Williams, Sampling bias is a challenge for quantifying specialization and network structure: lessons from a quantitative niche model, *Oikos*, vol. 125, pp. 502–513, Apr 2016.

[87] M. A. M. de Aguiar, E. A. Newman, M. M. Pires, J. D. Yeakel, D. H. Hembry, C. Boettiger, L. A. Burkle, D. Gravel, P. R. Guimarães, J. L. O'Donnell, T. Poisot, and M.-J. Fortin, Revealing biases in the sampling of ecological interaction networks, *bioRxiv*, 2018.

[88] P. Pöyhönen, A Tentative Model for the Volume of Trade between Countries, *Rev. World Econ.*, vol. 90, pp. 93–100, 1963.

[89] G. A. P. Carrothers, An Historical Bedew of the Gravity and Potential Concepts of Human Interaction, *J. Am. Inst. Plan.*, vol. 22, pp. 94–102, Jun 1956.

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

[91] T. Squartini, R. Mastrandrea, and D. Garlaschelli, Unbiased sampling of network ensembles, *New J. Phys.*, vol. 17, p. 023052, Feb 2015.

[92] S. Maslov and K. Sneppen, Specificity and Stability in Topology of Protein Networks, *Science*, vol. 296, pp. 910–913, May 2002.

[93] The Koblenz Network Collection (KONECT), `http://konect.cc/`. Accessed: 2019-08.

[94] A. R. Rao, R. Jana, and S. Bandyopadhyay, A Markov Chain Monte Carlo Method for Generating Random (0, 1)-Matrices with Given Marginals, *Sankhya*, vol. 58, pp. 225–242, Jun 1996.

[95] B. Fosdick, D. Larremore, J. Nishimura, and J. Ugander, Configuring Random Graph Models with Fixed Degree Sequences, *SIAM Rev.*, vol. 60, no. 2, pp. 315–355, 2018.

[96] J. L. Moreno and H. H. Jennings, Statistics of Social Configurations, *Sociometry*, vol. 1, pp. 342–374, Jan 1938.

[97] E. F. Connor and D. Simberloff, The Assembly of Species Communities: Chance or Competition?, *Ecology*, vol. 60, pp. 1132–1140, Dec 1979.

[98] B. Bollobás, A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs, *Eur. J. Comb.*, vol. 1, pp. 311–316, Dec 1980.

[99] G. Caldarelli and A. Chessa, *Data Science and Complex Networks*. Oxford University Press, 2016.

[100] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Generation of uncorrelated random scale-free networks, *Phys. Rev. E*, vol. 71, Feb 2005.

[101] Wikipedia entry on degree-preserving randomization, `https://en.wikipedia.org/wiki/Degree-preserving_randomization`. Accessed: 2019-08.

[102] Z. Li, P. J. Mucha, and D. Taylor, Network-Ensemble Comparisons with Stochastic Rewiring and Von Neumann Entropy, *SIAM J. Appl. Math.*, vol. 78, pp. 897–920, Mar 2018.

[103] C. J. Carstens and K. J. Horadam, Switching edges to randomize networks: what goes wrong and how to fix it, *J. Complex Netw.*, vol. 5, pp. 337–351, Oct 2016.

[104] Y. Artzy-Randrup and L. Stone, Generating uniformly distributed random networks, *Phys. Rev. E*, vol. 72, p. 056708, Nov 2005.

[105] igraph function for randomly rewiring links, `https://igraph.org/r/doc/keeping_degseq.html`. Accessed: 2019-08.

[106] NetworkX function for randomly rewiring links, `https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.swap.double_edge_swap.html`. Accessed: 2019-08.

[107] A. C. C. Coolen, A. De Martino, and A. Annibale, Constrained Markovian Dynamics of Random Graphs, *J. Stat. Phys.*, vol. 136, pp. 1035–1067, Sep 2009.

[108] B. Liu, S. Xu, T. Li, J. Xiao, and X.-K. Xu, Quantifying the Effects of Topology and Weight for Link Prediction in Weighted Complex Networks, *Entropy*, vol. 20, p. 363, May 2018.

[109] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Cut-offs and finite size effects in scale-free networks, *Eur. Phys. J. B*, vol. 38, pp. 205–209, Mar 2004.

[110] S. W. Emmons, The beginning of connectomics: a commentary on White et al. (1986) 'The structure of the nervous system of the nematode Caenorhabditis elegans', *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 370, Apr 2015.

[111] J. G. White, E. Southgate, J. Thomson, and S. Brenner, The structure of the nervous system of the nematode Caenorhabditis elegans, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 314, Nov 1986.

[112] S. J. Cook, T. A. Jarrell, C. A. Brittin, Y. Wang, A. E. Bloniarz, M. A. Yakovlev, K. C. Q. Nguyen, L. T.-H. Tang, E. A. Bayer, J. S. Duerr, H. E. Bülow, O. Hobert, D. H. Hall, and S. W. Emmons, Whole-animal connectomes of both Caenorhabditis elegans sexes, *Nature*, vol. 571, pp. 63–71, Jul 2019.

[113] F. Reif, *Fundamentals of statistical and thermal physics*. McGraw-Hill, Inc., 1965.

[114] M. Kardar, *Statistical Physics of Particles*. Cambridge University Press, 2007.

[115] W. T. Grandy, *Foundations of Statistical Mechanics*. D. Reidel Publishing Company, 1987.

[116] A. Rinaldo, S. Petrović, and S. E. Fienberg, Maximum lilkelihood estimation in the β-model, *Ann. Statist.*, vol. 41, pp. 1085–1110, Jun 2013.

[117] T. Squartini and D. Garlaschelli, Analytical maximum-likelihood method to detect patterns in real networks, *New J. Phys.*, vol. 13, p. 083001, Aug 2011.

[118] E. T. Jaynes, *Papers on probability, statistics and statistical physics*. Kluwer Academic Publisher, 1989.

[119] G. Bianconi and A.-L. Barabási, Bose-Einstein Condensation in Complex Networks, *Phys. Rev. Lett.*, vol. 86, pp. 5632–5635, Jun 2001.

[120] A. Nicolosi, M. L. C. Leite, M. Musicco, C. Arici, G. Gavazzeni, and A. Lazzarin, The Efficiency of Male-to-Female and Female-to-Male Sexual Transmission of the Human Immunodeficiency Virus: A Study of 730 Stable Couples, *Epidemiology*, vol. 5, pp. 570–575, Nov 1994.

[121] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 7794–7799, May 2005.

[122] K.-I. Goh, B. Kahng, and D. Kim, Universal Behavior of Load Distribution in Scale-Free Networks, *Phys. Rev. Lett.*, vol. 87, p. 278701, Dec 2001.

[123] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, Attack vulnerability of complex networks, *Phys. Rev. E*, vol. 65, p. 056109, May 2002.

[124] A. Barrat, M. Barthélemy, and A. Vespignani, The effects of spatial constraints on the evolution of weighted complex networks, *J. Stat. Mech: Theory Exp.*, vol. 2005, p. P05003, May 2005.

[125] M. Barthélemy, Spatial networks, *Phys. Rep.*, vol. 499, pp. 1–101, Feb 2011.

[126] R. Mastrandrea, T. Squartini, G. Fagiolo, and D. Garlaschelli, Enhanced reconstruction of weighted networks from strengths and degrees, *New J. Phys.*, vol. 16, p. 043022, Apr 2014.

[127] Openflights database, `https://openflights.org/data.html`. Accessed: 2019-08.

[128] Flightaware flight tracker, `https://flightaware.com`. Accessed: 2019-08.

[129] Brazilian institute of geography and statistics, http://www.ibge.gov.br. Accessed: 2019-08.

[130] Brazilian national agency of land transport, http://www.antt.gov.br. Accessed: 2019-08.

[131] Open data portal of the british goverment, https://data.gov.uk. Accessed: 2019-08.

[132] Information about regular bus networks from the spanish goverment, http://www.bus.es. Accessed: 2019-08.

[133] J. Wallinga, W. J. Edmunds, and M. Kretzschmar, Perspective: human contact patterns and the spread of airborne infectious diseases, *Trends Microbiol.*, vol. 7, pp. 372–377, Sep 1999.

[134] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. D'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, Statistical physics of vaccination, *Phys. Rep.*, vol. 664, pp. 1–113, Dec 2016.

[135] P. S. Bearman, J. Moody, and K. Stovel, Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks, *Am. J. Sociol.*, vol. 110, pp. 44–91, Jul 2004.

[136] A. C. Ghani, C. A. Donnelly, and G. P. Garnett, Sampling biases and missing data in explorations of sexual partner networks for the spread of sexually transmitted diseases, *Stat. Med.*, vol. 17, pp. 2079–2097, Sep 1998.

[137] K. A. Fenton, A. M. Johnson, S. McManus, and B. Erens, Measuring sexual behaviour: methodological challenges in survey research, *Sex. Transm. Infect.*, vol. 77, pp. 84–92, Apr 2001.

[138] M. Ajelli, S. Parlamento, D. Bome, A. Kebbi, A. Atzori, C. Frasson, G. Putoto, D. Carraro, and S. Merler, The 2014 Ebola virus disease outbreak in Pujehun, Sierra Leone: epidemiology and impact of interventions, *BMC Med.*, vol. 13, pp. 1–8, Dec 2015.

[139] L. A. Meyers, M. E. J. Newman, M. Martin, and S. Schrag, Applying Network Theory to Epidemics: Control Measures for Mycoplasma pneumoniae Outbreaks, *Emerg. Infect. Dis.*, vol. 9, p. 204, Feb 2003.

[140] K. T. D. Eames and M. J. Keeling, Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, pp. 13330–13335, Oct 2002.

[141] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, A high-resolution human contact network for infectious disease transmission, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, pp. 22020–22025, Dec 2010.

[142] L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, and R. C. Brunham, Network theory and SARS: predicting outbreak diversity, *J. Theor. Biol.*, vol. 232, pp. 71–81, Jan 2005.

[143] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds, Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases, *PLoS Med.*, vol. 5, p. e74, Mar 2008.

[144] G. De Luca, K. Van Kerckhove, P. Coletti, C. Poletto, N. Bossuyt, N. Hens, and V. Colizza, The impact of regular school closure on seasonal influenza epidemics: a data-driven spatial transmission model for Belgium, *BMC Infect. Dis.*, vol. 18, pp. 1–16, Dec 2018.

[145] L. Fumanelli, M. Ajelli, P. Manfredi, A. Vespignani, and S. Merler, Inferring the Structure of Social Contacts from Demographic Data in the Analysis of Infectious Diseases Spread, *PLoS Comput. Biol.*, vol. 8, pp. 1–10, Sep 2012.

[146] P. Fronczak, A. Fronczak, and M. Bujok, Exponential random graph models for networks with community structure, *Phys. Rev. E*, vol. 88, p. 032810, Sep 2013.

[147] J. M. Hilbe, *Negative Binomial Regression*. Cambridge University Press, 2011.

[148] S. Arregui, A. Aleta, J. Sanz, and Y. Moreno, Projecting social contact matrices to different demographic structures, *PLoS Comput. Biol.*, vol. 14, pp. 1–18, Dec 2018.

[149] N. D. Wolfe, C. P. Dunavan, and J. Diamond, Origins of major human infectious diseases, *Nature*, vol. 447, pp. 279–283, May 2007.

[150] J. N. Hays, *Epidemics and Pandemics: Their Impacts on Human History*. ABC-CLIO, 2005.

[151] V. Barras and G. Greub, History of biological warfare and bioterrorism, *Clin. Microbiol. Infect*, vol. 20, pp. 497–502, jun 2014.

[152] K. E. Nelson and C. M. Williams, *Infectious Disease Epidemiology: Theory and Practice*. Jones & Barlett Learning, 2014.

[153] A. Morabia, *A History of Epidemiologic Methods and Concepts*. Springer Basel AG, 2004.

[154] I. M. Foppa, *A Historical Introduction to Mathematical Modeling of Infectious Diseases*. Academic Press, 2016.

[155] K. Dietz, The First Epidemic Model: A Historical Note on P.D. En'ko, *Australian J. Stat.*, vol. 30A, pp. 56–65, May 1988.

[156] W. H. Hamer, The Miltor Lectures on Epidemic Disease in England - The Evidence of Variability and of Persistency of Type, *The Lancet*, vol. 167, pp. 733–739, Mar 1906.

[157] H. Heesterbeek, The Law of Mass-Action in Epidemiology: A Historical Perspective, in *Ecological Paradigms Lost*, pp. 81–105, Elsevier, 2005.

[158] F. Brauer, Mathematical epidemiology: Past, present, and future, *Infect. Dis. Model.*, vol. 2, pp. 113–127, May 2017.

[159] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Epidemic processes in complex networks, *Rev. Mod. Phys.*, vol. 87, pp. 925–979, Aug 2015.

[160] W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale, *BMC Infect. Dis.*, vol. 11, pp. 1–14, Dec 2011.

[161] Q. Zhang, K. Sun, M. Chinazzi, A. Pastore y. Piontti, N. E. Dean, D. P. Rojas, S. Merler, D. Mistry, P. Poletti, L. Rossi, M. Bray, M. E. Halloran, I. M. Longini, and A. Vespignani, Spread of Zika virus in the Americas, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, May 2017.

[162] H. Heesterbeek, R. M. Anderson, V. Andreasen, S. Bansal, D. De Angelis, C. Dye, K. T. D. Eames, W. J. Edmunds, S. D. W. Frost, S. Funk, T. D. Hollingsworth, T. House, V. Isham, P. Klepac, J. Lessler, J. O. Lloyd-Smith, C. J. E. Metcalf, D. Mollison, L. Pellis, J. R. C. Pulliam, M. G. Roberts, and C. Viboud, Modeling infectious disease dynamics in the complex landscape of global health, *Science*, vol. 347, Mar 2015.

[163] S. Arregui, J. Sanz, D. Marinova, C. Martín, and Y. Moreno, On the impact of masking and blocking hypotheses for measuring the efficacy of new tuberculosis vaccines, *PeerJ*, vol. 4, p. e1513, Feb 2016.

[164] W. W. C. Topley and G. S. Wilson, The Spread of Bacterial Infection. The Problem of Herd-Immunity, *J. Hyg. (Lond).*, vol. 21, p. 243, May 1923.

[165] C. E. G. Smith, Prospects for the Control of Infectious Disease, *Proc. R. Soc. Med.*, vol. 63, p. 1181, Nov 1970.

[166] P. Fine, K. Eames, and D. L. Heymann, "Herd immunity": a rough guide, *Clin. Infect. Dis.*, vol. 52, pp. 911–916, Apr 2011.

[167] H. J. Larson, L. Z. Cooper, J. Eskola, S. L. Katz, and S. Ratzan, Addressing the vaccine confidence gap, *The. Lancet*, vol. 378, pp. 526–535, Aug 2011.

[168] Public Health England: measles in England, `https://publichealthmatters.blog.gov.uk/2019/08/19/measles-in-england/`. Accessed: 2019-08.

[169] CDC report on measles cases in 2019, `https://www.cdc.gov/measles/cases-outbreaks.html`. Accessed: 2019-08.

[170] Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani, Measurability of the epidemic reproduction number in data-driven contact networks, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, pp. 12680–12685, Dec 2018.

[171] X. Wang, A. Aleta, D. Lu, and Y. Moreno, Directionality reduces the impact of epidemics in multilayer networks, *New J. Phys.*, vol. 21, p. 093026, Sep 2019.

[172] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics, *Proc. R. Soc. Lond. A*, Aug 1927.

[173] O. Diekmann, H. Metz, and H. Heesterbeek, The legacy of Kermack and McKendrick, in *Epidemic Models: Their Structure and Relation to Data* (D. Mollison, ed.), pp. 95–115, Cambridge University Press, 1995.

[174] S. Arregui, M. J. Iglesias, S. Samper, D. Marinova, C. Martin, J. Sanz, and Y. Moreno, Data-driven model for the assessment of Mycobacterium tuberculosis transmission in evolving demographic structures, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, Apr 2018.

[175] J. D. Murray, *Mathematical Biology: I. An Introduction*. Springer, 2002.

[176] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Complex networks: Structure and dynamics, *Phys. Rep.*, vol. 424, pp. 175–308, Feb 2006.

[177] H. McCallum, N. Barlow, and J. Hone, How should pathogen transmission be modelled?, *Trends Ecol. Evol.*, vol. 16, pp. 295–300, Jun 2001.

[178] R. M. Anderson, Discussion: The Kermack-McKendrick epidemic threshold theorem, *Bull. Math. Biol.*, vol. 53, p. 1, Mar 1991.

[179] H. E. Soper, The Interpretation of Periodicity in Disease Prevalence, *J. Royal Stat. Soc.*, vol. 92, no. 1, p. 34, 1929.

[180] M. S. Bartlett, Measles Periodicity and Community Size, *J. Royal Stat. Soc. A*, vol. 120, no. 1, pp. 48–70, 1957.

[181] T. Britton, Stochastic epidemic models: A survey, *Math. Biosci.*, vol. 225, pp. 24–35, May 2010.

[182] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis*. Springer-Verlag New York, 2000.

[183] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*. Charles Griffin & Company LTD, 1975.

[184] N. T. J. Bailey, An Improbable Path, in *The Craft of Probabilistic Modelling*, pp. 63–87, Springer New York, 1986.

[185] B. D. Dalziel, O. N. Bjørnstad, W. G. van Panhuis, D. S. Burke, C. J. E. Metcalf, and B. T. Grenfell, Persistent Chaos of Measles Epidemics in the Prevaccination United States Caused by a Small Change in Seasonal Transmission Patterns, *PLoS Comput. Biol.*, vol. 12, Feb 2016.

[186] A. G. McKendrick, Applications of Mathematics to Medical Problems, *Proc. Edinburgh Math. Soc.*, vol. 44, pp. 98–130, Feb 1926.

[187] A. Aleta, A. N. S. Hisi, S. Meloni, C. Poletto, V. Colizza, and Y. Moreno, Human mobility networks and persistence of rapidly mutating pathogens, *R. Soc. Open Sci.*, Mar 2017.

[188] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2007.

[189] W. J. Edmunds, C. J. O'callaghan, and D. J. Nokes, Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections, *Proc. R. Soc. Lond. B*, vol. 264, pp. 949–957, Jul 1997.

[190] J. M. Read, J. Lessler, S. Riley, S. Wang, L. J. Tan, K. O. Kwok, Y. Guan, C. Q. Jiang, and D. A. T. Cummings, Social mixing patterns in rural and urban areas of southern China, *Proc. R. Soc. B*, Jun 2014.

[191] G. Béraud, S. Kazmercziak, P. Beutels, D. Levy-Bruhl, X. Lenne, N. Mielcarek, Y. Yazdanpanah, P.-Y. Boëlle, N. Hens, and B. Dervaux, The French Connection: The First Large Population-Based Contact Survey in France Relevant for the Spread of Infectious Diseases, *PLoS One*, vol. 10, Jul 2015.

[192] Y. Ibuka, Y. Ohkusa, T. Sugawara, G. B. Chapman, D. Yamin, K. E. Atkins, K. Taniguchi, N. Okabe, and A. P. Galvani, Social contacts, vaccination decisions and influenza in Japan, *J. Epidemiol. Community Health*, vol. 70, pp. 162–167, Feb 2016.

[193] M. C. Kiti, T. M. Kinyanjui, D. C. Koech, P. K. Munywoki, G. F. Medley, and D. J. Nokes, Quantifying Age-Related Rates of Social Contact Using Diaries in a Rural Coastal Population of Kenya, *PLoS One*, vol. 9, Aug 2014.

[194] M. Ajelli and M. Litvinova, Estimating contact patterns relevant to the spread of infectious diseases in Russia, *J. Theor. Biol.*, vol. 419, pp. 1–7, Apr 2017.

[195] O. l. P. de Waroux, S. Cohuet, D. Ndazima, A. J. Kucharski, A. Juan-Giner, S. Flasche, E. Tumwesigye, R. Arinaitwe, J. Mwanga-Amumpaire, Y. Boum, F. Nackers, F. Checchi, R. F. Grais, and W. J. Edmunds, Characteristics of human encounters and social mixing patterns relevant to infectious diseases spread by close contact: a survey in Southwest Uganda, *BMC Infect. Dis.*, vol. 18, pp. 1–12, Dec 2018.

[196] A. Melegaro, E. Del Fava, P. Poletti, S. Merler, C. Nyamukapa, J. Williams, S. Gregson, and P. Manfredi, Social Contact Structures and Time Use Patterns in the Manicaland Province of Zimbabwe, *PLoS One*, vol. 12, p. e0170459, Jan 2017.

[197] K. Leung, M. Jit, E. H. Y. Lau, and J. T. Wu, Social contact patterns relevant to the spread of respiratory infectious diseases in Hong Kong, *Sci. Rep.*, vol. 7, pp. 1–12, Aug 2017.

[198] M. A. Behr, P. H. Edelstein, and L. Ramakrishnan, Revisiting the timetable of tuberculosis, *BMJ*, vol. 362, 2018.

[199] UN Population Division Database, `https://population.un.org/wpp/`. Accessed: 2019-08.

[200] P. van den Driessche and J. Watmough, Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission, *Math. Biosci.*, vol. 180, pp. 29–48, Nov 2002.

[201] J. Heffernan, R. Smith, and L. Wahl, Perspectives on the basic reproductive ratio, *J. Royal Soc. Interface*, vol. 2, pp. 281–293, Jun 2005.

[202] M. E. Halloran and B. R. Levin, Infectious diseases of humans: dynamics and control, *Trends in Microbiology*, vol. 1, pp. 202–203, Aug 1993.

[203] L. I. Dublin and A. J. Lotka, On the True Rate of Natural Increase, *J. Am. Stat. Assoc.*, vol. 20, p. 305, Sep 1925.

[204] R. M. Anderson, Transmission Dynamics and Control of Infectious Disease Agents, in *Population Biology of Infectious Diseases* (R. M. Anderson and R. M. May, eds.), Springer-Verlag, 1982.

[205] J. A. P. Heesterbeek, A Brief History of R0 and a Recipe for its Calculation, *Acta Biotheor.*, vol. 50, pp. 189–204, Sep 2002.

[206] J. A. P. Heesterbeek and K. Dietz, The concept of R0 in epidemic theory, *Statistica Neerlandica*, vol. 50, pp. 89–110, Mar 1996.

[207] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, On the definition and the computation of the basic reproduction ratio r0 in models for infectious diseases in heterogeneous populations, *J. Math. Biol.*, vol. 28, pp. 365–382, Jun 1990.

[208] P. van den Driessche, Reproduction numbers of infectious disease models, *Infect. Dis. Model.*, vol. 2, p. 288, Aug 2017.

[209] J. Wallinga and M. Lipsitch, How generation intervals shape the relationship between growth rates and reproductive numbers, *Proc. R. Soc. B*, Nov 2006.

[210] K. Dietz, The Incidence of Infectious Diseases under the Influence of Seasonal Fluctuations, in *Mathematical Models in Medicine* (J. Berger, W. J. Bühler, R. Repges, and P. Tautu, eds.), pp. 1–15, Springer Berlin Heidelberg, 1976.

[211] C. Viboud, L. Simonsen, and G. Chowell, A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks, *Epidemics*, vol. 15, pp. 27–37, Jun 2016.

[212] H. Nishiura, Correcting the Actual Reproduction Number: A Simple Method to Estimate R0 from Early Epidemic Growth Data, *Int. J. Environ. Res. Public Health*, vol. 7, p. 291, Jan 2010.

[213] L. F. White and M. Pagano, A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic, *Stat. Med.*, vol. 27, p. 2999, Jul 2008.

[214] J. Ma, J. Dushoff, B. M. Bolker, and D. J. D. Earn, Estimating Initial Epidemic Growth Rates, *Bull. Math. Biol.*, vol. 76, pp. 245–260, Jan 2014.

[215] G. Chowell, L. Sattenspiel, S. Bansal, and C. Viboud, Mathematical models to characterize early epidemic growth: A review, *Phys. Life Rev.*, vol. 18, pp. 66–97, Sep 2016.

[216] A. Cintrón-Arias, C. Castillo-Chávez, L. M. A. Bettencourt, A. L. Lloyd, and H. T. Banks, The estimation of the effective reproductive number from disease outbreak data, *Math. Biosci. Eng.*, vol. 6, pp. 261–282, Apr 2009.

[217] H. Nishiura and G. Chowell, *The Effective Reproduction Number as a Prelude to Statistical Estimation of Time-Dependent Epidemic Trends*, pp. 103–121. Dordrecht: Springer Netherlands, 2009.

[218] G. Chowell, C. Viboud, L. Simonsen, and S. M. Moghadas, Characterizing the reproduction number of epidemics with early subexponential growth dynamics, *J. R. Soc. Interface*, vol. 13, Oct 2016.

[219] S. Paine, G. N. Mercer, P. M. Kelly, D. Bandaranayake, M. G. Baker, Q. S. Huang, G. Mackereth, A. Bissielo, K. Glass, and V. Hope, Transmissibility of 2009 pandemic influenza A(H1N1) in New Zealand: effective reproduction number and influence of age, ethnicity and importations, *Eurosurveillance*, vol. 15, p. 19591, Jun 2010.

[220] J. Wallinga and P. Teunis, Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures, *Am. J. Epidemiol.*, vol. 160, pp. 509–516, Sep 2004.

[221] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, Superspreading and the effect of individual variation on disease emergence, *Nature*, vol. 438, pp. 355–359, Nov 2005.

[222] I. M. Longini, Jr., J. S. Koopman, A. S. Monto, and J. P. Fox, Estimating Household and Community Transmission Parameters for Influenza, *Am. J. Epidemiol.*, vol. 115, pp. 736–751, May 1982.

[223] S. Cauchemez, A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, and D. Swerdlow, Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, pp. 2825–2830, Feb 2011.

[224] F. Ball, D. Mollison, and G. Scalia-Tomba, Epidemics with two levels of mixing, *Ann. Appl. Probab.*, vol. 7, pp. 46–89, Feb 1997.

[225] L. Pellis, N. M. Ferguson, and C. Fraser, Epidemic growth rate and household reproduction number in communities of households, schools and workplaces, *J. Math. Biol.*, vol. 63, pp. 691–734, Oct 2011.

[226] N. G. Becker and K. Dietz, The effect of household distribution on transmission and control of highly infectious diseases, *Math. Biosci.*, vol. 127, pp. 207–219, Jun 1995.

[227] T. House and M. J. Keeling, Deterministic epidemic models with explicit household structure, *Math. Biosci.*, vol. 213, pp. 29–39, May 2008.

[228] C. Fraser, Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic, *PLoS One*, vol. 2, Aug 2007.

[229] M. Ajelli, S. Merler, L. Fumanelli, A. P. y. Piontti, N. E. Dean, I. M. Longini, M. E. Halloran, and A. Vespignani, Spatiotemporal dynamics of the Ebola epidemic in Guinea and implications for vaccination and disease elimination: a computational modeling analysis, *BMC Med.*, vol. 14, pp. 1–10, Dec 2016.

[230] S. Merler, M. Ajelli, A. Pugliese, and N. M. Ferguson, Determinants of the Spatiotemporal Dynamics of the 2009 H1N1 Pandemic in Europe: Implications for Real-Time Modelling, *PLoS Comput. Biol.*, vol. 7, Sep 2011.

[231] S. Merler and M. Ajelli, The role of population heterogeneity and human mobility in the spread of pandemic influenza, *Proc. R. Soc. B*, Feb 2010.

[232] M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, and L. Finelli, Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature, *BMC Infect. Dis.*, vol. 14, pp. 1–20, Dec 2014.

[233] B. J. Cowling, V. J. Fang, S. Riley, J. S. M. Peiris, and G. M. Leung, Estimation of the serial interval of influenza, *Epidemiology*, vol. 20, p. 344, May 2009.

[234] W. G. Cochran, The Statistical Analysis of Field Counts of Diseased Plants, *Suppl. J. Royal Stat. Soc.*, vol. 3, no. 1, pp. 49–67, 1936.

[235] M. Morris, Epidemiology and Social Networks:: Modeling Structured Diffusion, *Sociol. Methods Res.*, vol. 22, pp. 99–126, Aug 1993.

[236] K. Dietz, Epidemics and Rumours: A Survey, *J. Royal Stat. Soc. A*, vol. 130, no. 4, pp. 505–528, 1967.

[237] D. Mollison, Spatial Contact Models for Ecological and Epidemic Spread, *J. Royal Stat. Soc. B*, vol. 39, no. 3, pp. 283–326, 1977.

[238] B. Von Bahr and A. Martin-Löf, Threshold Limit Theorems for Some Epidemic Processes, *Adv. Appl. Probab.*, vol. 12, pp. 319–349, Jun 1980.

[239] P. Grassberger, On the critical behavior of the general epidemic process and dynamical percolation, *Math. Biosci.*, vol. 63, pp. 157–172, Apr 1983.

[240] N. T. J. Bailey, Introduction to the modelling of venereal disease, *J. Math. Biology*, vol. 8, pp. 301–322, Oct 1979.

[241] K. Dietz, Models for Vector-Borne Parasitic Diseases, in *Vito Volterra Symposium on Mathematical Models in Biology* (C. Barigozzi, ed.), pp. 264–277, Springer, Berlin, Heidelberg, 1980.

[242] H. W. Hethcote and J. A. Yorke, *Gonorrhea Transmission Dynamics and Control*. Springer, Berlin, Heidelberg, 1984.

[243] R. M. Anderson, G. F. Medley, R. M. May, and A. M. Johnson, A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS, *H. Math. Appl. Med. Biol.*, vol. 3, no. 4, pp. 229–63, 1986.

[244] R. M. May and R. M. Anderson, The transmission dynamics of human immunodeficiency virus (HIV), *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 321, pp. 565–607, Oct 1988.

[245] G. F. Bolz, Simulation on random graphs of the epidemic dynamics of sexually transmitted diseases - a new model for the epidemiology of AIDS, in *Stochastics, Algebra and Analysis in Classical and Quantum Dynamics. Proceedings of the IVth French-German Encounter on Mathematics and Physics.* (S. Albeverio, P. Blanchard, and D. Testard, eds.), 1988.

[246] A. S. Klovdahl, Social networks and the spread of infectious diseases: The AIDS example, *Soc. Sci. Med.*, vol. 21, pp. 1203–1216, Jan 1985.

[247] A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow, Social networks and infectious disease: The Colorado Springs study, *Soc. Sci. Med.*, vol. 38, pp. 79–88, Jan 1994.

[248] H. Andersson, Epidemics in a population with social structures, *Math. Biosci.*, vol. 140, pp. 79–84, Mar 1997.

[249] F. Cohen, Computer viruses: Theory and experiments, *Computers & Security*, vol. 6, pp. 22–35, Feb 1987.

[250] W. H. Murray, The application of epidemiology to computer viruses, *Computers & Security*, vol. 7, pp. 139–145, Apr 1988.

[251] J. O. Kephart and S. R. White, Directed-graph epidemiological models of computer viruses, in *Proceedings. 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, IEEE, May 1991.

[252] J. O. Kephart, G. B. Sorkin, D. M. Chess, and S. R. White, Fighting Computer Viruses, *Sci. Am.*, vol. 277, pp. 88–93, Nov 1997.

[253] R. Albert, H. Jeong, and A.-L. Barabási, Diameter of the World-Wide Web, *Nature*, vol. 401, pp. 130–131, Sep 1999.

[254] R. Pastor-Satorras and A. Vespignani, Epidemic Spreading in Scale-Free Networks, *Phys. Rev. Lett.*, vol. 86, pp. 3200–3203, Apr 2001.

[255] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg, The web of human sexual contacts, *Nature*, vol. 411, pp. 907–908, Jun 2001.

[256] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Critical phenomena in complex networks, *Rev. Mod. Phys.*, vol. 80, pp. 1275–1335, Oct 2008.

[257] M. E. J. Newman, Spread of epidemic disease on networks, *Phys. Rev. E*, vol. 66, p. 016128, Jul 2002.

[258] L. A. Meyers, M. E. J. Newman, and B. Pourbohloul, Predicting epidemics on directed contact networks, *J. Theor. Biol.*, vol. 240, pp. 400–418, Jun 2006.

[259] E. Cozzo, R. A. Baños, S. Meloni, and Y. Moreno, Contact-based social contagion in multiplex networks, *Phys. Rev. E*, vol. 88, p. 050801, Nov 2013.

[260] J. Sanz, C.-Y. Xia, S. Meloni, and Y. Moreno, Dynamics of Interacting Diseases, *Phys. Rev. X*, vol. 4, p. 041005, Oct 2014.

[261] C. Granell, S. Gómez, and A. Arenas, Dynamical Interplay between Awareness and Epidemic Spreading in Multiplex Networks, *Phys. Rev. Lett.*, vol. 111, p. 128701, Sep 2013.

[262] J. A. Drewe, Who infects whom? Social networks and tuberculosis transmission in wild meerkats, *Proc. R. Soc. B*, Feb 2010.

[263] J. L. Kool and R. A. Weinstein, Risk of Person-to-Person Transmission of Pneumonic Plague, *Clin. Infect. Dis.*, vol. 40, pp. 1166–1172, Apr 2005.

[264] V. P. Martinez, C. Bellomo, J. S. Juan, D. Pinna, R. Forlenza, M. Elder, and P. J. Padula, Person-to-Person Transmission of Andes Virus, *Emerg. Infect. Dis.*, vol. 11, p. 1848, Dec 2005.

[265] G. F. de Arruda, E. Cozzo, T. P. Peixoto, F. A. Rodrigues, and Y. Moreno, Disease Localization in Multilayer Networks, *Phys. Rev. X*, vol. 7, p. 011014, Feb 2017.

[266] L. Mercken, T. A. B. Snijders, C. Steglich, E. Vartiainen, and H. de Vries, Dynamics of adolescent friendship networks and smoking behavior, *Soc. Netw.*, vol. 32, pp. 72–81, Jan 2010.

[267] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini, The Role of Information Diffusion in the Evolution of Social Networks, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), pp. 356–364, ACM, 2013.

[268] M. Magnani and L. Rossi, The ML-Model for Multi-layer Social Networks, in *2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 5–12, Jul 2011.

[269] C. Machado, B. Kira, V. Narayanan, B. Kollanyi, and P. Howard, A Study of Misinformation in WhatsApp Groups with a Focus on the Brazilian Presidential Elections, in *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, (New York, NY, USA), pp. 1013–1019, ACM, 2019.

[270] M. Buchanan, *The Social Atom*. Bloomsbury, 2007.

[271] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, *Rev. Mod. Phys.*, vol. 81, pp. 591–646, May 2009.

[272] R. H. Knapp, A Psychology of Rumor, *Public Opin. Q.*, vol. 8, no. 1, p. 22, 1944.

[273] D. J. Daley and D. G. Kendall, Stochastic Rumours, *IMA J. Appl. Math.*, vol. 1, pp. 42–55, Mar 1965.

[274] S. Vosoughi, D. Roy, and S. Aral, The spread of true and false news online, *Science*, vol. 359, pp. 1146–1151, Mar 2018.

[275] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, The spread of low-credibility content by social bots, *Nat. Commun.*, vol. 9, pp. 1–9, Nov 2018.

[276] A. Aleta, J. O'Brien, J. Gleeson, and Y. Moreno, Dynamics of discussion threads in online boards: the case of Forocoches, *In preparation*, 2019.

[277] A. Aleta and Y. Moreno, The dynamics of collective social behavior in a crowd controlled game, *EPJ Data Sci.*, vol. 8, pp. 1–16, Jun 2019.

[278] Wikipedia entry on Internet Forums, https://en.wikipedia.org/wiki/Internet_forum. Accessed: 2019-08.

[279] Forum Software Timeline, https://www.forum-software.org/forum-software-timeline-from-1994-to-today. Accessed: 2019-08.

[280] H. Rheingold, *The virtual community*. Addison-Wesley, 1993.

[281] F. S. L. Lee, D. Vogel, and M. Limayem, Virtual Community Informatics: A Review and Research Agenda, *JITTA*, vol. 5, no. 1, p. 5, 2003.

[282] C. R. Sunstein, The Law of Group Polarization, *J. Political Philos.*, vol. 10, no. 2, pp. 175–195, 1999.

[283] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, Political Discourse on Social Media, in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, ACM Press, 2018.

[284] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and D. Boyd, The Arab Spring| The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions, *Int. J. Commun.*, vol. 5, p. 31, Sep 2011.

[285] Incels: a definition and investigation into a dark internet corner, https://www.vox.com/the-highlight/2019/4/16/18287446/incel-definition-reddit. Accessed: 2019-08.

[286] J. A. Berger and C. Heath, Idea Habitats: How the Prevalence of Environmental Cues Influences the Success of Ideas, *Cogn. Sci*, vol. 29, pp. 195–221, Mar 2005.

[287] D. Sperber, Anthropology and Psychology: Towards an Epidemiology of Representations, *Man*, vol. 20, pp. 73–89, Mar 1985.

[288] R. Axelrod, The Dissemination of Culture: A Model with Local Convergence and Global Polarization, *J. Confl. Resolut*, vol. 41, pp. 203–226, Apr 1997.

[289] J.-P. Onnela and F. Reed-Tsochas, Spontaneous emergence of social influence in online systems, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, pp. 18375–18380, Oct 2010.

[290] C. Dellarocas, Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms, *Manag. Sci*, vol. 52, pp. 1577–1593, Oct 2006.

[291] Facebook already speaks Spanish, https://elpais.com/tecnologia/2008/02/11/actualidad/1202722079_850215.html. Accessed: 2019-08, in Spanish.

[292] Twitter already speaks Spanish, https://elpais.com/tecnologia/2009/11/04/actualidad/1257328861_850215.html. Accessed: 2019-08, in Spanish.

[293] J. M. Miotto and E. G. Altmann, Predictability of Extreme Events in Social Media, *PLoS One*, vol. 9, p. e111506, Nov 2014.

[294] P. Sobkowicz, M. Thelwall, K. Buckley, G. Paltoglou, and A. Sobkowicz, Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?, *EPJ Data Sci.*, vol. 2, pp. 1–20, Dec 2013.

[295] Statistics of Forocoches, https://www.forocoches.com/index.php?p=media. Accessed: 2019-08, in Spanish.

[296] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, The Twitter of Babel: Mapping World Languages through Microblogging Platforms, *PLoS One*, vol. 8, Apr 2013.

[297] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, in *Seventh International AAAI Conference on Weblogs and Social Media*, Jun 2013. [Online; accessed 8. Sep. 2019].

[298] V. Kulkarni and W. Y. Wang, TFW, DamnGina, Juvie, and Hotsie-Totsie: On the Linguistic and Social Aspects of Internet Slang, *arXiv*, Dec 2017.

[299] F. Kooti, H. Yang, M. Cha, K. Gummadi, and W. Mason, The Emergence of Conventions in Online Social Networks, in *International AAAI Conference on Web and Social Media*, 2012.

[300] R. Amato, L. Lacasa, A. Díaz-Guilera, and A. Baronchelli, The dynamics of norm change in the cultural evolution of language, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, pp. 8260–8265, Aug 2018.

[301] Chitikia Insights: The Value of Google Result Positioning, http://info.chitika.com/uploads/4/9/2/1/49215843/chitikainsights-valueofgoogleresultspositioning.pdf, 2013. Accessed: 2019-08.

[302] M.-A. Rizoiu, Y. Lee, and S. Mishra, Hawkes processes for events in social media, in *Frontiers of Multimedia Research*, pp. 191–218, ACM, Dec 2017.

[303] A. N. Medvedev, J.-C. Delvenne, and R. Lambiotte, Modelling structure and predicting dynamics of discussion threads in online boards, *J. Complex Networks*, vol. 7, pp. 67–82, May 2018.

[304] A. Reinhart, A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications, *Stat. Sci.*, vol. 33, pp. 299–318, Aug 2018.

[305] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes - Volume I: Elementary Theory and Methods*. Springer-Verlag New York, 2003.

[306] A. G. Hawkes, Spectra of Some Self-Exciting and Mutually Exciting Point Processes, *Biometrika*, vol. 58, pp. 83–90, Apr 1971.

[307] P. J. Laub, T. Taimre, and P. K. Pollett, Hawkes Processes, 2015.

[308] V. Filimonov and D. Sornette, Quantifying reflexivity in financial markets: Toward a prediction of flash crashes, *Phys. Rev. E*, vol. 85, May 2012.

[309] M.-A. Rizoiu, S. Mishra, Q. Kong, M. Carman, and L. Xie, SIR-Hawkes, in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, ACM Press, 2018.

[310] Y. Ogata, Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes, *J. Am. Stat. Assoc.*, vol. 83, pp. 9–27, Mar 1988.

[311] Deep Dive - Y. Ogata's Residual Analysis for Point Processes, `https://simplystatistics.org/2017/09/04/deep-dive-ogata/`. Accessed: 2019-08.

[312] T. Ozaki, Maximum likelihood estimation of Hawkes' self-exciting point processes, *Ann. Inst. Stat. Math.*, vol. 31, pp. 145–155, Dec 1979.

[313] M. Lallouache and D. Challet, The limits of statistical significance of Hawkes processes fitted to financial data, *Quant. Finance*, vol. 16, pp. 1–11, Jan 2016.

[314] J. E. Cavanaugh and A. A. Neath, The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements, *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 11, Mar 2019.

[315] F. Papangelou, Integrability of expected increments of point processes and a related random change of scale, *Trans. Amer. Math. Soc.*, vol. 165, pp. 483–506, 1972.

[316] D. Abrams and M. A. Hogg, Collective Identity: Group Membership and Self-Conception, in *Blackwell Handbook of Social Psychology: Group Processes*, pp. 425–460, Blackwell Publishers Ltd, 2008.

[317] G. Le Bon, *The Crowd: A Study of the Popular Mind*. 1895.

[318] S. Reicher, The Psychology of Crowd Dynamics, in *Blackwell Handbook of Social Psychology: Group Processes*, pp. 182–208, Blackwell Publishers Ltd, Jan 2008.

[319] S. T. La Macchia and W. R. Louis, Crowd behaviour and collective action, in *Understanding peace and conflict through social identity theory: Contemporary global perspectives*, Springer International Publishing, 2016.

[320] S. Reicher, C. Stott, P. Cronin, and O. Adang, An integrated approach to crowd psychology and public order policing, *Policing*, vol. 27, no. 4, pp. 558–572, 2004.

[321] R. V. Kozinets, A. Hemetsberger, and H. J. Schau, The wisdom of consumer crowds: Collective innovation in the age of networked marketing, *J. Macromarketing*, vol. 28, no. 4, pp. 339–354, 2008.

[322] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, reCAPTCHA: Human-Based Character Recognition via Web Security Measures, *Science*, vol. 321, pp. 1465–1468, Sep 2008.

[323] R. F. Baumeister, S. E. Ainsworth, and K. D. Vohs, Are groups more or less than the sum of their members? The moderating role of individual identification, *Behav. Brain Sci.*, vol. 39, p. e137, 2016.

[324] B. Latané, The psychology of social impact., *Am. Psychol.*, vol. 36, no. 4, p. 343, 1981.

[325] A. J. Quinn and B. B. Bederson, Human Computation: A Survey and Taxonomy of a Growing Field, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, (New York, NY, USA), pp. 1403–1412, ACM, 2011.

[326] T. W. Malone, R. Laubacher, and C. Dellarocas, The collective intelligence genome, *IEEE Eng. Manage. Rev.*, vol. 38, pp. 38–52, Aug 2010.

[327] W. Mason and D. J. Watts, Financial incentives and the "performance of crowds", in *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '09*, ACM Press, 2009.

[328] C. Prendergast, The Provision of Incentives in Firms, *J. Econ. Lit.*, vol. 37, pp. 7–63, Mar 1999.

[329] J. Heyman and D. Ariely, Effort for payment: A tale of two markets, *Psychol. Sci.*, vol. 15, no. 11, pp. 787–793, 2004.

[330] U. Gneezy and A. Rustichini, Pay enough or don't pay at all, *Q. J. Econ*, vol. 115, no. 3, pp. 791–810, 2000.

[331] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, The future of crowd work, in *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, ACM Press, 2013.

[332] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, Beyond the Turk: Alternative platforms for crowdsourcing behavioral research, *J. Exp. Soc. Psychol*, vol. 70, pp. 153–163, May 2017.

[333] J. Cox, E. Y. Oh, B. Simmons, C. Lintott, K. Masters, A. Greenhill, G. Graham, and K. Holmes, Defining and Measuring Success in Online Citizen Science: A Case Study of Zooniverse Projects, *Comput. Sci. Eng.*, vol. 17, pp. 28–41, Jul 2015.

[334] L. von Ahn, Games with a Purpose, *Computer*, vol. 39, pp. 92–94, Jun 2006.

[335] F. Khatib, S. Cooper, M. D. Tyka, K. Xu, I. Makedon, Z. Popović, D. Baker, and F. Players, Algorithm discovery by protein folding game players, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, pp. 18949–18953, Nov 2011.

[336] M. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz, Collabio: A Game for Annotating People within Social Networks, in *UIST09*, Oct 2009.

[337] C. F. Salk, T. Sturn, L. See, S. Fritz, and C. Perger, Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game, *Int. J. Digital Earth*, vol. 9, pp. 410–426, Apr 2016.

[338] A. Birke, H. Schoenau-Fog, and L. Reng, Space Bugz!: A Smartphone-controlled Crowd Game, in *Proceeding of the 16th International Academic MindTrek Conference*, MindTrek '12, (New York, NY, USA), pp. 217–219, ACM, 2012.

[339] Description of The Button event, `https://perma.cc/HLV8-NTBL`. Accessed: 2019-08.

[340] T. F. Müller and J. Winters, Compression in cultural evolution: Homogeneity and structure in the emergence and evolution of a large-scale online collaborative art project, *PLoS One*, vol. 13, Sep 2018.

[341] J. Rappaz, M. Catasta, R. West, and K. Aberer, Latent Structure in Collaboration: The Case of Reddit r/place, *arXiv*, 2018.

[342] *Guinness World Records 2015 Gamer's Edition*. Guinness Book, Nov 2014.

[343] Chat logs and videos of the whole event, `https://archive.org/`. Accessed: 2019-08.

[344] Pokémon passes 300 million games sold, https://perma.cc/DC6H-GTMV. Accessed: 2019-08.

[345] T. Althoff, R. W. White, and E. Horvitz, Influence of Pokémon Go on Physical Activity: Study and Implications, *J. Med. Internet Res.*, vol. 18, p. e315, Dec 2016.

[346] M. Sjöblom and J. Hamari, Why do people watch others play video games? An empirical study on the motivations of Twitch users, *Comput. Hum. Behav*, vol. 75, pp. 985–996, Oct 2017.

[347] Twitch is 4th in Peak US Internet Traffic, https://perma.cc/5V9Y-YYPG. Accessed: 2019-08.

[348] B. C. B. Churchill and W. Xu, The Modem Nation: A First Study on Twitch.TV Social Structure and Player/Game Relationships, in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, pp. 223–228, IEEE, Oct 2016.

[349] M. Mallory, Community-based Play in Twitch Plays Pokémon, in *Well Played* (E. Flynn-Jones, ed.), vol. 3, Pittsburgh, PA: ETC Press, 2015.

[350] Twitch Plays Pokémon timeline, https://perma.cc/5V9Y-YYPG. Accessed: 2019-08.

[351] Time needed to finish Pokémon Red, https://perma.cc/6Y69-76KH. Accessed: 2019-08.

[352] M.-V. Lindsey, The Politics of Pokémon. Socialized Gaming, Religious Themes and the Construction of Communal Narratives, *Heidelberg Journal of Religions on the Internet*, 2015.

[353] D. Centola and A. Baronchelli, The spontaneous emergence of conventions: An experimental study of cultural evolution, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, pp. 1989–1994, Feb 2015.

[354] F. Galton, Vox populi (The wisdom of crowds), *Nature*, vol. 75, no. 7, pp. 450–451, 1907.

[355] J. Surowiecki, *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*, vol. 296. United States: Anchor Books, 2004.

[356] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin, Effective leadership and decision-making in animal groups on the move, *Nature*, vol. 433, no. 7025, p. 513, 2005.

[357] J. R. Dyer, C. C. Ioannou, L. J. Morrell, D. P. Croft, I. D. Couzin, D. A. Waters, and J. Krause, Consensus decision making in human crowds, *Animal Behav.*, vol. 75, no. 2, pp. 461–470, 2008.

[358] L. Muchnik, S. Aral, and S. J. Taylor, Social Influence Bias: A Randomized Experiment, *Science*, vol. 341, pp. 647–651, Aug 2013.

[359] L. B. Rosenberg, Human swarming, a real-time method for parallel distributed intelligence, in *2015 Swarm/Human Blended Intelligence Workshop (SHBI)*, IEEE, Sep 2015.

[360] L. Rosenberg, D. Baltaxe, and N. Pescetelli, Crowds vs swarms, a comparison of intelligence, in *2016 Swarm/Human Blended Intelligence Workshop (SHBI)*, IEEE, Oct 2016.

[361] R. L. M. Lee, Do online crowds really exist? Proximity, connectivity and collectivity, *Distinktion*, vol. 18, pp. 82–94, Jan 2017.

[362] B. Kirman, C. Lineham, and S. Lawson, Exploring mischief and mayhem in social computing or: how we learned to stop worrying and love the trolls, in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pp. 121–130, ACM, 2012.

[363] H. L. Paul, N. D. Bowman, and J. Banks, The enjoyment of griefing in online games, *J. Gam. Virt. W.*, vol. 7, no. 3, pp. 243–258, 2015.

[364] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, Trolls just want to have fun, *Pers. Individ. Differ*, vol. 67, pp. 97–102, 2014.

[365] D. G. Rand, A. Peysakhovich, G. T. Kraft-Todd, G. E. Newman, O. Wurzbacher, M. A. Nowak, and J. D. Greene, Social heuristics shape intuitive cooperation, *Nat. Commun.*, vol. 5, p. 3677, 2014.

[366] T. Yamagishi, Y. Matsumoto, T. Kiyonari, H. Takagishi, Y. Li, R. Kanai, and M. Sakagami, Response time in economic games reflects different types of decision conflict for prosocial and proself individuals, *Proc. Natl. Acad. Sci. U.S.A.*, 2017.

[367] T. Kameda, T. Tsukasaki, R. Hastie, and N. Berg, Democracy under uncertainty: the wisdom of crowds and the free-rider problem in group decision making, *Psychol. Rev.*, vol. 118, pp. 76–96, Jan 2011.

[368] A. A. Hung and C. R. Plott, Information Cascades: Replication and an Extension to Majority Rule and Conformity-Rewarding Institutions, *Am. Econ. Rev*, vol. 91, pp. 1508–1520, Dec 2001.

[369] A strategy to traverse the ledge, `https://perma.cc/FFK8-9CHG`. Accessed: 2019-08.

[370] R. W. White, Motivation Reconsidered: The Concept of Competence, *Psychol. Rev.*, vol. 66, no. 5, pp. 297–333, 1959.

[371] C. Klimmt, T. Hartmann, and A. Frey, Effectance and Control as Determinants of Video Game Enjoyment, *Cyb. Psy. Behav.*, vol. 10, Dec 2007.

[372] E. D. Mekler, J. A. Bopp, A. N. Tuch, and K. Opwis, A systematic review of quantitative studies on the enjoyment of digital entertainment games, in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 927–936, ACM, 2014.

[373] R. Bartle, Hearts, clubs, diamonds, spades: Players who suit MUDs, Jun 1996.

[374] W. van den Hoogen, K. Poels, W. IJsselsteijn, and Y. de Kort, Between Challenge and Defeat: Repeated Player-Death and Game Enjoyment, *Media Psychol.*, vol. 15, no. 4, pp. 443–459, 2012.

[375] R. Spears and T. Postmes, *Group Identity, Social Influence and Collective Action Online: Extensions and Applications of the SIDE Model*, pp. 23–46. Handbooks in communication and media, Chichester,UK: Wiley-Blackwell, 2015.

[376] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, Political polarization on twitter, in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, vol. 133, pp. 89–96, 2011.

[377] T. J. Gray, C. M. Danforth, and P. S. Dodds, Hahahahaha, Duuuuude, Yeeessss!: A two-parameter characterization of stretchable words and the dynamics of mistypings and misspellings, *arXiv*, 2019.

[378] A. Nylund and O. Landfors, Frustration and its effect on immersion in games: A developer viewpoint on the good and bad aspects of frustration, Master's thesis, Umeå University, 2015.

[379] The rise of masocore gaming, https://perma.cc/HKD3-KP2Q. Accessed: 2019-08.

[380] Games with Busted Physics and Controls, https://perma.cc/Z33L-NUSW. Accessed: 2019-08.

A
Resumen en español

En el último cuarto del siglo XX se cuestionó la visión reduccionista que había predominado en el desarrollo de la física hasta entonces. En su lugar, se propuso que los sistemas se organizan en jerarquías de forma que el nivel superior tiene que seguir las reglas del nivel inferior y, al mismo tiempo, puede exhibir sus propias leyes, las cuales no pueden inferirse a partir de las de sus componentes fundamentales. Esta observación llevó a la creación de un nuevo campo conocido como sistemas complejos. No obstante, esta nueva visión no estaba restringida a sistemas puramente físicos. Se observó que sistemas muy diferentes provenientes de una gran cantidad de campos distintos, desde ecología a sociología o economía, podían ser analizados como sistemas complejos. Es más, permitió a los físicos contribuir con sus conocimientos y herramientas al desarrollo de la investigación en dichas áreas.

En esta tesis abordamos problemas de tres áreas de los sistemas complejos: redes, que son una de las principales herramientas matemáticas utilizadas para estudiar sistemas complejos; difusión de epidemias, que ha sido uno de los campos en los que la aplicación de una perspectiva de sistemas complejos ha resultado más exitosa; y el estudio del comportamiento colectivo, el cual ha atraído una gran atención en los últimos años dado la gran cantidad de datos que están ahora disponibles gracias a las redes sociales. De hecho, los datos serán también el hilo conductor de la discusión sobre las otras dos áreas previamente mencionadas. En particular, usamos nuevas fuentes de información para desafiar algunas de las asunciones clásicas que se han hecho tanto en el estudio de las redes como en el desarollo de los modelos de difusión de epidemias.

En el caso de las redes, estudiamos el problema de los modelos nulos utilizando herramientas provenientes de la física estadística. Gracias a ello, demostramos que anomalías encontradas en algunas redes pueden ser simplemente una consecuencia de una excesiva simplificación de los modelos utilizados para estudiarlas. A continuación, extendemos este marco para generar redes de contacto para el estudio de la propagación de epidemias en poblaciones en las que tanto la estructura de contactos como la distribución de edad de la población son importantes.

Posteriormente, seguimos el desarrollo histórico de la epidemiología matemática y repasamos las asunciones que se hicieron cuando no había suficientes datos sobre el verdadero comportamiento de este tipo de sistemas. Mostramos que una de las cantidades más utilizadas en este tipo de estudios, el *basic reproduction number*, no está adecuadametne definido en sistemas reales. De forma similar, extendemos el marco teórico utilizado para estudiar la propagación de epidemias en redes dirigidas a sistemas multicapa. Además, mostramos que el reto de incorporar datos a los modelos no está restringido únicamente al problema de obtenerlos, sino que también es muy importante ser conscientes de sus características para hacerlo adecuadamente.

Finalmente, concluímos la tesis estudiando dos ejemplos de comportamiento colectivo utilizando datos extraídos de sistemas *online*. Para ello, utilizamos técnicas que fueron desarrolladas originalmente para estudiar otro tipo de sistemas, como la predicción de terremotos. No obstante, demostramos que también pueden ser utilizados para estudiar este nuevo tipo de sistemas. Es más, demostramos que a pesar de sus características únicas,

poseen propiedades similares a las que han sido observadas en el mundo *offline*. Esto no solo implica que las sociedades modernas están entrelazadas con el mundo *online*, sino que señala que, si queremos entender los sistemas socio-técnicos, una visión holística, como la propuesta por los sitemas complejos, es indispensable.

B
Conclusiones en español

Comenzamos esta tesis mostrando que normalmente la ciencia avanza en pequeños pasos, en lugar de en grandes saltos. Esta observación es todavía más acertada en el caso de los sistemas complejos, donde una teoría unificada actualmente no existe (si es que lo hace alguna vez) y todo lo que tenemos es una colección de historias cortas.

En el capítulo 2 nos centramos en estudiar una de las herramientas más importantes utilizadas para el estudio de los sistemas complejos, las redes. En particular, abordamos el problema de cómo crear modelos nulos adecuados para redes en función de la cantidad de datos disponible. Recordemos que la propuesta más simple era utilizar redes aleatorias, como algo con lo que comparar las redes reales. Sin embargo, este procedimiento tiene dos grandes desventajas. Primero, es posible que la estructura microscópica de la red sea tal que resulte en estructuras de orden superior no presentes en las redes aleatorias. Aunque esto claramente es información útil, puede llevarnos a pensar que el sistema en consideración evolucionó específicamente para crear dichas estructuras. Sin embargo, es posible que sean simplemente una consecuencia directa de las propiedades de orden inferior, las cuales deverían ser entonces el objetivo de la investigación.

El otro principal problema es que, dado que las redes puedes ser utilizadas en una gran variedad de sistemas de características muy diversas, comparar una red real con una red completamente aleatoria puede llevarnos a pensar que es más rara de lo que realmente es. Por ejemplo, se ha observado que en las redes de amistad el número de triángulos (i.e., si A y B son amigos y B y C también, entonces A y C son amigos también) es mucho mayor que en redes aleatorias. Esto es ciertamente una propiedad importante de estas redes. No obstante, dada una red de amistad, es posible que no estemos interesados en medir si el número de triángulos es mayor el esperado al azar porque ya sabemos que seguramente lo será. En su lugar, puede resultar más adecuado comprobar si dicho número es superior al de otras redes de amistad, o en comparación con modelos nulos que reflejen las características comunes de estas redes. Este ejemplo puede ser extendido a una gran cantidad de sistemas. En consecuencia, hay prácticamente un modelo nulo para cada aplicación en la que podamos pensar.

Cuando estudiamos las anomalías presentes en la *betweenness* de redes de transporte, podríamos haber utilizado un modelo nulo que específicamente tuviese en cuenta las características de estos sistemas, como la población que reside en cada municipalidad, su tamaño, etc. En su lugar, elegimos seguir el marco inspirado por Jaynes según el cual la física estadística puede ser entendida como un problema de información, lo que resulta en el modelo de redes aleatorias exponenciales. Este marco general nos permitió determinar que las anomalías observadas eran simplemente una consecuencia de la distribución de pesos. Por tanto, en lugar de preguntarnos por qué estas redes muestran estas anomalías, el objetivo debería ser estudiar los mecanismos que llevan a dichas distribuciones de pesos. Es más, este estudio también demostró la importancia de poseer la cantidad adecuada de datos.

Concluímos el capítulo aplicando el formalismo para crear redes de contacto multicapa utilizando tanto datos sorbe la distribución de contactos como sobre los patrones de interacción por edad. Mostramos que dicha información puede ser extraída de bases de

datos reales e introducida en el modelo para generar redes de contacto realistas. Cabe destacar que en este caso no queríamos crear un modelo nulo para realizar comparaciones, sino para construir redes de forma no sesgada dados los datos disponibles. Por consiguiente, otra ventaja del modelo de redes aleatorias exponenciales es que unifica varios problemas bajo una única formulación.

En el capítulo 3, nos centramos en el estudio matemático de la propagación de epidemias. Seguimos el desarrollo histórico del campo, desde la aproximación más simple hasta los modelos más detallados. Es más, en cada paso, comprobamos el efecto de incluir más datos en los modelos. En el primer caso, mostramos que el reto de incorporar datos en los modelos no se reduce simplemente a conseguir la información, sino que también es muy importante ser conscientes de sus características. En particular, vimos que las matrices de contactos por edad no pueden aplicarse directamente a cualquier población, ya que poseen información implícita sobre la misma. Por tanto, si la población cambia, las matrices también deberán cambiar.

A continuación, creamos un modelo númerico altamente realista para estudiar la propagación de epidemias similares a la gripe y mostramos que las asunciones teóricas comúnmente aceptadas tal vez no sean lo suficientemente buenas como para capturar la complejidad del proceso. En particular, observamos que la definición de una de las cantidades más importantes en epidemiología, el *basic reproduction number*, no captura adecuadamente la verdadera dinámica del sistema. La razón era que la que la definición matemática se sostenía sobre una serie de asunciones que son contradichas por los datos.

Por otra parte, el tercer estudio estaba centrado en extender los modelos teóricos de propagación de epidemias en redes multicapa al caso en el que la dirección de los enlaces es conocida. Mostramos que las poblaciones en las que la red de contactos posee alguna direccionalidad son más resistentes de cara a una epidemia que aquellas que son completamente no dirigidas. Ciertamente, gracias a las plataformas sociales *online*, resulta más sencillo conseguir este tipo de datos para sistemas sociales. En cualquier caso, la formulación básica puede ser adaptada fácilmente para analizar la propagación de información. Por tanto, nuestros resultados también implican que el papel que pueden jugar las plataformas en las que la comunicación es no dirigida es muy diferente del que pueden jugar las que son dirigidas.

Finalmente, analizamos las consecuencias de usar diferentes modelos de epidemias en función de la disponibilidad de datos, con particular énfasis en las redes que fueron creadas al final del capítulo 2. Mostramos que cuantos más datos, mejor, pero que para algunas tareas los modelos más simples con menos datos también pueden aportar información relevante. En concreto, aunque conocimiento sobre la red de contactos subyacente es crucial para determinar el *epidemic threshold*, tener información sobre la estructura de edad de la población es crucial para definir correctamente los grupos de riesgo.

Para concluir, en el capítulo 4, analizamos dos ejemplos de comportamiento colectivo utilizando datos extraídos de fuentes muy diferentes. Primero, mostramos que los procesos de Hawkes pueden ser utilizados de forma efectiva para el análisis de foros de discusión *online* como Forocoches. Además, fuimos capaces de distinguir dos tipos muy diferentes de actividad, uno que era independiente de la actividad del resto de usuarios y otra en la que el componente social del proceso es indispensable. Terminamos el capítulo estudiando un evento de masas *online*, Twitch Plays Pokémon. Observamos que a pesar de sus singulares características, algunas de las propiedades de los grupos *online* son similares a las de los grupos *offline* mostrando, una vez más, que las sociedades modernas están interlazadas con el mundo *online*.

En resumen, hemos revisado una pequeña fracción de los sistemas complejos, con especial énfasis en el rol que las nuevas fuentes de datos pueden tener en problemas que van desde los trabajos más teóricos a simulaciones computacionales altamente detalladas.

Confiamos en que esta colección de historias cortas mostrará la gran diversidad de problemas que siguen abiertos en el campo de los sistemas complejos y, al mismo tiempo, arrojará algo de luz sobre ellos.

B.1 Perspectivas

Existen múltiples formas en las que los resultados de esta tesis pueden ser extendidos, algunas de ellas para incrementar nuestro conocimiento sobre sistemas particulares y otras para aumentar nuestro entendimiento global de los sistemas complejos. Actualmente algunas ideas ya han comenzado a desarrollarse, mientras que otras son simplemente proyectos.

En el capítulo 2 nos centramos en estudiar el modelo de redes aleatorias exponenciales. A pesar de sus grandes ventajas, es importante tener en cuenta que también posee algunos inconvenientes. Por ejemplo, los recursos computacionales necesarios para obtener numéricamente los parámetros del modelo pueden ser muy elevados, en función del tamaño y las características de las redes. Además, actualmente existen numersos investigadores, provenientes de muy diversas áreas, que quieren utilizar redes pero que carecen del conocimiento técnico necesario para poder entender este modelo. Por ello, uno de nuestros próximos proyectos será hacer una gran revisión de los modelos nulos que existen en la literatura, estudiando sistemáticamente sus ventajas y desventajas. El objetivo de este trabajo será servir de referencia a aquellos investigadores trabajando en sistemas complejos que no estén acostumbrados a estudiar redes y que, por tanto, desconocen los inconvenientes que las técnicas más simples poseen.

En cuanto al capítulo 3, cabe destacar que el hilo conductor de los cuatro estudios eran los datos: (1) cómo manejar los datos; (2) teoría frente a simulaciones basadas en datos; (3) mejorar las teorías a la luz de nuevos datos; y (4) combinar datos. Por tanto, en futuros trabajos seguiremos explorando nuevas fuentes de datos, a veces utilizándolas para mejorar los modelos teóricos, otras veces para crear simulaciones más realistas en las que múltiples fuentes de datos sean combinadas. Por ejemplo, actualmente estamos trabajando con un nuevo conjunto de datos que posee información sobre las rutinas diarias de los trabajadores en un hospital, con el objetivo de diseñar estrategias efectivas para la reducción de la propagación de las infecciones asociadas con la atención sanitaria.

Finalmente, en el capítulo 4, vimos dos ejemplos de comportamientos colectivo *online*. En el caso de Forocoches, todavía hay mucho trabajo por hacer. En lo que respecta a su dinámica, podemos añadir intensidades de fondo no constantes para caracterizar más adecuadamente el comportamiento de los hilos, o explorar más a fondo la relación entre el éxito de un hilo y su contenido (o con los usuarios que participan en él). Es más, estos datos pueden ser utilizados también para estudiar la creación y evolución de *memes* que brevemente hemos mencionado, o como complemento para el análisis de eventos que actualmente se estudían principalmente usando datos de Twitter. Por otra parte, el trabajo de Twitch Plays Pokémon puede ser considerado, por el momento, cerrado, aunque ha dado lugar a una serie de ideas sobre imitar las reglas del juego en sistemas más sencillos, con el objetivo de realizar experimentos controlados sobre el comportamiento y la organización de los grupos humanos. No obstante, este capítulo ha mostrado también las numerosas oportunidades de investigación que existen los sistemas *online*, en ocasiones conectadas con el mundo *offline* "clásico", en otras con características completamente diferentes. Estaremos vigilantes y cuando lleguen nuevos datos y se descubran nuevos fenómenos, los exploraremos.