Exploring patterns of empirical networks

Luis Enrique Correa da Rocha

Doctoral Thesis
2011

Department of Physics
Umeå University

Department of Physics
Umeå University
SE-90187 Umeå, Sweden

ISBN: 978-91-7459-241-2

Printed by Print & Media, Umeå 2011

Abstract

We are constantly struggling to understand how nature works, trying to identify recurrent events and looking for analogies and relations between objects or individuals. Knowing patterns of behavior is powerful and fundamental for survival of any species. In this thesis, datasets of diverse systems related to transportation, economics, sexual and social contacts, are characterized by using the formalisms of time series and network theory. Part of the results consists on the collection and analyzes of original network data, the rest focuses on the simulation of dynamical processes on these networks and to study how they are affected by the particular structures. The majority of the thesis is about temporal networks, i.e. networks whose structure changes in time. The new temporal dimension reveals structural dynamical properties that help to understand the feedback mechanisms responsible to make the network structure to adapt and to understand the emergence and inhibition of diverse phenomena in dynamic systems, as epidemics in sexual and contact networks.

Sammanfattning

Vi är ständigt kämpar för att förstå hur naturen fungerar, försöker identifiera återkommande evenemang och söker analogier och relationer mellan objekt eller individer. Veta beteendemönster är kraftfull och grundläggande för överlevnad av arter. I denna avhandling, dataset av olika system i samband med transporter är ekonomi, sexuella och sociala kontakter, som kännetecknas av att använda formalismer av tidsserier och nätverk teori. En del av resultatet utgörs av insamling och analys av ursprungliga nätdata, fokuserar resten på simulering av dynamiska processer i dessa nätverk och att studera hur de påverkas av de särskilda strukturer. Huvuddelen av avhandlingen handlar om tidsmässiga nät, i.e. nät vars struktur förändringar i tid. Den nya tidsdimensionen avslöjar strukturella dynamiska egenskaper som hjälper till att förstå den feedback mekanismer som ansvarar för att göra nätverksstruktur att anpassa sig och förstå uppkomsten och hämning av olika företeelser i dynamiska system, epidemier i sexuella och kontaktnät.

Resumo

Constantemente nos esforçamos para entender como a natureza funciona, tentando identificar eventos recorrentes e procurando por analogias e relações entre objetos ou indivíduos. Conhecer padrões de comportamento é algo poderoso e fundamental para a sobrevivência de qualquer espécie. Nesta tese, dados de sistemas diversos, relacionados a transporte, economia, contatos sexuais e sociais, são caracterizados usando o formalismo de séries temporais e teoria de redes. Uma parte dos resultados consiste na coleta e análise de dados de redes originais, a outra parte concentra-se na simulação de processos dinâmicos nessas redes e no estudo de como esses processos são afetados por determinadas estruturas. A maior parte da tese é sobre redes temporais, ou seja, redes cuja estrutura varia no tempo. A nova dimensão temporal revela propriedades estruturais dinâmicas que contribuem para o entendimento dos mecanismos de resposta responsáveis pela adaptação da rede, e para o entendimento da emergência e inibição de fenômenos diversos em sistemas dinâmicos, como epidemiais em redes sexuais e de contato pessoal.

Publications

The thesis is based on the following publications (reprinted with the permission of the publishers):

I L E C da Rocha. *Structural evolution of the Brazilian airport network*. Journal of Statistical Mechanics: Theory and Experiments (2009), P04020.

II L E C Rocha and P Holme. *The network organisation of consumer complaints*. Europhysics Letters **91** 2 (2010), 28005.

III L E C Rocha, F Liljeros and P Holme. *Information dynamics shape the sexual networks of Internet-mediated prostitution*. Proceedings of the National Academy of Sciences of the USA **107** 13 (2010), 5706–5711.

IV L E C Rocha, F Liljeros and P Holme. *Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts*. Public Library of Science Computational Biology **7** 3 (2011), e1001109.

V S Lee, L E C Rocha, F Liljeros and P Holme. *Exploiting temporal network structures of human interaction to effectively immunize populations*. Submitted to Public Library of Science Computational Biology (2011).

Other publications by the author not included in the thesis:

- L E C da Rocha. *Notícia ruim corre depressa! (Bad news travels fast!).* Revista Brasileira de Ensino de Física **31** (2009), 3303.

- L E C Rocha, F Liljeros and P Holme. chapter *in* Handbook of the Economics of Prostitution. Oxford University Press. *Communication and sexual networks of Internet mediated prostitution*, forthcoming (2011).

- L da F Costa, O N Oliveira Jr, G Travieso, F A Rodrigues, P R Villas Boas, L Antiqueira, M P Viana and L E C Rocha. *Analyzing and modeling real-world phenomena with complex networks: A survey of applications.* Advances in Physics **60** 3 (2011), 329–412.

- S Bernhardsson, L E C da Rocha, and P Minnhagen. *The meta book and size-dependent properties of written language.* New Journal of Physics **11** (2009), 123015.

- S Bernhardsson, L E C da Rocha, and P Minnhagen. *Size dependent word frequencies and translational invariance of books.* Physica A: Statistical Mechanics and its Applications **389** 2 (2010), 330–341.

- M Zanin, L E C Rocha and S Boccaletti. *Dynamics of roles in complex networks.* Submited (2011).

Preface

The story of this thesis started long ago, after I received an email early in the morning with my acceptance for the PhD program. Since then, life was like riding a roller coaster, which means, ups and downs, loops and screams, but fun overall. But before getting to another amusement ride, it is time to write down a summary of my scientific years in Umeå. The idea of this thesis is to give a gentle introduction and shortly review relevant literature to motivate and contextualize the research presented in the papers that I co-authored. Whenever possible, I try to give my own perspective of network science rather than simply reproducing what one can find in other monographs and books.

In general, my research was driven by data analysis and computer simulations. After all, I decided to include five original papers in this thesis, besides six other papers not necessarily original. Each one has a particular story and importance in my carreer. During the PhD program, I tried to not only write and publish research papers but also make a connection with laypeople about my research. Initially, I wrote an article for a physics pedagogical journal in Brazil. The focus was more on teaching information spreading on networks than on producing original research. Later on, this connection was indirect; three papers that I co-authored got the media attention, from regular newspapers to popular science magazines in different languages. For me, it was really important to establish this connection to people who is actually funding our research, and surprisingly exciting to see the non-scientific feedback. Within the scientific community, some of our papers crossed the borders of physics and were noticed by Scott Cunningham, an economist who kindly invited us to contribute a book chapter to the Handbook of the economics of prostitution. More recently, together with my Brazilian collaborators, I had the pleasant experience to get a review paper published in a prestigious physics journal. I do hope this review also cross the borders of physics and be useful for diverse scientists out there.

During my earlier years as a PhD student, I have being involved in a project about books. The main hypothesis was that when an author writes down a text, he or she picks the words randomly from a large dictionary in his or her brain. As a consequence, the word-frequency distribution depends on the length of the written text and is characteristic of each author. That was certainly an interesting project but a little detour of my research goals, especially, from the social and technological

systems I wanted to study. Therefore, to wrap up the thesis focusing on networks, I decided to skip the two papers related to the meta-book concept. I am also skipping a project related to tracing the evolution of network structure by using node similarity measures. This project is within the scope of the thesis, but since was concluded during the writing of the thesis, I chose to simply acknowledge it here.

The following chapters are supposed to be as much non-technical as possible and as such, readable by a scientist of any area of expertise. Indeed, even non-scientists might find it pleasant to read. Technical details appear in the text but are mostly left to the attached manuscripts. I start the thesis (Chapter 1) explaining the thesis title and motivating the relevance of pattern detection in nature. In the next chapter (Chapter 2), I introduce definitions and basic concepts of networks. In Chapter 3, some issues and challenges related to data collection are discussed together with some thoughts about the potential of experimental network science. Chapters 4 and 5 are dedicated to introduce concepts and measures directly related to the published papers, including a brief historical overview, and a summary of each paper. The thesis closes with a chapter (Chapter 6) summarizing and connecting the main contributions, and with little speculation about future research directions. I finally suggest the reader to get a comfy seat, make a *caipirinha*[a] and enjoy the reading.

Luis E C Rocha
July 1, 2011
Umeå, Sverige

[a]There are strict (official) recommendations about the *caipirinha* making, but essentially it consists on adding ice cubes in a lowball glass, a shoot of Brazilian *cachaça*, and limes smashed with sugar (from sugarcane!).

Acknowledgment

The knowledge and experience accumulated during my doctoral program go much beyond the publication list. A great deal of them comes in form of collaboration and support from several people. The contributions are not always obvious, sometimes, a little comment made me change radically my style and behavior; some other times, patient people continuously supported my babe steps improvement. I am grateful for all those who helped me during these years, even though, my lousy memory will possibly forget some names over here.

My wife, Luana, deserves a special spot. More than love and moral support, as a good physicist, she was always the first referee of my work and the one that always heard my ideas and helped me to mature (or trash) them. The patience and love of my closest family, João and Elizabete, Marcos and Paula, Maria, Regina and Adilson, and of all other Correa da Rocha, Antonini, Greco, de Freitas Nascimento, and Roveran, are needless to acknowledge.

Certainly Petter Holme was a key player in this story, his flexibility and inspiring scientific attitude guided and provided me great experiences during these years. My closer collaborators, Fredrik Liljeros, Sungmin Lee, Luciano da F Costa, Martin Rosvall, Matheus Viana, Massimiliano Zanin, Sebastian Bernhardsson and Petter Minnhagen provided me invaluable lessons about science and research practices. All other IceLab members and friends, Alcides Esquivel, Atieh Mirshahvalad, Beom Jun Kim (unofficial member), Etsuko Nonaka, Fariba Karimi, Henrik Sjödin, Hong Zhang, Jan Ohst, Magnus Lindh, Sang Hoon Lee, Seung Ki Baek and Han Hwa Choung (unofficial member), Zhi-Xi Wu, Zinzhu Meng, and Åke Brännström, for all cool "fika"s, dinners, scientific and non-scientific conversations. The feedback of Tamás Nepusz, Gonzalo Travieso, Aaron Clauset, Devon Brewer, and Diego Rybski contributed to improve the quality of my papers and should be remembered. I am also grateful to Marco Aurélio Ubirajara Garcia Gomes for providing me an opportunity several years ago that surely contributed to many decisions of my life, especially related to working outside Brazil.

I am in debt to the helpful colleagues of the physics department, in special, Ann-Charlott Dalberg, Katarina Hassler, Lena Burström, Jörgen Eriksson, Kjell Rönnmark, and Peter Olson. Good lunch times and coffee-breaks were enjoyable with many other people from the physics department. Funny moments will be also

remembered with all those people I met during the IKSU ski-trips, in scientific conferences and schools, or in the daily-life at university; I simply have to skip all those names here.

After all, life in Sweden gave me some closer friends that participated in my life in different ways. Apart from some names already cited, thanks to Luca Mana and Louise Löwenberg, Mitja and Alëna Kobyakov, Robert and Emma Saers, Michael Bradley, Alice and George Bezett, André and Rebecca Haraldsson, Sandra and Joachim Sandström, Francisco and Fernanda Rivera, Aleksandra and Piotr Matyba, Ligia and Kjell Lundin, and Anna Lundin.

I wanna thank Filip Vanhavere and his research group in SCK–CEN for friendliness, and for giving me a spot to work while visiting Luana in Belgium. Hawoong Jeong and Kim Sneppen provided the same hospitality during my visits to KAIST in South Korea and Niels Bohr in Denmark. My transition to a post-doc life was made much smoother thanks to the welcome attitude of Vincent Blondel to Louvain-la-Neuve in Belgium.

A special thanks goes to the Swedish taxpayers, who supported my job here in Sweden, and to the multitude of volunteer programmers engaged to provide and maintain Linux systems and other free software that I have extensively adopted over the years. I should also thank Microsoft Research and SJCKMS foundations for partially supporting my participation, respectively, in NetSci2008 and in NetSci2011.

Contents

Chapter 1

Introduction

The human being has an inherent capacity of detecting patterns of all sorts. It actually seems that we have a natural impulse to hunt for patterns. Identifying patterns is indeed a key to success and overall, a necessity of daily life; once the pattern is detected, in principle, the future can be predicted. That is powerful. Ordinary people constantly struggles in this quest, they have to learn how much time is needed to heat a cup of water, how frequent the buses depart from one station to another, and more important, to please their peers, they have to learn how they behave, and so on. Interestingly, humans are not so good on identifying random events[a]. Experimental psychology has extensively studied such phenomena, a classic example is the Gambler's fallacy, where subjects believe that an increasing succession of looses in a game, makes a win situation more likely [1]. This apparent contradiction creates, sometimes, a blurred border between randomness and order that is difficult to distinguish. Mathematics, however, helps to identify the border between these two states [2].

> *"A mathematician, like a painter or a poet, is a maker of patterns."*
> *G H Hardy, 1941 [3]*

To detect a pattern, first one has to define a system to study and observe its behavior under certain circunstances. This observation is usually performed through an experiment, where the experimenter controls some parameters and observes the outcome [4]. Sometimes, there is no control, simply observation of the phenomenon. Taking one or another direction depends on several factors as cost, time, technology, ethical issues, etc. The most common approach is to take small parts of a system and study their properties separately. This works well indeed. The problem appears when one wants to understand the emergent phenomena created by collective behav-

[a]There is an entertaining short article illustrating simple experiments where we fail to detect random patterns, the name is "The illusion of randomness" written in 2001 by Richard A Muller, just google it on the web.

ior when these parts interact, that is generally not simply the sum of the individual parts [5]. What happens then is that the observation has to happen at another level, which is more difficult to control due to the number and complexity of parameters that an experimenter would have to deal with. These challenges motivate the observational approach where the system is observed as it is. A different configuration of the system is obtained by the natural system conditions and evolution, and not by the experimenter control.

From a pragmatic point of view, an observation can be seen as the act of collecting a set of properties of an entity at a certain moment in time and in space. The trick is to map these properties into numbers and organize them in a meaningful structure.

> *"... if you graph the numbers of any system, patterns emerge. Therefore, there are patterns everywhere in nature."*
> *Max Cohen, main character of the movie PI, 1998 [6]*

There are some ways of organizing these numbers such that patterns emerge, one vastly adopted method is to arrange the measured values in mathematical series that is a sequence of ordered numbers. Another idea is to organize them in a graph, which is simply a collection of objects connected pairwise by a line representing the relation between them. Once these structures are defined, the goal is to find regular or irregular patterns, investigate their meaning and the reason of their emergence. By knowing these patterns, ideally, one can predict and interfere in the future state of the system.

This thesis is essentially about identifying patterns in diverse systems using mathematical methods, in special network structures. The goal is to not only measure them, but whenever possible, use that information to understand the dynamics of society both at the individual and at the technological level.

Chapter 2

Basics of networks

In the colloquial sense, a network is simply a bunch of connections. Networks emerge when connections exist between objects. This is such a general concept that one finds networks everywhere, some are more conspicuous, as the fisherman net or a spider web, others more obscure, such as a network of related diseases [7] or subsequent notes in a melody [8]. In daily life, people are continuously thinking in networks; when one goes to work, he or she is aware of which pathways are possible, and that some choices are faster than others. It is usually not difficult to identify gossip super spreaders among the peers, and to cite a historical situation, even the medieval Catholic church realized that certain individuals had disproportionate influence to spread heresies [9].

Nonetheless, the network view of daily life was formalized, apparently for the first time, by Leonard Euler in his classic study of the Königsberg bridges in 1735. A reader familiar with Latin can appreciate the original manuscript about this problem in ref. [10], others can simply check ref. [11] for an overview. Afterwards, networks in one form or another arose in different fields of the science, from Kekulé's diagrams to Kirchhoff's circuit laws. Only in 1878 the term graph (the mathematical term for networks) was introduced by James J Sylvester to describe such general structures visible everywhere [12]. Since then, mathematicians worked hard to study properties of general graphs, and more recently, random graphs [13]. In parallel, mostly after the 1950s, social scientists applied network ideas to understand the impact of social ties on human relations [14]. Needless to say, science is a continuous activity strongly influenced by the past. Even breakthroughs happen due to a combination of previous results and other factors, which create the propitious environment for new ideas to emerge. Such conditions were apparently met in 1999, when Barabási and Albert published a seminal paper about the growth mechanism of the world wide web [15]. Though networks have been reinvented and applied in different fields along the years, after this publication, network science took the form as it has nowadays. Yet, graph theory and social networks have their own path, many times, overlapping or complementing network science.

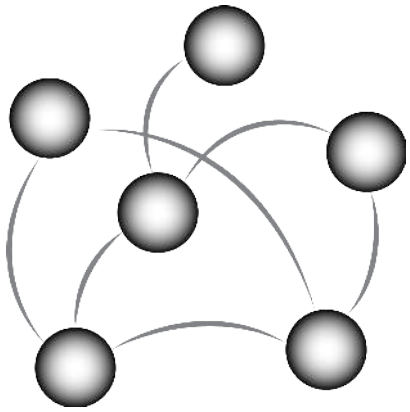

Figure 2.1: *A simple network with 6 vertices and 7 edges.*

2.1 Definition

Due to historical reasons, the terminology and many definitions used in network science are borrowed from graph theory and social networks [13, 14]. Since there are some differences though, it is safer to define terms as they are used by the so-called complex network community [16].

In its simplest form, a *network* Γ is formally defined as a *graph*, by a set $V(\Gamma) = \{i\colon i = 1, 2, \ldots, N\}$ of *vertices* (or nodes) and a set $E(\Gamma) = \{(i, j)\colon i \text{ and } j \in V(\Gamma)\}$ of *edges* (or links, or ties). Sometimes one refers to a stub, i.e. a piece of an edge; two connected stubs form an edge. A vertex typically represents an object (or an individual), while an edge represents a relation between two objects (or the same object). A pictorial view of a graph with 6 vertices and 7 edges is presented in Figure 2.1. The number of vertices and edges in the network are given by the size of the sets, respectively, $N = |V(\Gamma)|$ and $E = |E(\Gamma)|$. A *sub-network* κ (or group of vertices) of Γ is defined as a set of vertices $V(\kappa)$, such that $V(\kappa) \subseteq V(\Gamma)$ and $E(\kappa) \subseteq \{(i, j)\colon (i, j) \in E(\Gamma) \text{ and } i, j \in V(\kappa)\}$. A subnet can contain, for example, only one vertex, the original network, or a null number of vertices. In the following chapters, a mathematically informal description of the network is preferred.

2.1.1 General networks

The previous definition is simple, elegant, and contains all necessary ingredients to define a network for any system. Nevertheless, it can be extended to include more specific information of the system. In general, both the vertices and edges can have quantities related either to intrinsic properties of the objects and relations they are representing, or to the network structure itself.

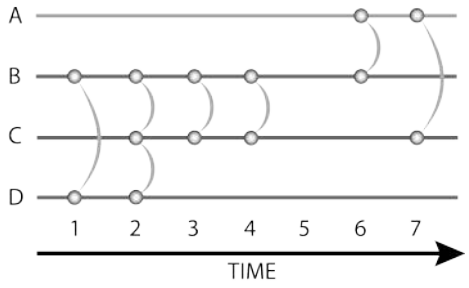

Figure 2.2: *An illustration of temporal network; vertices correspond to horizontal lines, and the vertical curved edges are available only at certain moments in time.*

Since a vertex represents an object, it can have an identity[a], an age, a geographic position, or belong to a specific class of vertices. If the network has two classes of vertices, for example, representing man and woman, this network is said to be bipartite. In a dynamical description of a network, a vertex can be active or inactive at certain moments in time. Generally speaking, any observable property of an object can be mapped into a vertex intrinsic quantity.

The edges in turn can be directed, to represent an asymmetry in the relation between two vertices, or undirected to disregard this factor. A typical example is that a person A might consider person B as a friend but the reverse is not necessarily true. Many times, asymmetric edges are removed to obtain a stronger definition of connectivity, i.e. reciprocal edges are stronger than the unidirectional ones. Yet another way to define weaker and stronger edges is by mapping real values, termed weights, to edges in order to represent the strength of ties. To specify the moment an edge is available or not, a time stamp can be associated to an edge (Fig. 2.2). Note that these definitions are very general and by simple operations, one representation can be converted to another. For example, a thresholding operation converts a weighted edge into a simple unweighted edge, or the removal of time-stamps converts a temporal network to its static version. Note, however, that once the information is lost, the reversed operations are not as simple.

Networks can be also seen as multiple layers of vertices (the same or not) connected by different edges [17, 18, 19, 20]. Such multi-layered networks are also called multiplex. A typical situation is the fact that people have different relations (friendship, romantic, workmate); each of them is a different contact network but they can be combined into one multi-layered network (Fig. 2.3). The multi-layered network is adequate when one wants to study the interdependence of different layers, for example, the fact that firefighters organize themselves into communication networks to control forest fire [18], or to study the Internet breakdown due to cascading power grid blackout [20].

[a]The true identity is usually replaced by aliases due to privacy issues.

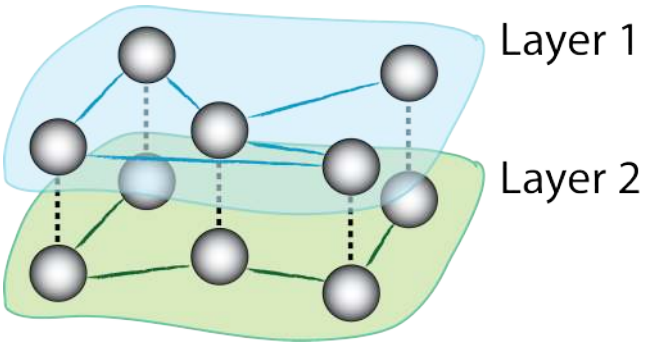

Figure 2.3: *An illustration of a 2-layer or multiplex network. Both layers contain the same 5 vertices, but the connections between them differ on each layer.*

2.1.2 Data structure

Although networks are defined by using sets (from the formalism of graph theory, see §2.1), working with them can be rather cumbersome on network problems. A more suitable network representation is obtained, either as a matrix, or a list. For practical reasons, the choice of one or another structure depends on the purpose of the study. In the following, benefits and limitations of these data structures are discussed.

Matrix

There are more than one way of mapping a network into a matrix. The most common one is, perhaps, the *adjacency matrix*. In this representation, a NxN matrix A is defined such that each row i or each column j represents one single vertex ($i = j$, the same vertex). If two vertices are connected, the respective entry $a_{ij} = 1$, otherwise $a_{ij} = 0$. There is a natural extension for weighted networks when the binary value can be replaced by a real value $\mathbb{N}$, resulting in the so-called *weighted matrix* W. An example of a weighted network and the respective weighted matrix is given in Fig. 2.4.

The directed network is represented by making $a_{ij} \neq a_{ji}$. For the special case of bi-partite networks, to reduce the matrix size, rows and columns can represent vertices of different types, and the direction in the edges can be included by using plus or minus signs in the respective entries.

Another less common representation in network science but more popular in social networks is the *incidence matrix* I. In this case, the network is mapped into a matrix of size NxE where the entry $i_{ij} = 1$ if the vertex and edge are incident, and $i_{ij} = 0$ otherwise.

A dynamic network can be represented either by a 3-dimensional matrix where the new dimension represents different time steps, or in case of knowing the mech-

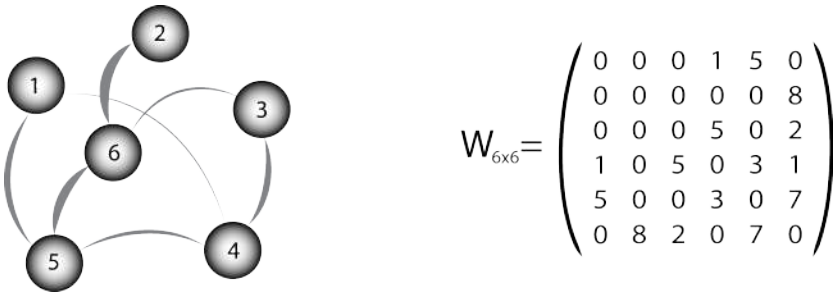

Figure 2.4: *An illustration of a weighted network on the left, edge thickness represents different weights. On the right, the corresponding mapping of the network into a weighted matrix.*

anism of edge evolution, a temporal function can be defined for a 2−dimensional adjacency matrix rather than the usual single value, i.e. $a_{ij} = f_{ij}(t)$.

The matrix format is usually adopted due to the viability to perform matrix operations, both in mathematical and in computational contexts [21]. The most important shortcoming of the matrix data structure is that since the network is usually sparse ($E \ll N^2$), the majority of the entries are zeros and, for large networks, this representation wastes too much computer memory.

List

A widely used solution to deal with sparse matrices is to convert them into lists. In general, direct matrix operations are more difficult to perform but several other operations, e.g. queries, are faster. Technically, the *adjacency list* is an array of lists, where each list corresponds to a different vertex i and contains the vertices that are connected to i[b]. In such a structure, directed networks are represented by simply removing a vertex from the respective list of vertex i. It is not so trivial to include weights in this structure, but depending on the network, some mathematical tricks can be used, for instance, by multiplying the vertex identity by 1000 and adding values (i.e. weights) in the range $[1, 999]$. The original identity can be recovered by dividing by 1000 and retaining only the integer part; the weight is the remainder of the division. This structure is more useful when one wants to rapidly identify the neighbors of a vertex (without weight). An example of adjacency list from the unweighted version of the network in Fig. 2.4 is presented as table 2.1.

In the *incidence list*, however, each entry corresponds to an edge, and thus contains the identity of the pair of connected vertices and respective edge-weights (if any). This structure is very practical and generally used to store network data into files. Incidence lists are especially suitable to represent networks with temporal in-

[b]One can think in this array as a matrix where each row corresponds to one vertex i and the columns k contain the identity j of the vertices adjacent to i.

Vertex	Neighbor 1	Neighbor 2	Neighbor 3
1	4	5	∅
2	6	∅	∅
3	4	6	∅
4	1	3	5
5	1	4	6
6	2	3	5

Table 2.1: *An example of adjacency list corresponding to the unweighted version of the network in Fig. 2.4. Note that ∅ corresponds to an empty entry.*

Vertex 1	Vertex 2	Time	Weight
B	D	1	1.5
B	C	2	2.2
C	D	2	0.5
B	C	3	1
B	C	4	5.1
A	B	6	2
A	C	7	1.3

Table 2.2: *An ordered list corresponding to the temporal network of Fig. 2.2. The weights are random numbers shown to illustrate the data structure.*

formation. Since each entry corresponds to an edge, the instant the edge is active can be represented by simply setting the time-stamp in front of the connected pair. An *ordered list*, therefore, can be obtained such that the initial entry is older than the final entry, as table 2.2.

2.2 Models

A network model is aimed to reproduce the structure in some way. In essence, it is a sequence of instructions to decide which vertices are connected to whom. For practical purposes, network models can be either *empirical* or *theoretical*. In my terminology, the empirical models are simply the graph representation of the observed relation between objects. In other words, the system is converted into a structure described (modeled) by a graph. The theoretical counterpart includes: generative models intended to create synthetic graphs to study the dependence of network structure; mechanistic models for experimenting about hypothetical mechanisms behind the network evolution; and null models, for hypothesis testing of the empirical structures. It is common in the literature to only refer to the theoretical

models as network models, while the empirical models (as I call) are simply the empirical network, represented mathematically as a graph. Although these definitions might rise some philosophical debate, the terminologies are equivalent in practice.

2.2.1 Empirical

In principle, it is possible to build a network for any natural or artificial system. One only needs to define a set of vertices, corresponding to specific objects of the system, and a rule to connect those vertices. There is indeed a plethora of networks of the most diverse systems. In an attempt to collect a reasonable sample of the state-of-the-art, in a *tour de force*, together with some collaborators, I co-authored an extensive paper dedicated to review networks constructed from real data [22]. To illustrate the empirical modeling, in the following sections some examples of empirical networks are presented, separated in broad categories according to the general method adopted to construct them [22]. This categorization is for pedagogical reasons and certainly incomplete. The original datasets used in this thesis are described separately in §3.1.

Proximity

The proximity method is based on the idea of connecting objects that are spatially or temporally close. A typical example is the spatial network formed by connecting the center of regions defined by a Voronoi tessellation or connecting overlapping crossing points of a Delaunay triangulation [23], the first method being extensively used by network practitioners to create geographical networks. A more abstract proximity network is obtained by connecting subsequent words (the vertices) in a text [24], or subsequent notes (the vertices) in a melody [8]. In this category one can also include situations where two individuals are in contact during a certain period, as the case of sexual contacts [25].

Coexistence

The coexistence method is performed by connecting a group of objects that shares a common existence. This method generates a bi-partite network because the object is represented by one type of vertex while the "common existence" is regarded as another type of vertex. The original bipartite network can be converted into a unipartite structure by simply connecting objects that share the common place, or vice-versa. The bipartite representation is more informative but at the same time more difficult to analyze. This method has been extensively adopted by the network science community, for example, to form networks of individuals (vertex type A) connected because they co-authored a paper (vertex type B) [26], or because they shared the same ward (vertex type B) in a hospital [27].

Communication

Networks created by using this methodology are also very common in the literature. The idea is straightforward, simply connect two individuals or devices, if there is a communication channel between them. Popular case studies are based in mobile phone calls [28] or email exchange [29]. In those systems, the mobile phone or the email owner defines a vertex, and the edge is established if one individual called or emailed another one. These networks typically contain the number of calls (emails) between the same pair of individuals as weights. Recently, phone calls and emails have been used in the context of temporal networks (see §2.2.2 - Temporal) because the time-stamps of the calls(emails) are commonly available in the datasets.

Confluence

The confluence method resembles the proximity method in some aspects. In this case, vertices are defined by the crossing of different pathways such that the pathways define the edges and the crossing points define the vertices of the network. Road [30] and digital integrated circuit [31] networks are natural examples, where the intersection of two or more roads, streets, or electronic circuits (the edges) define a crossing point and thus a vertex.

Correlation

This method is more unconventional and is based on representing a time-series (see §4.3 for a definition of time-series) by vertices and connecting these vertices according to the correlation between the respective series. The mechanism has been used in fields where time-series are popular and dataset abundant, as financial [32] and climate [33] systems. A general method to construct networks from time-series, including a review of the literature, is presented in ref. [34]. This method captures density variations in the phase-space of the underlying dynamical system into network structures [34].

Reference

In the reference network model, objects or individuals are connected if they refer to each other. These networks are especially important to understand how ideas and influence spread (or is created) in some environments, as in science or the virtual world. A classical example is the citation networks, where new articles refer to published articles, and consequently, create directed edges between the corresponding vertices [35]. Another example is the world wide web, where websites refer to each other through links [36].

2.2.2 Theoretical

In order to understand how the empirical structures arise and to reproduce them, it is a common practice to create a minimalistic model containing the essential ingredients to obtain those structures. Theoretical network models are generally stochastic processes aimed to reproduce certain structures of interest, not necessarily observed in empirical networks. As part of the cycle of science, many times the models are validated empirically by comparing with the structure of real networks [37]. As with empirical networks, there is a full course menu of theoretical network models, for some important models see ref. [16]. In the following sections, however, only some relevant methods are reviewed.

Random

Most theoretical network models are indeed random and created by a stochastic process. Nevertheless, one usually refer to the *random network* as the network formed by connecting pairs of vertices with a uniform probability, or as in Solomonoff and Rapoport original words [38]:

> *"Consider an aggregate of points, from each of which issues some number of outwardly directed lines (axones). Each axone terminates upon some point of the aggregate, and the probability that an axone from one point terminates on another point is the same for every pair of points in the aggregate. The resulting configuration constitutes a random net."*

This network model popularized after the formal proposal and subsequent study of random graphs by Erdős and Rényi [39] among others [13]. To distinguish this uniform probability model to other random models, many researchers refer to this protocol as the Erdős-Rényi (or ER) model. The beauty of this model is its simplicity, it only requests the number of vertices and a probability defining how likely two of the vertices are connected. One of its important characteristics is the fact that the vertices are, on average, connected to the same number of other vertices. The dark side is that empirical structures simply do not follow these random structures and in practice the random assumption is used only as a null model (see bellow about null models).

Preferential attachment

The history of preferential attachment (also know as cumulative advantage or richer-get-richer paradigm) models is relatively old; the first formal study dates back to G Udny Yule who created a process aimed to explain the distribution of number of species per genus of flowering plants [40]. This is quite extensive and difficult paper to read; a more readable general derivation of a preferential attachment model is proposed by Herbert A Simon to explain power-law distributions in diverse systems,

e.g. the city-population, frequency of words in a book, income [41]. These models, nonetheless, are general mechanisms not directly related to networks. The first known attempt of a *preferential attachment network* model is due to Barabási and Albert (also known as BA network). In 1999, they proposed a model, now known to be a special case of Simon's proposal [42], where they not only rediscover the preferential attachment process but also creates a network structure not accounted in previous models [15]. This seminal paper marks the breakthrough of modern network science.

Their mechanism is relatively simple, intuitive, and creates networks with a peculiar property where the distribution of edges per vertex obeys a power-law distribution, i.e. there is no characteristic scale as in the random model. In their original publication [15], the network growth is proposed in the following way:

> *"... starting with a small number (m_0) of vertices, at every time step we add a new vertex with $m(\leq m_0)$ edges that link the new vertex to m different vertices already present in the system. To incorporate preferential attachment, we assume that the probability p_i that a new vertex will be connected to vertex i depends on the connectivity k_i of that vertex, so that $p_i(k_i) = k_i / \sum_j k_j$."*

Connectivity here is the number of edges connected to a vertex (see more about connectivity in §4.1). One inconvenient of this model is that older vertices are more likely to have more edges, which might not be the case in some networks.

Dynamic

Networks that change the structure over time are called *dynamic networks*. Such networks have been studied in the context of spread of diseases and opinions [43, 44], and game theory [45]. One example is the model proposed by Volz and Meyers where the vertices are initially randomly connected and over time, they change partners with a constant probability such that the number of edges (and vertices) are maintained [43]. Within this class, there is also the so-called *adaptive networks* where the network topology depends on the dynamics on the vertices and a feedback of the new network structure leads to new dynamical states [44]. A beautiful biological example is the vessels (the network) controlling the blood flow in the body (the dynamics); when there is some restriction of blood supply (ischemia), new arteries are formed to improve the local flow, which implies in a new network structure [46]. An illustrative theoretical example is the study of epidemics affecting the network structure in the model by Gross and co-authors. Starting with a random network, a susceptible individual breaks an edge with infected vertices (with a certain probability) and then randomly rewire to another susceptible [47]. Dynamic networks are rather unexplored field, one potential application is to use them to model the influence of human behavior on the spread of diseases [48], and human reaction to other critical events as traffic jams or cascade failures.

Temporal

A *temporal network* is also a dynamic network but with a different characteristic; the edges and vertices can be in active or inactive states, which means that the number of edges and vertices available vary throughout time. In paper IV [49], a random temporal network is proposed where vertices are assumed to be constantly available but the edges appear only at certain instants in time, otherwise, they are idle and cannot be used. One procedure to create this temporal network is to initially define a rate of activity, or in other words, a probability p_i for each vertex i to get an active stub. Afterwards, following the set of rules bellow during a desired time interval, one obtains a network structure.

For each time step:

1. *For all vertices: with probability p_i, activate a stub of vertex i*
2. *Collect all active stubs in a box and randomly connect pairs of vertices in this box*
   ⇒ If one stub is left alone, keep it for the next time step
   ⇒ This procedure generates an instantaneous network at this time step
3. *Inactivate all the edges and return to 1) for the next time step*

This mechanism is very general and the choice of p_i defines the structure of the network in terms of the number of edges per vertex (or degree distribution, see §4.1). If $p_i = \text{stubs}_i/\Delta T$ (where stubs_i is the number of contacts – or degree §4.1 – vertex i has during an interval ΔT), or in other words, given a distribution of number of edges per vertex, one can create a simple temporal network by distributing these edges uniformly in time.

Null networks

Network practitioners usually compare measured structures with null models to assess if a certain structure is representative or could be simply expected by chance. The concept of null models come from hypothesis testing in statistics [50] and was originally coined by Ronald Fischer in 1935 [51]. The null hypothesis, usually a default position assumed to be true, is challenged against a new hypothesis based on empirical evidences; if the difference between the hypotheses are statistical significant (typically, if the new hypothesis is expected to be observed more than 5% of the time), one rejects the null hypothesis, otherwise, no conclusion is made. An illustrative simple example is to compare the average score μ of pupils in two schools, the null hypothesis H_0 is that they are the same ($H_0 : \mu_1 = \mu_2$) and the new hypothesis may be that one is larger than the other ($H_1 : \mu_1 > \mu_2$).

This formal analysis has been used to identify overrepresentation of motifs in

biological circuits [52], but in practice, has not been extensively adopted by the network science community in physics, computer science, and engineering. The typical approach (the same followed in this thesis) is to simply compare the first and second moments of empirical network quantities with null networks, and discuss the differences. For practical purposes, a gentle definition of a null model is given by Gotelli and Graves [53]:

> *"A null model is a pattern-generating model that is based on randomization of ... data or random sampling from a known or specified distribution. The null model is designed with respect to some ... process of interest. Certain elements of the data are held constant, and others are allowed to vary stochastically to create new assemblage patterns. The randomization is designed to produce a pattern that would be expected in the absence of a particular ... mechanism."*

The choice of the random network model depends in which structure one wants to study. At the same time that one model is adequate to study, for example, the pattern of number of edges per vertex, other models are more appropriate to analyze the existence of a group of highly connected vertices. The choice of a null model is indeed fundamental to properly assess the significance of observable structures. It is still unanswered, for instance, the question of which are the proper null models for studying temporal networks.

The previous theoretical models presented in this thesis can be certainly used as null models. In the early years of network science, a common practice was to compare the empirical network structure with the random network (see above the definition). This is an adequate null hypothesis if one wants to show that the vertices do not have a characteristic number of edges. Nevertheless, it is known nowadays that the number of edges per vertex constrains the connectivity of vertices and thus creates correlations in the network structure. The current approach is therefore to reshuffle the edges maintaining the number of stubs per vertex fixed. One way to perform this efficiently is to select two pairs of vertices and swap the reciprocal pairs, repeating the procedure 4 times the number of edges in the network [54]. A similar procedure can be performed to obtain a null model for temporal networks; two edges are selected at random and their respective time stamps swapped accordingly [55, 49].

Other models

It is important to cite some other simple network models widely used not only in network science but in other areas of science. The following models are not random but follow deterministic rules. The most popular perhaps is the lattice structure. Lattice network structures are usually characterized by a high degree of regularity of the connections and are found for example in crystal structures [56],

hyperbolic structures [57] or when the space is discretized into grids [18]. Another popular structure is the fully connected network, also called panmixing or well-mixed assumption in some contexts as in epidemiology [58, 59, 60]; in this case all vertices are connected to all other vertices. In both cases, the network structure is relatively trivial but still, these network models are widely used since they are relatively simpler to work analytically in comparison to the previous random models. As such, they are frequently used as a first attempt to include network structure in a mathematical model.

Chapter 3

Experimental network science

Network science is generally seen as a theoretical discipline although often relies on empirical datasets. Independently of labels, collection of empirical data requires an experimental procedure and consequently, brings a plethora of new challenges and limitations that many times some individuals are not aware of. Since this thesis is based on the study of original empirical networks, it is reasonable to dedicate a chapter to shortly discuss the data collection procedures and limitations of the studied datasets. Not going deeply in the subject, a short essay speculating about a potential experimental network science is presented.

3.1 Data collection

The collection of network data follows different protocols depending on the source of data and goal of the project. Biological data usually come from databases which are voluntarily updated with the outcome of individual (many times, at a small scale) experiments, for instance, the MIPS database for protein-protein interaction [61] or the KEGG database for molecular interaction and reaction networks [62]. On the other hand, in social sciences, interviews or questionnaires are performed in selected samples of the population [63, 64]. Many data sets used in network science are acquired through *observational studies*, sometimes called *natural experiments*. In those cases, data are collected during a specific event where the researcher usually has no control of the variables. This is especially common in large scale social, economic, technological, ecological and epidemiologic problems. In practical terms, the network variables of a system (e.g. the pairs of connected vertices) are simply recorded during the event of interest, for example, the structure of contact pairs can be collected during a disease outbreak, and for comparison, under a disease-free state.

Information acquired by observing a representative subset of a population at a defined time is called *cross-sectional data*. In contrast, *case-control studies* are concerned about studying a group of individuals with a specific characteristic. This

last category is relatively common in network research. Both methodologies give rise to the concept of static network since usually one wants to learn about the network structure irrespectively of the order of the contacts. *Longitudinal studies*, however, consists on repeated observations of the same subjects (or objects) over periods of time. This type of research has increased in recent years and is connected to temporal networks. In this case, it is possible, for instance, to record the number of edges of a vertex over time. Another form of longitudinal study is the *cohort study*. A cohort is a group of individuals sharing a common characteristic or experience within a period of time.

To provide simple practical examples of the data collection process, the datasets discussed in this thesis are shortly described in the following sections together with the methodology adopted to extract network information from each of them.

Airports and flights

Domestic and international flights are regulated by government agencies that usually keep track of the regular flights within their jurisdiction. In case of Brazil, this information is online and publicly available at an annual basis since 1995. The files are in .pdf (www.adobe.com) and .doc (www.microsoft.com) formats. The files consist of lists containing the routes between origin and destination airports, together with the number of flights, passengers, cargo and post carried over the year. Since the number of flights and airports are relatively small (Table 3.1), the files were converted to plain text and edited by hand, converting airport names to a numerical vertex identity. Different airports serving the same location (Metropolitan area) are merged into one vertex. An edge was defined by connecting the vertices representing the origin and destination airport, and the different quantities mapped as edge-weights.

Consumer complaints

The Brazilian Department of Justice collects and maintains a list of resolved and unresolved complaints received by companies during the previous year. At least for the 2009 dataset, the system covered only about 59.2% of the country population. The list is available as a .pdf file containing the name of the company and the respective complaints, organized in number of events and category of complaint. This file was converted to plain text and the text was automatically parsed by a code written using Python programming language (www.python.org). A bipartite network was formed by connecting companies (vertex type A) that received a certain complaint (vertex type B) (Table 3.1). Most companies had a unique code number that could be used to guarantee the single identity. The original identity was converted to a numerical vertex identity. The same companies reported in different states are grouped as a single vertex. The complaints were in small number and could be visually inspected to remove de-duplications.

Type	No. Vertices	No. Edges	Time / Resolution
Airports and flights	~ 200	$\sim 1{,}000$	12 years / 1 year
Consumer complaints	8,979	17,979	1 year / 1 year
Sexual contacts	16,730	50,185	2,232 days / 1 day

Table 3.1: *Magnitude of the original networks presented in this thesis.*

Sexual contacts

Using a website about commercial sex and sexuality, information about sexual encounters with escorts (sex-sellers) was collected. The website is a forum and as such, escorts are organized in threads created by the forum members (or sex-buyers) reviewing their encounters with a specific escort. New encounters with a previous reviewed escort are reported in an existing thread about that same escort. The webpages were downloaded and parsed using Python. Only anonymous information is used. A bipartite network is straightforwardly obtained by connecting a member who had a sexual encounter with an escort (Table 3.1). Information from the reviews, e.g. type of sex and date, is mapped into vertices and edges accordingly. To maintain the reliability of the Internet forum, the moderators claim to have a constant and intensive screening policy to remove fake reviews, and de-duplication of members and escorts.

3.2 Sampling and accuracy

Every empirical study relies on sampling a real system. One natural consequence is that measured properties and conclusions are (or at least, should be) limited to the observed sample. This apparent trivial statement is often neglected and many times, one observation is (incorrectly) taken as a "proof" of a certain phenomena[a]. Empirical evidence validates a theory, but new evidences can either trash out or strengthen the same theory. Therefore, caution is mandatory when extending conclusions from a specific sample to the whole population. In other words, having studied the network of flights in Brazil does not necessarily mean, *a priori*, that the same structures are observable in the US flight-network.

The sampling procedure itself can be tricky. Sampling edges (by selecting random edges and storing all vertices connected to them) or vertices (by selecting all edges connected to randomly chosen vertices) can lead to different network structures [66]. In special, Stumpf and collaborators discussed the fact that samples of scale-free networks (sampling by choosing random vertices and the respective edges) are not

[a]Having a physicist background, I see *proof* as a rather strong statement (popular during my mathematical analysis courses) and as such, frequently misused in scientific contexts, see ref. [65] for a pedagogical explanation.

necessarily scale-free and this effect is more pronounced for larger scaling exponents [67]. Another common sampling method (e.g. to crawl the world wide web or to track chains of tested positive infected individuals) is the snowball sampling, in this case, one starts in a seed vertex and goes collecting all contacts of the seed; the contacts become the new seeds and the process is repeated until a certain number of vertices or edges are collected. This sampling is highly biased because individuals with many contacts tend to be selected more often and hard-to-reach groups tend to be underrepresented [66, 68]. A method to overcome this bias, widely used in social sciences, is the respondent-driven sampling, which combines the snowball sampling with a weighting function to compensate for the non-random selection of subjects (or objects) [68]. Another approach to assess sampling limitations is to go into the other direction and try to predict missing edges and vertices in the already collected sample. The prediction of missing edges from a given empirical network is indeed an open research question. Some earlier ideas are based on comparing the (intrinsic or topological) similarity of vertices and filling in edges between similar vertices [69]. Another recent approach consists in generating a set of hierarchical random graphs that fits an empirical network, and connect the pairs of vertices that have an average high probability of being connected in this hierarchical structures but are originally not connected in the empirical network [70].

3.3 Experiments

Observational studies draw inferences about a system and frequently are the first step in the cycle of scientific analysis. The next one is the theoretical formulation, i.e. an hypothesis to explain the observation, and a model for prediction based on the hypothesis, followed by experimental validation [4]. The current network research is mostly based in the two initial steps, observation and theoretical analysis, but little in the experimental validation.

Many network studies rely on comparing the network of a particular system configuration with null network models for the same system, and then estimate how much this configuration affects the network structure. Networks obtained during critical or unlikely events are especially of interest when available. Another method is to split the dataset into two parts, one for modeling and prediction, and another for model validation. Computer simulations can be also used to perform experiments (usually called *in silico* experiments). A simple *in silico* network experiment, for instance, is to remove vertices with some desired topological characteristics and then investigate the consequences for simulated disease spread in this scenario in comparison to the original network (i.e. without removal of vertices). Strictly speaking, computer experiments are indeed variations of the parameters of the model, or simply different models for different scenarios. This abuse of the concept of experimentation is completely acceptable but sometimes lead researchers to conclude that

computer experiments are equivalent to real experiments.

Some people can argue that this experimental gap is due to the difficulty to arrange large-scale complex network experiments, mainly because of cost, logistic limitations, ethical issues [71, 72], control of variables, among other restrictions [73]. This is actually true in many aspects; most people would agree, for instance, that infecting a student with a virus to track its spread within a school would be significantly unethical. Nevertheless, besides the studies in biology, some interesting social experiments have been proposed in the literature, for example, the classic Milgram experiment, where random chosen individuals in USA are requested to send letters to a common destinatary (whose name and other info are known), solely by passing the letters to acquaintances that supposedly are more likely to know the target than themselves [74]. A similar experiment, at a global scale, was later performed by using emails rather than regular post with similar results, i.e. about six steps – intermediate contacts – separate any participant [75]. Yet Iribarrena and Moro performed another experiment to study the effect of human activity in the diffusion of information. Subscribers of an online newsletter were rewarded for recommending it; the spread of the offering email was tracked at every step and a chain of contacts were recorded [76]. These few examples illustrate the potential of experimental research.

The availability of mobile sensors has also motivated several observational research about mobility and contact patterns in relatively small environments, and seem to be a promising intermediate step towards experimental setups [77, 78, 79]. However, the increasing coverage of smartphones with their extensive range of applets has not been exploited [80]. To go further on experimental network science, another reasonable idea, not yet fully exploited too, is to use web 2.0 applications [81, 82]. There is a multitude of social networks and virtual environments where people constantly interact and are eager to participate in games and other collaborative activities. Although the "virtual life" does not cover all the scope of network science, it might provide fruitful insights about the interplay of network structure and dynamical processes on the network.

Chapter 4

Network structure

In the previous chapters, general aspects of how to build networks following different mechanisms, and some limitations of networks to describe real systems were discussed. In this chapter, the focus is in methods to quantify structures of networks. Quantifying structures is the first step towards detecting patterns in the network. Once a method to measure some network structure is established, the empirical network can be compared against null models in an attempt to identify non-trivial correlations in the network structure.

Network measures can be roughly divided into at least two categories, those intended to measure static structures and those aimed to characterize the temporal profile of the network. Static measures are rather developed nowadays but still constitute the main interest of most of the network science community. On the other hand, though some methods for dynamic networks have been proposed throughout the years, only recently this branch of network science started to get more attention. In the following sections, some relevant methods and measures from the literature are reviewed, and to complement the chapter, there is a description of a multivariate method applied to the analysis of static network structure.

4.1 Static

Static network measures have been mainstreaming network research since the early days [16]. After all, the network itself is a static structure and as such has to be characterized. To obtain a static network, one has to define a time interval and collect all edges observable within this interval. When multiple time intervals are defined and the respective networks recorded, one refers to each sample as snapshots of the whole network.

Paths and Distances

A *path* σ'_{ij} of length σ'^{d}_{ij}, connecting vertices i and j, is defined as a sequence $S = \{i, \ldots, m, m+1, \ldots, j\}$ of $\sigma'_{ij} - 1$ vertices from the set V, such that there is an edge connecting vertex m with vertex $m+1$ in the sequence.

The *shortest path* σ_{ij} between vertices i and j is defined as the geodesic and sets the topological *distance* (number of edges) between two vertices. The normalized average shortest-path σ over all possible vertices pairs is given in eqn. 4.1. The *diameter* D is defined as the largest geodesic in the network, which is the minimum distance necessary to move between any pair of vertices in the entire network.

$$\sigma \equiv \frac{1}{N(N-1)} \sum_{i \neq j} \sigma^{d}_{ij} \tag{4.1}$$

By definition, any communication process (e.g. a phone call, a walker, an infection) occurring between one vertex and another has to pass through a sequence of intermediate vertices and edges. In this aspect, the shortest path corresponds to the optimal route in terms of shortest distance between any two vertices. Consequently, the diameter characterizes which is the minimum number of hops one has to take to go from any vertex in the network to any other.

Connected Components

If there is a path between vertices i and j, it means that j is reachable from i. If a sub-graph of an undirected network contains vertices such that any two of them are reachable from each other, then, the sub-graph is named a *connected component*. The size of a connected component is given by its number of vertices N_{cc}. A giant component exists when a connected component is much larger than the other connected components in the network. A related concept is the *strongly connected component*, which means that there is always a path between any two vertices in the respective sub-graph of a directed network.

Betweenness

One can measure the centrality of a vertex h by counting the number of shortest paths passing through this single vertex, i.e. $|\sigma_{ij}(h)|$. The potential of vertex h to intermediate information passing between vertices i and j can be defined as the probability that h falls on a randomly selected geodesic between these vertices. Therefore, considering all pairs of vertices, one gets eqn. 4.2 [83].

$$B'(h) \equiv \sum_{i,j:i<j \text{ and } i \neq j \neq h}^{N} \frac{|\sigma_{ij}(h)|}{|\sigma_{ij}|} \tag{4.2}$$

The measure was first proposed by Freeman in the context of social networks and is frequently referred to as *Freeman betweenness centrality* to distinguish among other centrality metrics [83]. This measure can be further normalized by the total number of possible paths passing through h between any two vertices in the network (eqn. 4.3).

$$B(h) = \frac{B'(h)}{(N-1)(N-2)} \tag{4.3}$$

A related concept to the Freeman betweenness is the edge-betweenness. In that case, one counts the geodesics passing through an edge rather than through a vertex [84].

Degree and Strength

The *neighbors* n_i of a vertex i are those vertices at distance one from i. The number of edges between neighbors of i is given by e_i. Higher order neighborhoods can be defined and are frequently called hierarchies [85]. The *degree* k_i of a vertex is a measure of centrality and is given by its number of neighbors (eqn. 4.4), where a_{ij} is the respective entry in the adjacency matrix (see §2.1.2). Vertices with degree much higher than the average network degree are called *hubs*.

$$k_i \equiv \sum_{i,j}^{N} a_{ij} \tag{4.4}$$

In a directed network, in- (k_i^{in}) and out-degrees (k_i^{out}) are defined, respectively, to edges pointing inwards and outwards to the vertex. In weighted networks, one counts all weights w_{ij} of the edges connecting to vertex i to obtain its *strength* s_i (eqn. 4.5).

$$s_i \equiv \sum_{i,j}^{N} a_{ij} w_{ij} \tag{4.5}$$

k-cores

A connected component containing vertices with at least degree k is called k-*core* of a network. A simple procedure to obtain this sub-network is to start with the original network and iteratively remove all vertices of degree less than k. The resulting sub-network(s) has(have) all vertices with degree higher or equal k and provides a measure of network cohesion [86].

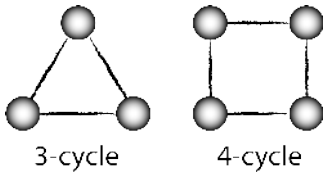

Figure 4.1: *Illustration of the shortest cycles in simple and bipartite networks.*

Clustering and cycles

The term cluster is used to refer to a group of vertices more connected between themselves than with vertices outside the group. There are different ways of measuring clustering, the choice depends on which type of clustering one wants to quantify. The most popular methods include counting the number of short cycles or detecting the community structure of a network.

Considering only the closest neighbors of a vertex, the local *clustering coefficient* cc_i (or transitivity in social networks) can be defined by counting the proportion of actual edges between common neighbors of a vertex i relative to the number of possible edges between them (eqn. 4.6).

$$cc_i \equiv \frac{2e_i}{n_i(n_i - 1)} \tag{4.6}$$

Another common method is to count the number of cycles. A d-cycle is defined as a closed path σ_{ii} of size σ_{ii}^d, that starts and ends at the same vertex but does not pass more than once through the same intermediate vertices. The smallest informative cycle depends on the type of network. A cycle of size 2 does not bring more information than the edge itself. Cycles of size 3, in simple networks, quantify the connectivity between common neighbors of a vertex. In bipartite networks, in turn, the smallest cycle has size 4 and shows how likely two vertices of one type have two vertices of another type in common (Fig. 4.1).

Communities

Communities, modules, or clusters[a] are different names typically used to describe the same property, a group of vertices sharing some common characteristics or playing similar roles in the network [87]. There are several ways of defining and detecting communities using networks. The interested reader should look at ref. [88] for a thorough review of the state-of-the-art community detection algorithms and methods.

[a]Sometimes, the word cluster is avoided in the context of communities, where the word modules are preferred.

The most naïve method to group similar vertices is to simply look on the dataset and separate vertices according to intrinsic properties (when available) of them, e.g. male and female vertices. Nevertheless, one typically wants to automatically detect hidden communities (unknown from visual data inspection) by using network structure. One idea is to calculate the structural similarity between vertices, and afterwards, organize them hierarchically according to the level of similarity. Such approach is used, for instance, in paper III and briefly described in §4.2.

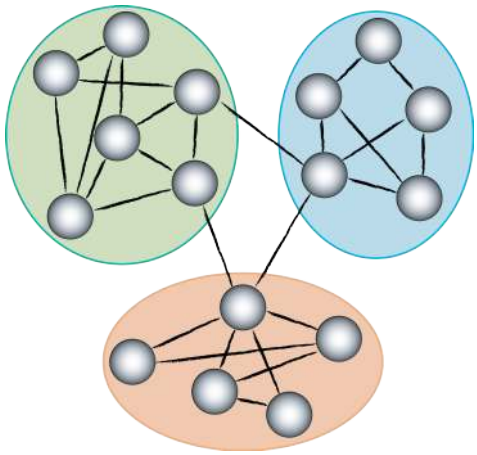

Figure 4.2: *Pictorial network with three communities containing* 6 *(right),* 5 *(left), and* 5 *(bottom) vertices. Each pair of communities is connected by one edge.*

The majority of methods, however, refers to communities as groups of vertices more interconnected between themselves than with other vertices (Fig. 4.2). This definition is as lousy as the definition of hubs, because there is no universally accepted way (threshold or other quantity) to define the border distinguishing between communities. One common method, however, to quantify the over-connectivity of vertices is through *modularity*, i.e. by comparing the density of edges in a subnet of the sampled network, with a subnet (with the same set of vertices) where the edges are randomly connected (null network). If a null network is defined by randomizing the edges but maintaining the vertex-degree fixed, the modularity is written as eqn. 4.7, where g_i corresponds to the group of i and $\delta(g_i, g_j)$ is the Kronecker delta. The goal then is to fragment the network into subnets such that Q is maximized. One such method, important for historical reasons, is to split the network by iteratively removing edges with high edge-betweenness [84].

$$Q = \frac{1}{2E} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2E} \right) \delta(g_i, g_j) \tag{4.7}$$

A more efficient method (at least on benchmark networks [89]) is to let a random walker explores the network and then measure the probability of jumps between

vertices. In principle, highly interconnected vertices tend to trap the walker such that it stays longer within that group (community) rather than moving between less connected vertices [90].

Assortativity

Whether similar vertices are connected or not is a fundamental question. Given any vertex attribute [16], one can measure the level of assortativity (likeness) of the network vertices. In networks, a feature of interest is the degree. If vertices of similar degree are preferably connected between themselves, the network is said to be assortative. Conversely, if the connections are preferably between vertices of opposite degrees, the network is disassortative. This network property can be quantified by measuring the *coefficient of assortativity* r (eqn. 4.8), which ranges between -1 (disassortative) and $+1$ (assortative). Note that δ_{ij} is the Kronecker delta.

$$r \equiv \frac{\sum_{ij}(a_{ij} - k_i k_j/2E)k_i k_j}{\sum_{ij}(k_i\delta_{ij} - k_i k_j/2E)k_i k_j} \tag{4.8}$$

It is also possible to calculate the correlation between the degrees of the vertices on both sides of an edge. For bipartite networks with vertices of types A and B, this correlation is given by eqn. 4.9. The averages in this equation are taken by summing over the different edges.

$$r \equiv \frac{\langle k_A k_B\rangle - \langle k_A\rangle\langle k_B\rangle}{\sqrt{\langle k_A^2\rangle - \langle k_A\rangle^2}\sqrt{\langle k_B^2\rangle - \langle k_B\rangle^2}} \tag{4.9}$$

Reciprocity

In directed networks, a vertex i pointing to a vertex j does not imply that the reverse is also true. This reciprocity is important to measure the symmetry of the relations. The symmetry of a single pair of vertices can be measured by a simple binary yes or no, but for the entire network, the normalized reciprocity (in respect to a random network) is given by eqn. 4.10 [91].

$$R \equiv \frac{\sum_{i\neq j}^{N}(a_{ij} - \overline{a})(a_{ji} - \overline{a})}{\sum_{i\neq j}^{N}(a_{ij} - \overline{a})^2} \qquad \text{where: } \overline{a} \equiv \frac{\sum_{i\neq j}^{N} a_{ij}}{N(N-1)} \tag{4.10}$$

Histograms

A histogram is a function of the number of occurrences of some quantity x (e.g. the degree) within a range of values. If the histogram is normalized by the size of the sample (total number of occurrences), then it transforms into the frequency or

probability distribution $p(x)$. The shape and exponents of this distribution permit to identify universality classes of networks. Random networks, for instance, have Poissonian degree distributions (eqn. 4.11a) in the limit of $N \gg E$, while the generative preferential attachment model results in networks with power-law degree distributions (eqn. 4.11b)[b].

$$p(k) = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle} \qquad (a) \qquad\qquad p(k) \propto k^{-\alpha} \qquad (b) \tag{4.11}$$

Therefore, rather than looking to summary statistics (e.g. mean and variance) of some quantity, plotting the histograms can be quite informative to identify patterns in the data. Since noisy distributions are specially common in small networks, it turns to be difficult to infer the functional form of the distribution even by using advanced statistical methods [92]. A common approach (by physicists) is to simply plot the distribution of the desired measure and fit the curve using a least-square method [50]. To reduce the noise before the fitting, typically the cumulative distribution, $P(x* \geq x) = \sum_{x=x*}^{x_{max}} p(x)$ is considered. Another method sometimes adopted is to use logarithmic data binning; in this case, the histogram is binned on increasing powers of 2 such that the binned values appear equally spaced in a logarithmic plot (convenient to visualize power-law distributions).

Entropy

There are several statistics to reduce the information of the distribution function into a single number; the most commons are the first and second moments, known as the mean and variance. Another statistic is the entropy of the distribution (eqn. 4.12). The entropy is a measure of disorder and it has been adopted in different contexts (see e.g. [93]). Here, entropy is used as a simple measure of heterogeneity of a degree distribution. In the limit of no heterogeneity, i.e. if the degree distribution has only one value, $p(x = x_1) = 1$ and $p(x \neq x_1) = 0$, then $S_{\text{degree}} = 0$. On the other hand, S increases for increasing heterogeneity of the distribution.

$$S_{\text{degree}} \equiv -\sum_{k=1}^{k_{max}} p_{\text{degree}}(k) \log_\epsilon p_{\text{degree}}(k) \tag{4.12}$$

Minimum spanning tree

A *tree* is a connected graph with no cycles in the structure. The number of edges in the tree is thus equal to the number of vertices minus one. A star-like and line graphs are limiting cases of tree structures. A *spanning tree* is a tree containing all vertices of a graph and some of its vertices, therefore, any connected network,

[b] Although k is a discrete quantity, the approximation of k to a continuous variable is acceptable in this context.

weighted or not, can have multiple spanning trees. The *minimum spanning tree* (MST) of a weighted network contains the minimum set of edges necessary to connect all vertices such that the sum of the edge-weights is minimal. There are several algorithms to extract this tree from a network. A relatively easy to understand is the solution proposed by Prim [94]:

Start with $V(Tree) = \{i\}$*, where the vertex* i *is an arbitrary starting vertex of the network* Γ *and* $E(Tree) = \{\emptyset\}$

1. *Until* $V(Tree) = V(\Gamma)$*, repeat:*
2. *Choose an edge* (i, j) *with minimal weight such that* i *is in* $V(Tree)$ *and* j *is not*
   $\Rightarrow$*If there are multiple edges with the same weight, pick any*
3. *Add* j *to* $V(Tree)$*, and* (i, j) *to* $E(Tree)$

This apparent mathematical amusement, the minimum spanning tree, can indeed be useful in applied network problems. Suppose, for instance, that protesters are concentrated in one point and they want to take over the central part of a city and the plan is to have all crossing streets under control at the shortest time. Suppose also, that each street has a different length and that they can only move in a group, following the streets (the actual network) at the same speed [c]. The minimum spanning tree therefore provides the minimum total path that the group should follow to reach all points in the shortest time. Note that there might exist a shortest path between two crossing points that is not included in the final outcome, but the MST only takes the global optimal solution.

Summary of paper I

In the early 2000s in Brazil, news about delays and chaos in the domestic airports were increasingly being reported by the media. After becoming a network enthusiast, I wondered how much of that was due to the network of flights and its continuous changes. In the paper entitled "Structural evolution of the Brazilian airport network", I construct annual networks of domestic regular flights between airports in Brazil and study how local and global structures change during a period of 12 years. I find that, in terms of network quantities such as degree, strength, and betweenness, vertices change their relative rank and absolute values considerably throughout the years. Thus, some airports become more central than others. Surprisingly though, the degree and betweenness distributions are described by the same

[c] For outsiders, this example probably turned to be more like a physicist amusement than a proper illustration of an application.

functional form irrespectively of the year, but with different parameter values. In other words, while local structures change significantly, the emergent network structure apparently continues to follow the same organizational principles. At a global scale, an optimization process seems to take place, with the number of routes (or edges) decreasing while the traffic, passengers and cargo carried (or weight) more than doubled in the 12-year period. Profitable routes are maintained in contrast to the less popular ones that are simply removed.

4.2 Multivariate analysis

To analyze multiple variables simultaneously one uses multivariate methods. These methods are particularly interesting for network science because, due to construction, networks naturally have multiple degrees of freedom that can be used as different input variables for the analysis. Nevertheless, though popular in different fields, they have not been applied much to study network structure [95, 96]. In the following sections, the concepts of feature vector and vector similarity are presented in the context of networks, and then, a multivariate method, namely principal component analysis, is introduced applied to network analysis.

Feature vector

Any object can be described by a vector of features $\mathbf{v}$, i.e. a vector where each entry contains a value corresponding to a different attribute of the object. In case of networks, each vector $\mathbf{v}_i$ corresponds to one vertex i, and these properties can be either topological quantities as those described in §4.1 (e.g. degree, betweenness) or intrinsic properties known from the dataset (e.g. the population size or income per-capita, in case the vertex represents a city).

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 5 \\ 0 \end{bmatrix} \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 8 \end{bmatrix} \mathbf{v}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 5 \\ 0 \\ 2 \end{bmatrix} \mathbf{v}_4 = \begin{bmatrix} 1 \\ 0 \\ 5 \\ 0 \\ 3 \\ 0 \end{bmatrix} \mathbf{v}_5 = \begin{bmatrix} 5 \\ 0 \\ 0 \\ 3 \\ 0 \\ 7 \end{bmatrix} \mathbf{v}_6 = \begin{bmatrix} 0 \\ 8 \\ 2 \\ 1 \\ 7 \\ 0 \end{bmatrix} \tag{4.13}$$

Going for the network structure, one simple idea is to use the vertex neighbors and the respective edge-weights as the features of the vector. Considering the entire network with N vertices, a vector with N-dimensions is created for each single vertex i such that each of these dimensions corresponds to one other vertex j[d]. In other

[d]Note that according to this general definition, self-loops, i.e. edges connecting the vertex with itself, are accepted.

words, each entry j in the vector i corresponds to the edge (i, j); for example, the network of Fig. 2.4 are represented by a set of vectors in eqn. 4.13:

Vector similarity

Since the network features are mapped into the vector by real numbers, the similarity of these vectors can be quantified by different methods. When the data are dense or continuous, a good choice is to simply measure the Euclidian distance between the two vectors; if the data is linearly correlated, then, the Pearson correlation is an appropriate choice. A general suitable method for sparse non-linear data is the cosine similarity. The cosine similarity between vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ is given by eqn. 4.14, where $|\cdot|$ is the magnitude of the vector. The cosine similarity captures the trend and disregards the magnitude of the vectors; as a result, it provides a similarity scale ranging from -1 (least similar) to 1 (most similar).

$$\text{cosine similarity} = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{|\mathbf{v}_i|\,|\mathbf{v}_j|}. \tag{4.14}$$

Principal component analysis

The vectors can be correlated in this multi-dimensional space. If some dimensions are correlated, it means that a variation in one dimension is equivalent to a variation in another dimension. Therefore, if the correlation is strong, only one of the dimensions suffices to describe the data points, and the other can be discarded without loosing much information but with the benefit of a simpler representation (the use of less dimensions).

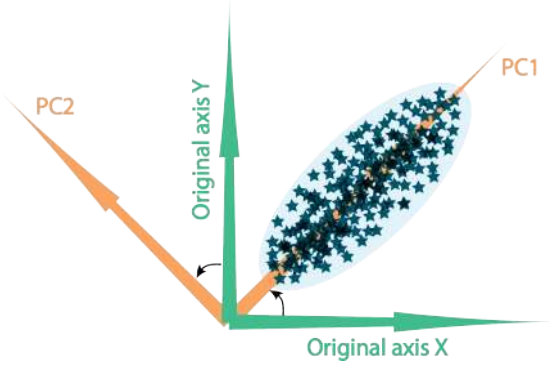

Figure 4.3: *Example of rotation to reduce the dimensionality of data. Due to large correlation, the initial two-dimensional data points can be fairly explained by using one-dimensional projection in the principal component 1 (PC1).*

The multivariate method *principal component analysis* (PCA) is a general method for this type of decorrelation and data reduction [97]. In the PCA, the axes in the

original N-dimensional space are rotated to point towards the maximum variability of data. The rotation matrix is the covariance matrix of the different dimensions and the eigenvalues of this matrix provide a scale of data variability in the directions corresponding to the eigenvectors (named *principal components*, PCs) of the rotation matrix[e]. Applying this rotation, the original set of points are projected into the new N-dimensional space (Fig. 4.3). Consequently, PCs corresponding to small variability (small eigenvalues) and thus, little associated information, can be discarded without loss of relevant information.

Before performing the PCA, the data points are replaced by their Z-scores such that all dimensions have zero mean and standard deviation equal to one. By doing this standardization, the sum of all eigenvalues become equal to the number of points, which corresponds to maximum variability. This procedure highlights the variability contribution of the new axes (PCs) after performing the rotation, where all eigenvalues larger than one provide more information than any of the original axes. The contribution of each eigenvalue in the total variability is obtained by dividing its value by the sum of all eigenvalues. The optimal number of PCs is chosen by taking, for instance, the eigenvalues explaining a certain percentage of the variability, say 95%, and discarding the rest. Another common methodology is to simply take the two highest eigenvalues and project the data into their respective PCs.

Summary of paper II

I frequently get this feeling that I am the only person unsatisfied with a certain company, since others are enthusiastically typing their credit card pin-codes around. Rather than complaining, I used to simply try the company next door, but I quickly realized that it does not help much on many occasions. In "The network organisation of consumer complaints" I collect data about registered consumer complaints about Brazilian companies and develop, together with Petter, a method to detect relations between these companies. We extend the idea of feature vectors presented in the thesis and apply to bipartite networks. Since a vertex of one type only connects to vertices of another type, the vectors represent vertices of type A while the dimensions represent the vertices of the opposite type. With our data, each company vertex has a feature vector in which we associate a category of complaint to each dimension. According to the network terminology, we use information about contacts between the vertex (company) and its first neighborhood (the complaints). Afterwards, we measure the vector similarity between pairs of vertices (i.e. pairs of companies)

[e]This method is also known as singular value decomposition.

to obtain a new fully connected network where the edge-weights correspond to the similarity level between the companies. Then, we extract the minimum spanning tree of the new network and compare (with meta-data from the dataset) if same sector companies are clustered together (in branches of the tree). Since in this multidimensional space there are lots of correlations, we perform the principal component analysis to remove them and are able to identify that the best clustering of same sector companies is obtained by using only 27 *of the principal components (out of the initial* 8827 *dimensions). Furthermore, analyzing the bipartite network structure, we identify that companies indeed fix the complaints. The fraction of resolved complaints is independent but the number of complaints increases sub-linearly with the company's market value. Another interesting finding is that the complaint categories appear to be organized in a hierarchical way. Companies only get complaints of lower degree if they have already got complaints of higher degree, in other words, when the company receives few complaints, these complaints are common across the system (i.e. received by other companies, and thus classified as high-degree complaints). Nonetheless, companies with large number of complaints necessarily also get those unusual (specific) complaints which are shared by few companies.*

4.3 Dynamic

Research on networks has been devoted to quantify the static structure. Measuring the dynamic changes is overall a relatively recent subject and consequently still lacks an abundant menu of metrics. The traditional methodology has been to use the static measures discussed in §4.1 to characterize the structures at different instants in time and then analyze their temporal evolution [98]. Recently though, some researchers are redefining static measures taking into account the temporal order of the connections such that the edge exists only at certain moments in time and is absent otherwise [99, 100, 101, 102]. In this section, some methods to characterize the degree evolution are reviewed.

Time series

To study discrete time dependent data, one has to use a time series. A time series $g(t)$ is a sequence of values a_t displaced at regular intervals in time (where $a_t \in \mathbb{R}$ and t is a non-negative integer). These values can be related to any structural or non-structural network quantities. The time series formalism provides a way to organize and analyze how such quantities change in time.

Fourier analysis

The analysis of a time series can reveal several regularities on the data. One such pattern is the cyclic nature of several phenomena, for example, due to daily (circadian) or weekly rhythms. Using the time series formality, temporal cycles are regular repetitions of values in the sequence a_t such that $a_t = a_{t+\Delta T}$, where ΔT is a fixed interval of time, e.g. seven days.

A classic method to identify cycles is by using Fourier analysis. Fourier analysis is based in the Fourier transform (eqn. 4.15) which is a method to convert a function in the time (or space) domain to frequency or other reciprocal space. Performing a Fourier analysis means that one assumes that the original time series is periodic and can be approximated by a sum of sines of different frequencies and amplitudes [f]. Therefore, the Fourier transform provides the spectrum of frequencies corresponding to regular oscillations in the time series. This method is mostly adequate to identify *a priori* sinusoidal signals, but more rectilinear shapes are reproduced by using the fundamental frequency and its high order harmonics. The fundamental frequency sets the length of the cycle (e.g. seven days) and the higher harmonics the shape of the cycle (e.g. square-like).

$$G(f) = \sum_{t=0}^{T-1} g(t) \exp\left(-i2\pi \frac{f}{T} t\right) \tag{4.15}$$

Detrended fluctuation analysis

Correlation functions can be used to detect memory effects in time series. In special, the autocorrelation measures the similarity of the time series with its own temporal shifted version. If the autocorrelation function decays exponentially with the time lag, the series has short-range memory. No memory is characterized by zero autocorrelation for any time lag and is equivalent to a random walk process. The more conspicuous scenario happens when the autocorrelation function decays as a power-law, meaning that the series has long-range correlations (in simple words, it means that even for a large time-lag, the series has a degree of similarity with the original series).

One way to measure these correlations is by measuring the self-similarity of the integrated version of the original signal (with mean zero). Self-similarity means that the signal seems similar after a re-escaling operation, i.e. the variance grows linearly with the length of the signal [103, 104]. The most popular and efficient method to quantify the self-similarity is the *detrended fluctuation analysis*. In this method, the original signal is separated into $N_{\max}$ boxes of size L each, and local trends are

[f] Roughly speaking, it is an approximation method in the same spirit of expanding any function in a Taylor series around a point, i.e. one assumes that a function can be approximately described by a sum of polynomials at that point.

removed (by subtracting from the signal its nth-order fit, $y_{\text{fit}}(t)$) before the variance is calculated (eqn. 4.16). The process is repeated for boxes of different sizes and a line is obtained in a log-log plot of $F(L)$. There is a direct correspondence of the scaling exponent α with autocorrelation properties, if $\alpha = 0.5$, there is none or short-range correlations in the series, while $0.5 < \alpha < 1$ indicates long-range correlations. The DNA sequences [103] and heart beat of health subjects [105] are classic examples of signals with long-range correlations; recent applications include observation of long-range correlations in email communication [104].

$$F(L) = \sqrt{\frac{1}{N_{\max}} \sum_{t=1}^{N_{\max}} [y(t) - y_{\text{fit}}(t)]^2} \tag{4.16}$$

Preferential attachment

The mechanism of preferential attachment leads to a network with power-law degree distribution (§2.2.2). One can go into the other direction and, for a given network, quantify the intensity of the preferential attachment. This can be done by measuring the probability (at a certain time t) of a vertex with degree $k_i(t)$ to get a new edge at the time $t+1$ [106]. By fitting δ in eqn. 4.17 with maximum-likelihood estimates for all vertices, one can infer the scaling of the growth, linear ($\delta = 1$), sub- ($\delta < 1$) or super-linear ($\delta > 1$).

$$Prob[k_i(t+1) = k_i(t) + 1] = \frac{k_i(t)^\delta}{\sum_j k_j(t)^\delta} \tag{4.17}$$

According to the Barabási-Albert model, linear preferential attachment creates networks with scale-free degree distributions with exponent $\alpha = 3$ in eqn. 4.11b [15]. Sub-linear preferential attachment, however, results in networks which degree distributions are better described by stretched exponentials [107], as in eqn. 4.18.

$$p(k) \propto \exp(-k/k_0)^\alpha \tag{4.18}$$

Summary of paper III

I do like polemics and prostitution is certainly a topic that provokes heated debates with strong positions on both sides. But rather than taking a judgmental attitude, as a physicist, I decided to search for data to understand a little the dynamics of the activity. I crawled the web and found a forum where sex-buyers review their encounters with sex-sellers. In the paper "Information dynamics shape the sexual networks of Internet-mediated prostitution", Petter and I asked the help of Fredrik to study the dynamics of the forum and how it was connected with real-life commercial activity.

We focus on the section about members who had sexual encounters with escorts, an expensive form of prostitution. Analyzing the temporal profile of the reviews, we notice that well graded escorts have a higher tendency to attract new customers, which in turn, write good reviews if satisfied with the encounter. Contrary, unsatisfactory encounters are reflected into bad reviews, which decrease the number of future customers. This feedback between on and offline activity indicates the importance of the forum for this type of business. A remarkable finding is that, both escorts and buyers get into more risky behavior with experience. Although the offer of multiple risky services benefits the escort by increasing its popularity and grades, it also brings higher chances of getting an infectious disease. The activity in the forum follows weekly cycles but exhibits broad distribution of inter-contact time at a day-level, which means that an encounter may trigger subsequent encounters with other members. In fact, we measure a linear preferential attachment for short intervals, while slightly sub-linear scaling is observed for longer time intervals. By connecting a member reviewing a certain escort, we naturally obtain a sexual network. Thanks to the popularity of the web site, the network is large with 16,730 vertices and 50,185 contacts. Surprisingly though, this sexual network is connected even though spanning over 12 cities far apart in Brazil (at least 400 *km). Nevertheless, city boundaries create clustered structures reflected in the large number of 4-cycle and in the large diameter relative to the respective random networks. Contacts between cities follow the inverse-square law observed in trading and communication patterns. Within a city, the number of sellers scales sub-linearly with city population suggesting that this type of prostitution does not benefit as much from an increasing concentration of people.*

Chapter 5

Dynamics on networks

In many systems, the network is the actual underlying structure where transmission processes occur. The network of roads, for instance, defines the possible pathways for vehicles to move around. An elegant example is the network of neurons in the brain, where the neuronal cells (neurons) connected by synapses define, respectively, the vertices and edges of the neuronal network; electric and chemical impulses in turn travel between cells creating what is usually named consciousness [108]. There are cases though, that the network structure is defined by the actual dynamical process, for instance, the propagation of Internet chain letters creates a structure of letter-senders connected to letter-receivers [109].

In this section, the focus is in the first scenario, i.e. in the impact of underlying topological structures in dynamical processes taking place on top of, or simply, on these network structures. Understanding how quantities propagate within a network is fundamental to create methods to improve (e.g. in marketing) or to reduce (e.g. in case of disease) the spreading.

5.1 Disease spreading

The spread of infectious diseases (e.g. HIV, Influenza) is a continuous challenge of public health. A significant large number of people is infected every year irrespective of age, sex or income [110]. These infections spread through pathogens by direct physical contact or close proximity, body fluids, vector organisms, among other means. To reduce the impact of these diseases on the population, researchers focus on either improving patient treatment (e.g. developing new medicines) or on control of epidemics[a] (e.g. by vaccination or informational campaigns). However, to control an epidemic, one has to first understand how these diseases actually spread throughout the population.

[a]Epidemics are essentially a regime where a more than usual number of individuals is infected (where usual depends on the disease and period). This topic is discussed in details in §5.2.

5.1.1 Homogeneous networks

Mathematical models have been applied in epidemiology as early as the 18th century. Daniel Bernoulli, a well-known figure for physicists [b], proposed an age-dependent model to study the gain in life-expectancy if small-pox was eliminated as a cause of death. A readable review of his original article is available in ref. [111]. A particular class of mathematical models, much used nowadays, are the compartmental models, which roots date back to the work of Hamer [112] in 1906; he proposed that the population dynamics of a disease is proportional to the probability of one individual being infected, the probability of another being susceptible, and the number of contacts between them per unit time (contact-rate). In a modern formulation, the idea is to divide the population into compartments containing individuals at different stages of the infection dynamics and allowing them to move only between some of the compartments. The basic assumption is that people within a compartment are well mixed; this means that the population of a compartment can be described by a density function. Therefore, the chance of a contact between individuals of different types, i.e. the chance of an infection event, is proportional to the respective density of infected (I) and susceptible (S) individuals, and happens with a certain probability ρ [58, 60].

It is a common practice to choose simple disease models with as few compartments as possible but still capturing the main properties one wants to study. The SI (Susceptible-Infected) model is the simplest, yet is adequate to study the early phase of an outbreak over shorter time scales than the duration of the disease. This model also gives an upper bound of the number of reachable individuals, since once infected the individual continues in this state. The mathematical framework of compartmental models is general enough to permit more realistic and detailed extensions, including different stages to mimic specific diseases. A classic and perhaps the most important example is the Kermack-McKendrick model. In this model, also known as SIR (Susceptible-Infected-Removed), individuals obey the SI dynamics, but once in the infected state, they move to a removed or recovered compartment (with density R) with probability μ, where $1/\mu = T_{\text{infection}}$, the period of infection [113].

$$\begin{aligned} \frac{dS}{dt} &= -\rho SI \\ \frac{dI}{dt} &= \rho SI - \mu I \\ \frac{dR}{dt} &= \mu I \end{aligned}$$

Since S, I and R represent densities, the sum over all compartments, $N =$

[b] He proposed the now famous Bernoulli's principle, stating that the pressure of a fluid decreases with the increase in the speed of a fluid.

$S+I+R$, is equal to one. Figure 5.1 shows the evolution of the number of individuals in each compartment for $\rho = 0.01$ and $\mu = 0.1$ and initial conditions $S(0) = 0.99$, $I(0) = 0.01$ and $R(0) = 0$. The density of infected individuals during a period of time is defined as the *prevalence* of the infection. The number of new cases of infection, on the other hand, is called *incidence*.

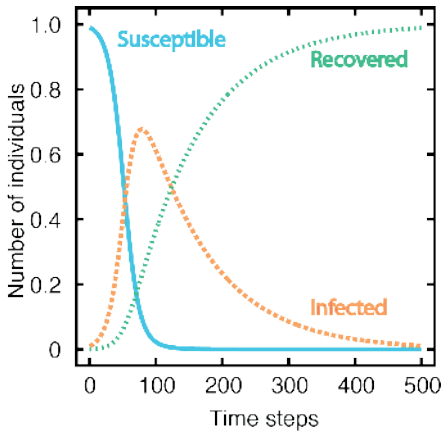

Figure 5.1: *Evolution of the density of individuals in each stage of SIR disease spread model assuming well-mixed conditions and* $\rho = 0.01$ *and* $\mu = 0.1$.

The equation describing the variation of the number of infected individuals has a peak at $dI/dt = 0$ (eqn. 5.1). This means that the number of infected individuals may increase $dI/dt > 0$ or decrease $dI/dt < 0$ according to the parameters ρ and μ.

$$\rho SI - \mu I = 0 \qquad \Rightarrow \qquad \frac{\rho SI}{\mu} = I \tag{5.1}$$

For a susceptible population with initial density of infected individuals $I(0) \ll S(0)$, the approximation $S(0) \approx 1$ is valid and the ratio between infection and recovery rates can be defined as:

$$R_0 = \frac{\rho}{\mu} = 1 \tag{5.2}$$

Therefore, if $R_0 > 1$, the number of secondary infections (number of individuals infected by an initial source) has a non-zero probability of growing and the infection can reach a large fraction of the population. Conversely, if $R_0 < 1$, the density of infected individuals vanishes exponentially. Note, however, that its validity is restricted to the early disease outbreak (due to the chosen approximations). It turns out that this quantity R_0, the *reproduction rate*, is of fundamental importance for modern epidemiology and typically take different forms for each disease spreading model [58, 60]. Finally, for a disease transmitted through contacts, the *endemic*

steady-state is given by $R_0 \mathrm{x} S = 1$, which means that the disease survives within a population without external inputs.

5.1.2 Heterogeneous networks

The homogeneous assumption (well-mixed population) is a simplified approach that waves the fact that individuals interact following highly heterogeneous contact patterns. To overcome these limitation, population structure has been introduced in these models to account for different characteristics of subgroups of individuals, for example, age structure[114] and risk behavior [115]. Another variation includes models of metapopulation, where individuals are well-mixed within independent groups and the interaction happens through the movement of individuals between the groups [116, 117].

At an individual level, heterogeneity is introduced by using contact networks [63, 118, 25]. The first attempt to model (a simple) disease spreading considering the (heterogeneous) network structure is perhaps the study of site percolation in a one dimensional small-world network model performed by Newman and Watts [119]. Noting that several empirical networks (e.g. world wide web or sexual networks) relevant to the spread of virus/infections follow power-law like degree distributions, Pastor-Satorras and Vespignani showed, by using a mean-field analysis of the SIR model (eqn. 5.3, where k is the degree and p_k is the degree distribution), that scale-free networks with exponent $2 < \alpha \leq 3$ lack epidemic threshold, meaning that any small number of initially infected individuals can infect the whole network [120].

$$\frac{dI_k}{dt} = -\mu I_k + \rho k(1 - I_k)\Phi(\lambda) \qquad \text{where:} \qquad \Phi(\lambda) = \sum_k \frac{k p_k I_k}{\sum_s s p_s} \tag{5.3}$$

The properties of the SIR model in static random networks are nowadays much studied and several properties are derived analytically, as for example, the critical epidemic threshold (eqn. 5.4) and the average outbreak size [121].

$$\rho_{\text{critical}} = \frac{\sum_k k p_k}{\sum_k k(k-1) p_k} \tag{5.4}$$

It is generally hard to write down equations for disease spreading in networks with more complex structures than random contact patterns, for example, in case of degree-degree correlations, community structure [122], or existence of connected components. One way to overcome these difficulties is to study the system by numerical simulations. Numerical simulations are also used to validate analytical predictions [120, 121] and to study disease spreading directly in the empirical network [49]. The general framework is to initially let all vertices susceptible, except by one random vertex chosen as a source of infection. At each time step, an infected vertex transmits the infection to all of its contacts with probability ρ. This probability includes the per-contact infection probability $\rho_{\text{infection}}$ and the actual probability that

the contact exists at that time p_{contact}. Typically $p_{\text{contact}} = 1$ and thus $\rho = \rho_{\text{infection}}$; but the time scale can be renormalized by using $p_{\text{contact}} = C/ET \neq 1$, where C and E are respectively the total number of contacts and the total number of unique pairs in the network, and T is the time window.

5.1.3 Dynamic networks

The static network is adequate to represent contact structures that change more slowly than the actual dynamical process taking place on the network [43, 120]. In case of disease spreading, it is assumed that the static representation is a good approximation for rapidly spreading pathogens that can cause acute infection [123]. Nevertheless, although human daily movement, for example, seems to have a high potential of predictability [124], contact patterns are dynamic and depending on the time scale, the emerging contact structure can be quite different (ruling out the static approximation) with certain individuals being more active every other time [125, 126, 25].

Temporal information can be included in different forms as edge-weights, usually by aggregating all contacts during a time interval [125, 127], or for example, by counting the number of consecutive times that two individuals stayed in contact [127]. Other temporal correlations, like concurrency [128][c], the order of contacts [64, 98, 49], or seasonal effects [130], may affect disease spread as well. These effects are lost in the static representation, unless line graphs are used [131], but then, the network structure is removed, which is not desirable either.

During recent years, different mathematical models have been proposed to study some consequences of network structure co-evolving with disease spreading [43, 123, 132, 133]. The model of Volz and Meyers, for instance, assumes that vertices change partners with a certain rate but the number of partners is maintained fixed throughout time. The rate of changing partners ϕ provides intermediate scenarios between the border cases of static ($\phi = 0$) and well-mixed ($\phi = 1$) interaction; one observation is that the static approximation is valid when ϕ is small [43]. The complexity of analytical models aimed to capture both network structure and dynamic evolution can be appreciated in the set of eqns. 5.5, which corresponds to a general model devised by Kamp [132]. In this model, for SIR epidemics, vertices born (at rate ν_1) and die (at rate ν_2), and it is implicitly assumed that new vertices connect randomly to vertices already in the system[d]; preferential attachment and cluster structures can be included after some mathematical endeavor [132].

[c] Although under debate [129], concurrency is usually claimed as a key player in the high prevalence of HIV in e.g. sub-Saharan Africa.

[d] In this set of coupled differential equations, $\bar{p}_k$ is the probability of an individual entering the population to have k contacts; $\bar{g}$ is the probability generating function of $\bar{p}_k$; and p_{AB} is the probability of an edge to point from vertex A to vertex B. See more details in ref. [132]

$$\begin{aligned}
\frac{dS_k}{dt} &= -\rho p_{\mathrm{SI}} k S_k && \text{new infections} \\
&\quad +\eta_1 N \bar{p}_k - \eta_2 S_k && \text{natural birth and death} \\
&\quad +\eta_1 \bar{g}'(1,t)(S_{k-1} - S_k) && \text{contacts made with new vertices} \\
&\quad -\eta_2[kS_k - (k+1)S_{k+1}] && \text{conts. lost from dying vertices} \\
&\quad -\mu p_{\mathrm{SI}}[kS_k - (k+1)S_{k+1}] && \text{conts. lost from vertices dying from infection}
\end{aligned} \tag{5.5}$$

$$\begin{aligned}
\frac{dI_k}{dt} &= +\rho p_{\mathrm{SI}} k S_k && \text{new infections} \\
&\quad -(\eta_2 + \mu) I_k && \text{death} \\
&\quad +\eta_1 \bar{g}'(1,t)(I_{k-1} - I_k) && \text{contacts made with new vertices} \\
&\quad -\eta_2[kI_k - (k+1)I_{k+1}] && \text{conts. lost from dying vertices} \\
&\quad -\mu p_{\mathrm{II}}[kI_k - (k+1)I_{k+1}] && \text{conts. lost from vertices dying from infection}
\end{aligned}$$

5.1.4 Temporal networks

Research on dynamical processes in temporal networks is quite recent and has been essentially based on numerical simulations. One way to simulate the disease spreading models in a temporal network is to first map the original network into a time-ordered list, as discussed in §2.1.2. After selecting a (single or multiple) source of infection, a pointer goes through the ordered list updating the state of the vertices according to the chosen model of disease spreading. The states can be updated at each time step or at the available temporal resolution, for instance, after a day or an hour. If opting for the first updating method, the order of the contacts within a time unit should be randomized and averages taken to remove the bias (due to the construction of the ordered list) of the order of contacts within a day (or within an hour). Since real datasets usually have network structure spanning within a limited time interval, boundary conditions have to be defined when the pointer reaches the last contact (t_{F}). One boundary condition is to simply finish the simulation at this point, another is to adopt periodic boundary conditions which consists in the repetition of the network after the end of the interval, i.e. the network at $t_{\mathrm{F}} + 1 + \Delta T$ is the same as the network at $t_0 + \Delta T$[e]. An algorithm to simulate different disease spreading models (SI, SIR and $\mathrm{SI_1I_2}$, the last containing two periods of different infectivity, ρ_1 and ρ_2) in temporal networks is presented bellow:

[e]This periodic boundary condition is equivalent to transforming any time-series into a periodic series. In this case, the network is replicated periodically.

1. *Take one sample of the network (e.g. with X days)*
2. *Select the infection source in the sample*
3. *Go through the ordered list*
   - 3.1 *If vertex i is infective*
     - 3.1.1 *Infects its contact j at that day with probability $\rho_1 = \rho$*
     - 3.1.2 *Sets the starting day t_0^j of infection of vertex j*
4. *After each day t*
   - 4.1 *Counts the number of infective vertices*
   - 4.2 *For all vertices k*
     - *for SI:*
       do-nothing
     - *for SIR:*
       If duration of infection $T_{infection} = t - t_0^j$
       $\Rightarrow$ Remove vertex k
     - *for SI_1I_2:*
       If duration of (first period of) infection $T_{infection-1} = t - t_0^j$
       $\Rightarrow$ Replace $\rho_1 = \rho_2$
   - 4.3 *Go back to step* 3

Summary of paper IV

When I got the network about the sex-buyers and -sellers, my primary idea was to study the spread of infections on that sexual network. After all, many infections spread through sexual contacts and so far, connected sexual networks at such large scale was unavailable. The same trio team up again and in the paper "Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts" we study, by using simulated SI and SIR models, how the evolving sexual network affects the infection dynamics. The innovative aspect of this project is that rather than considering the static network structure, we use the actual time stamps on the edges, i.e. if a contact happens at time δ, the respective edge is available at that time and unavailable otherwise. Therefore, the network structure co-evolves with the infection dynamics. The temporal network model reflects the contact's temporal heterogeneities. In this type of networks, the number of possible paths if one wants to move between two vertices is much smaller than in the static model. One consequence is that the distribution of the number of

reachable vertices (from a random source) has a characteristic value, however, skewed towards smaller values and with a surprisingly high chance of small and null outbreaks. We also identify that the broad distribution of inter-contact times causes the infection to spread faster than if the contacts are regular in time. This result is the opposite of some recent research using different networks, i.e. on other datasets the broad distribution of inter-contact times slowdown the spread. Our findings can be (at least partially) explained by the fact that in our network most of the vertices join the network during short intervals, in other words, the vertices having contacts during, say, the initial quarter of the network do not appear in the final quarter. Nonetheless, not only the inhomogeneous inter-contact patterns but also the network structure plays a role; clusters constrain the dynamics such that the infection takes longer to spread in comparison to random structures. One surprising fact is that the infection dynamics in this contact structure has well-defined, high epidemic thresholds for both SI and SIR models. To complement the paper, we study a more realistic model of HIV spread and our simulations suggest that the specific type of commercial sex of our dataset is not a reservoir of major importance for HIV.

5.2 Control of epidemics

The definition of epidemics in the literature of epidemiology is quite general and indeed depends on the specific case of study [134]. A working definition according to "A dictionary of epidemiology" [59] is:

> *The occurrence in a community or region of cases of an illness, specific health-related behavior, or other health-related events clearly in excess of normal expectancy.*

In the physics literature, however, an epidemic regime emerges when the outbreak, i.e. the number of infected individuals (or vertices), is different than zero. A critical threshold usually characterizes the border between epidemic and non-epidemic regimes. One goal of epidemics control is to increase this threshold; another is to simply reduce the size of the outbreak. This is typically achieved by vaccination or educational campaigns. In the physics literature, these campaigns are usually indistinguishable and one generally refers to vaccination as the process of targeting an individual to immunize or to remove from the dynamics.

Individuals behave differently and some possess more risks than others to either get infected or transmit an infection. Due to limited resources and frequent inability to reach all population in society, it is thus desirable to increase herd (population) immunity by targeting as few individuals as possible. In terms of networks, this

different behavior is reflected in the connectivity of the vertices[f]. Therefore, the goal of the network studies is to identify such high-risk vertices by looking to the vertex connectivity in the network. Within the physics community, vaccination methods are closely connected to percolation problems [135]. In simple words, when the water pass from the top of the coffee powder (in the filter) to the bottom, there is a percolation (this scenario – above the percolation threshold – corresponds to the non-zero probability of epidemics). Rigorous mathematical percolation studies involve combinatorics and go beyond the scope of this thesis [136].

5.2.1 Topological methods

There is a great interest in exploring local and global network connectivity to develop better immunization strategies. Efficient immunization strategies (and disease spreading research) have been regarded as promising applications of network science to epidemiology. The fundamental idea is to identify central vertices that are more important to maintain the network connected, or to intermediate transmission processes, and simply remove them. The resulting fragmented network is expected to disrupt the propagation of infections between the several emerging connected components.

Random

A first attempt of a vaccination protocol is to uniformly select a fraction of vertices and remove them. There is evidence that this protocol is very ineffective, especially in networks where the number of contacts is broadly distributed [137, 135]. This protocol does not consider the information about the degree of the vertices, but in networks with broad degree distribution, most vertices have low-degree; the removal of them affects little the connectivity of the network.

Hubs

A natural extension of the random method is to include information about the local connectivity of the vertex. Vertices with larger degrees (hubs) have a diversity of neighbors and are expected to be more central because they connect different parts of the network. To illustrate the importance of the vertex-degree, in a simple model of infection spread, the importance of a vertex is proportional to its degree squared (k^2), since k neighbors can infect the vertex, which in turn, can propagate the infection to k other vertices. It has been shown in empirical and theoretical networks that the removal of vertices in order of highest degree is a quite effective method

[f]Risk behavior can be also modeled by varying the per-contact probability of transmitting an infection but to focus on structural effects and study the worst case scenario in terms of transmissibility, this probability is generally assumed to be at maximum, i.e. equal to one.

[137, 135]; after removal of few hubs, the network gets quite fragmented such that an emergent infection unlikely propagates between the connected components. The major criticism of this protocol is the hardness to identify hubs in real life, particularly in case of human contacts. On the other hand, they can be straightforwardly detected in electronic or artificial systems where, in principle, the entire network structure is known in advance.

Betweenness

Other centrality measures can be used in the same way as the degree to identify vertices to remove. As discussed in §4.1, the betweenness measures how important a vertex is to connect different clusters of a network. Therefore, targeting them turns to be quite effective in case of highly clustered networks. A typical case is related to vertices corresponding to travelers, these vertices make contacts in several locations and act as bridges between those places [138]. Another interesting example is the case of doctors, nurses and patients in hospitals. According to some results, doctors tend to have higher betweenness due to high mobility within a hospital, which implies in contacts with a diversity of other individuals [118].

Acquaintance immunization

In case of complete ignorance of the network structure, Cohen and collaborators devised a protocol that indirectly explore the connectivity of the network by making random choices [139]. In this protocol, a vertex is selected at random, and then, a random neighbor (or acquaintance) of this vertex is picked and vaccinated. This method explores the fact that randomly selected acquaintances possess more edges than the randomly selected vertices [140]. Especially in disassortative networks, the first random step are more likely to select a low-degree vertex that, in turn, is probably connected to a high-degree vertex (the random neighbor). This method and the previous ones can be improved by iteratively vaccinate the neighbors of the previously vaccinated individual [141].

k-cores

Sometimes, due to the community structure, high degree vertices are not as central as they would be in the same network with random edges, e.g. a randomized network version with same degree distribution as the original. In that case, the k-cores measure can capture which hubs are in the core and which are in the periphery of the network. Kitsak and collaborators showed that if an infection starts at vertices with high k-core, it gets more pervasive, on empirical networks, than if it starts in hubs or vertices with high betweenness [142]. Surprisingly, in their analysis, hubs are still more critical than high k-core vertices in case of multiple sources of infection.

5.2.2 Temporal methods

Since most of previous efforts have been based on static network studies, there is a great potential to explore temporal information to control epidemics. Contact patterns are not regular and indeed individuals are generally more active during short periods followed by absence of interactions. Furthermore, under some circumstances people tend to repeat their behavior [25, 79, 143], which means that one can learn from the past to forecast an individual activity. This information can be used to select temporally strategic persons to vaccinate.

Seasonal

At some scales, temporal patterns are characterized by regular variations due to seasonal or circadian rhythms. A simple vaccination protocol is thus to target all vertices active during these periods of higher activity, preferably at the beginning of the cycle to avoid future outbreaks of the infection. This method is regularly adopted, for instance, on annual Influenza vaccination campaigns or on educational campaings before major festivities where people are expected to have more sexual contacts. Although this method uses temporal information, it relies on targeting random individuals (within a specific cohort) of the static network.

Weight

An extension of the seasonal method is based on recording the historical activity of a certain individual and then to assume that the behavior repeats afterwards. This means that vaccinating and individual's strong acquaintances (i.e. pairs that occur more often) is equivalent to remove potential frequent future contacts (i.e. pairs that are likely to connect). It turns out that selecting the most frequent contact is more effective for weighted networks than simply choosing a neighbor at random [144].

Recent

When looking at a higher resolution in time, human contact behavior appears non-regular and actually, exhibit bursts of activity. Bursts mean that, during a certain period, a person is more active than at other periods, and the inter-event time between two contacts has no characteristic scale. Since communication and sexual activity seem to follow skewed distribution of inter-event times [145, 25], it is likely that a contact will happen again soon after a previous contact event. One way to explore this temporal structure is to simply select vertices at random and target the vertex's most recent contact to remove [144].

Summary of paper V

Since in paper IV, we observed that temporal correlations affect the dynamics of infection propagation, we decided to investigate vaccination methods using this extra (temporal) information. After some preliminary tests on the sexual network, the team, now led by Sungmin, ended up proposing, in the paper entitled "Exploiting temporal network structures of human interaction to effectively immunize populations", two methods of vaccination that, for diverse networks, appear to have higher control efficiency in comparison to known methods. The idea is to capture temporal patterns of activity, by learning an individual's behavior, and use that information to vaccinate acquaintances of randomly chosen individuals within the population. Therefore, in this paper, the network is split into two parts, one to learn the contact patterns of each person, and the other to actually simulate the propagation of infections. A number of individuals is vaccinated, at the first time after the learning period, by choosing a random vertex and removing one of its contacts according to the weight or the recent protocols discussed in the thesis. The weight protocol appears to be adequate when the contacts are more regularly spread in time, as for instance, in the email network where people are continuously contacting each other. On the other hand, when activity lasts only during short intervals, as the case of contact networks, the recent protocol seems more appropriate. This protocol takes advantage of the fact that, on average, people that have met recently are more likely to be socially active and thus meet again in the near future. The main contribution of this paper is the idea of using temporal information to efficiently immunize the population by knowing only local structural information, rather than the entire contact network.

5.3 Other dynamical processes

There are several other dynamic processes that are modeled by using networks as underlying structures that would be too extensive to include in this short section [16, 146, 147]. Differential equation models (disease-like) can be formulated to represent spreading of ideas, rumors [146], emotions, social behavior [148], acceptance of products in the marketing, or population dynamics [149]. There is also interest in opinion dynamics, i.e. how does one person changes its own opinion according to its contact's opinion, using spin-glass models [150, 151]. Methods based on the random walker process have been extensively proposed to study different properties of networks, for example, centrality measures [152], community detection [90], and cascading failures [153].

Most previous research has been focused on static networks, but studies considering dynamic and temporal networks have appeared recently. These papers

are mostly motivated by the temporal correlation due to the power-law distribution of inter-event times. Disease-like dynamics on email network [76, 154, 155] and mobile phone communication [55] suggest that temporal and cluster structures slowdown the spread of information. In a voter model, consensus takes longer to be achieved if power-law inter-event time is used rather than the usual exponential distribution[156].

Chapter 6

Conclusions and perspectives

Networks have gathered considerable attention since the late 1990s. The simplicity and generality of network theory are pervasive enough to be applied in different fields of science, from biology to archaeology, passing through arts, sports and military. Not unexpected, physicists like me jumped in and have continuously contributed to the different research fronts of fundamental aspects of networks, applications of networks, and in empirical studies to detect patterns or validate network models.

The main goal of this thesis is to explore empirical networks; both by detecting and studying non-trivial topological patterns, and by using these empirical networks as underlying structures where dynamical processes occur. The innovative aspects include collecting original networks and modeling them using further levels of complexity than typically adopted, as for instance, multiple-layers, bipartivity, edge-weigths and temporal information. Paper II is such an example, based on multivariate analysis, we propose a method to group similar vertices according to the bipartivity structure and edge-weight information. We apply the method to a network of consumer complaints about companies. In all other papers, temporal information is nevertheless the central actor. It is safe to say that this thesis is well placed in the history of network science. Research on dynamical networks is currently receiving much attention and the thesis adds important contributions to the field. Papers I and III are related to the analysis of temporal structures of networks, in the first case, I use simple static network measures to study how the network of flights and airports evolves annually, and in the second paper, we apply recent developed methods, specifically designed to characterize the network temporal profile, to understand the dynamics of a web-community, and as a consequence of the nature of the web-community, we obtain a large dynamic connected sexual network. Papers IV and V take a different direction, by using the knowledge from our previous results, the research moves to the study of epidemics co-evolving with network structure.

Theoretical models are undoubtedly elegant. By setting some simple rules or parameters, one can study the system evolution considering idealistic scenarios, or

perhaps, its response to perturbations. For a specific system, increasing model complexity is needed. However, this is not the usual approach in our field, where the goal is to understand general aspects of the system by using simple models. Empirical research, nevertheless, has its own beauty as well. Measuring signals and structures, finding relations, reveal how our universe actually behaves. This thesis takes the second path. By empirical analysis, we conclude that network structure does change. This might sound obvious at a first glance, but the non-trivial aspect is that structure changes in such irregular way that the static network misses important structures in some cases. A classic example is the order the contacts are made. We show in paper IV, for example, that a particular order of contacts increases the incidence of simulated infectious diseases. The bursty activity, i.e. the fact that individuals are highly active during short periods of time and idle otherwise, can be explored in simple efficient protocols to control epidemics. We show that potential against other methods in paper V. In papers I and III, we have observed that some networks are affected and consequently shaped by external factors through a feedback mechanism. The flights network, for example, showed a decreasing number of routes and flights, but an increasing number of carried passengers and cargo along the years; this suggests an optimization process where previously central airports transfer their importance to new ones. More remarkable is the interplay between online and offline activity studied in paper III. The quality of services offered by sex workers, rated in a web-forum, directly interferes in their future sexual network. While this observation might sound trivial, it is a direct measure of the pervasiveness of Internet in our social and economic life in the 2000s.

Empirical research leads us to experiments. As a physicist, I am fascinated by experiments, although, usually I do not appreciate the smell of some labs. Some fundamentalists claim that the "real physics" is in the experiments, the analytical analysis is essentially mathematics, but let us avoid these taxonomies and go further. Many researchers argue that experiments are too difficult in complex system research. It involves a plethora of ethical and logistic issues, sometimes high cost, that simply turn observational research a better and easier choice. I agree that can be difficult, but being difficult does not mean they are not doable. We have been identifying structural changes in the network due to external factors, but usually, in a passive observational way. The challenge for the near future is to have more control of these external factors by performing controlled experiments. Electronic gadgets, smart phones, and virtual environments, like online games and social networks, already cover a significant sample of the population and permit us to create original experiments, but of course, with some effort.

Overall, I am optimistic about network science and I believe that although much has been done, there is a lot yet to be explored, for example, in terms of applications, temporal network analysis, and feedback mechanism between structure and dynamics. Finally, to be a little speculative, I have hopes that large scale experimental network science might be one exciting new direction.

Bibliography

[1] R Falk and C Konold. Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, 104(2):301–318, 1997.

[2] M N Fried. Mathematics as the science of patterns. *Loci: Convergence*, 2011.

[3] G H Hardy. *A mathematician's apology*. University of Alberta Mathematical Sciences Society, first electronic edition, 1940.

[4] A Franklin. Experiment in physics. In Edward N Zalta, editor, *The Stanford encyclopedia of philosophy*. Spring edition, 2010.

[5] M Mitchell. *Complexity: A Guided Tour*. Oxford University Press, USA, 2009.

[6] www.pithemovie.com. PI, The Movie, 1998.

[7] K-I Goh, M E Cusick, D Valle, B Childs, M Vidal, and A-L Barabási. The human disease network. *Proceedings of the National Academy of Sciences of USA*, 104(21):8685–8690, 2007.

[8] D C Correa, J H Saito, and L da F Costa. Musical genres: Beating to the rhythms of different drums. *New Journal of Physics*, 12:053030, 2010.

[9] P Ormerod and A P Roach. The medieval inquisition: Scale-free networks and the suppression of heresy. *Physica A*, 339:645–652, 2004.

[10] L Eulero. Solvtio problematis ad geometriam sitvs. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

[11] A-L Barabási. *Linked: How Everything is Connected to Everything else and What it Means for Business, Science, and Everyday Life*. Plume Books, 2003.

[12] N L Biggs, E K Lloyd, and R J Wilson. *Graph Theory 1736–1936*. Clarendon Press, 1986.

[13] B Bollobás. *Modern Graph Theory*. Springer-Verlag, New York, 1998.

[14] S Wasserman and K Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, 1994.

[15] A-L Barabási and R Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[16] M E J Newman. *Networks: An Introduction*. Oxford University Press, USA, 2010.

[17] M Kurant and P Thiran. Layered complex networks. *Physical Review Letters*, 96(13):138701, 2006.

[18] L E C da Rocha and L da F Costa. 2d pattern evolution constrained by complex network dynamics. *New Journal of Physics*, 9(108), 2007.

[19] L E C Rocha. Structural evolution of the Brazilian airport network. *Journal of Statistical Mechanics*, (4):P04020, 2009.

[20] S V Buldyrev, R Parshani, G Paul, H E Stanley, and S Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464:1025–1028, 2010.

[21] E Estrada and D J Higham. Network properties revealed through matrix functions. *SIAM Review*, 52(4):696–714, 2010.

[22] L da F Costa, O N Oliveira Jr, G Travieso, F A Rodrigues, P R Villas Boas, L Antiqueira, M P Viana, and L E C Rocha. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[23] M de Berg, O Cheong, and M van Kreveld. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, third edition, 2008.

[24] S N Dorogovtsev and J F F Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London B*, 268:2603–2606, 2001.

[25] L E C Rocha, F Liljeros, and P Holme. Information dynamics shape the sexual networks of Internet-mediated prostitution. *Proceedings of the National Academy of Sciences of USA*, 107(13):5706–5711, 2010.

[26] M E J Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):16131, 2001.

[27] F Liljeros, J Giesecke, and P Holme. The contact network of inpatients in a regional healthcare system. A longitudinal case study. *Mathematical Population Studies: An International Journal of Mathematical Demography*, 14:269–284, 2007.

[28] R Lambiotte, V D Blondel, C de Kerchove, E Huens, C Prieur, Z Smoreda, and P van Dooren. Geographical dispersal of mobile communication networks. *Physica A*, 387:5317–5325, 2008.

[29] J P Eckmann, E Moses, and D Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of USA*, 101(40):14333–14337, 2004.

[30] B A N Travençolo and L da F Costa. Hierarchical spatial organization of geographical networks. *Journal of Physics A*, 41(22):224004–224004, 2008.

[31] R Ferrer i Cancho, C Janssen, and R V Solé. Topology of technology graphs: Small world patterns in electronic circuits. *Physical Review E*, 64(4):046119, 2001.

[32] R N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11:193–197, 1999.

[33] A A Tsonis, K L Swanson, and P J Roebber. What do networks have to do with climate. *Bulletin of the American Meteorological Society*, 87(5):585–596, 2006.

[34] R V Donner, Y Zou, J F Donges, N Marwan, and J Kurths. Recurrence networks - a novel paradigm for nonlinear time series analysis. *New Journal of Physics*, 12:033025, 2010.

[35] D J de S Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[36] R Albert, H Jeong, and A-L Barabási. Diameter of the world wide web. *Nature*, 401:130, 1999.

[37] S Bernhardsson and P Minnhagen. Selective pressure on metabolic network structures as measured from the random blind-watchmaker network. *New Journal of Physics*, 12:103047, 2010.

[38] R Solomonoff and A Rapoport. Connectivity of random nets. *Bulletin of Mathematical Biology*, 13(2):107–117, 1951.

[39] P Erdős and A Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

[40] G U Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B*, 213:21–87, 1925.

[41] H A Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[42] S Bornholdt and H Ebel. World wide web scaling exponent from Simon's 1955 model. *Physical Review E*, 64:035104(R), 2001.

[43] E Volz and Lauren A Meyers. Susceptible-infected-recovered epidemics in dynamic contact networks. *Proceedings of the Royal Society of London B*, 274:2925–2933, 2007.

[44] T Gross and B Blasius. Adaptive coevolutionary networks: A review. *Journal of the Royal Society Interface*, 5:259–271, 2008.

[45] S van Segbroeck, F C Santos, and J M Pacheco. Adaptive contact networks change effective disease infectiousness and dynamics. *PLoS Computational Biology*, 6(8):e1000895, 2010.

[46] W Schaper and D Scholz. Factors regulating arteriogenesis. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 23:114, 2003.

[47] T Gross, C J D D'Lima, and B Blasius. Epidemic dynamics on an adaptive network. *Physical Review Letters*, 96:208701, 2006.

[48] S Funk, M Salathé, and V A A Jansen. Modelling the influence of human behavior on the spread of infectious diseases: A review. *Journal ot the Royal Society Interface*, 7:1247–1256, 2010.

[49] L E C Rocha, F Liljeros, and P Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *Public Library of Science Computational Biology*, 7(3):e1001109, 2011.

[50] G Casella and R L Berger. *Statistical Inference*. Duxbury Press, second edition, 2001.

[51] R A Fisher. *The Design of Experiments*. Oliver and Boyd, second edition, 1937.

[52] U Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, first edition, 2006.

[53] N J Gotelli and G R Graves. *Null Models in Ecology*. Smithsonian Institute Press, 1996.

[54] S Maslov and K Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[55] M Karsai, M Kivela, R K Pan, K Kaski, J. Kertész, A-L Barabási, and J Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83:025102(R), 2011.

[56] N W Ashcroft and N D Mermin. *Solid State Physics*. Brooks Cole, 1976.

[57] S K Baek, P Minnhagen, and B J Kim. Percolation on hyperbolic lattices. *Physical Review E*, 79:011124, 2009.

[58] R M Anderson and R M May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, 1992.

[59] M Porta, editor. *A Dictionary of Epidemiology*. Oxford University Press, fifth edition, 2008.

[60] E Vynnycky and R G White. *An Introduction to Infectious Disease Modelling*. Oxford University Press, Oxford, 2010.

[61] P Pagel, S Kovac, M Oesterheld, B Brauner, I Dunger-Kaltenbach, G Frishman, C Montrone, P Mark, V Stümpflen, H W Mewes, A Ruepp, and D Frishman. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[62] www.genome.jp/kegg/. KEGG database, 2011.

[63] F Liljeros, C R Edling, L A N Amaral, H E Stanley, and Y Åberg. The web of human sexual contacts. *Nature*, 411:907–908, 2001.

[64] J Moody. The importance of relationship timing for diffusion. *Social Forces*, 81:25–56, 2002.

[65] D E Simanek. A glossary of frequently misused or misunderstood physics terms and concepts, 2004, http://www.lhup.edu/ dsimanek/glossary.htm.

[66] S H Lee, P-J Kim, and H Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:016102, 2006.

[67] M P H Stumpf, C Wiuf, and R M May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of USA*, 102(12):4221–4224, 2005.

[68] D D Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199, 1997.

[69] D Liben-Nowell and J Kleinberg. The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[70] A Clauset, C Moore, and M E J Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[71] D Butler. Data sharing threatens privacy. *Nature*, 449:644–645, 2007.

[72] Editorial. A matter of trust. *Nature*, 449:637–638, 2007.

[73] D Lazer, A Pentland, L Adamic, S Aral, A-L Barabási, D Brewer, N Christakis, N Contractor, J Fowler, M Gutmann, T Jebara, G King, M Macy, D Roy, and M van Alstyne. Computational social science. *Science*, 323:721–723, 2009.

[74] S Milgram. The small world problem. *Psychology Today*, 1(1):60–67, 1967.

[75] P S Dodds, R Muhamad, and D J Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003.

[76] J L Iribarren and E Moro. Impact of human activity patterns on the dynamics of information diffusion. *Physical Review Letters*, 103:038702, 2009.

[77] C Cattuto, W van den Broeck, A Barrat, V Colizza, J-F Pinton, and A Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *Public Library of Science One*, 5(7):e11596, 2010.

[78] 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. *Proceedings of the National Academy of Sciences of USA*, 107(51):22020–22025, 2010.

[79] T Takaguchi, M Nakamura, N Sato, K Yano, and N Masuda. Predictability of conversation partners. *arXiv:1104.5344*, 2011.

[80] N Eagle, A Pentland, and D Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of USA*, 106(36):15274–15278, 2009.

[81] W S Bainbridge. The scientific research potential of virtual worlds. *Science*, 317:472–476, 2007.

[82] M Szella, R Lambiotte, and S Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences of USA*, 107(31):13636–13641, 2010.

[83] L C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[84] M E J Newman and M Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[85] L da F Costa and L E C da Rocha. A generalized approach to complex network. *The European Physical Journal B*, 50:237–242, 2006.

[86] S B Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

[87] H C White and R L Breiger. Pattern across networks. *Society*, 12(5):68–74, 1975.

[88] S Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[89] A Lancichinetti and S Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80:056117, 2009.

[90] M Rosvall, D Axelsson, and C T Bergstrom. The map equation. *The European Physical Journal - Special Topics*, 178(1):13–23, 2009.

[91] D Garlaschelli and M I Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93:268701, 2004.

[92] A Clauset, C R Shalizi, and M E J Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4), 2009.

[93] J N Kapur and H K Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, 1992.

[94] T H Cormen, C E Leiserson, R L Rivest, and C Stein. *Introduction to Algorithms*. MIT Press, third edition, 2009.

[95] L da F Costa, F A Rodrigues, G Travieso, and P R Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56:167–242, 2007.

[96] L E C Rocha and P Holme. The network organisation of consumer complaints. *Europhysics Letters*, 91(2):28005, 2010.

[97] I T Jolliffe. *Principal Component Analysis*. Springer series in statistics. Springer, second edition, 2002.

[98] P Holme. Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174, 2004.

[99] J Tang, S Scellato, M Musolesi, C Mascolo, and V Latora. Small-world behavior in time-varying graphs. *Physical Review E*, 81:055101(R), 2010.

[100] J Tang, M Musolesi, C Mascolo, V Latora, and V Nicosia. Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, pages 1–6, New York, USA, 2010. SNS'10.

[101] R K Pan and J Saramäki. Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, 84:016105, 2011.

[102] P Grindrod and M C Parsons. Communicability across evolving networks. *Physical Review E*, 83:046120, 2011.

[103] C-K Peng, S V Buldyrev, S Havlin, M Simons, H E Stanley, and A L Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E*, 49:1685–1689, 1994.

[104] D Rybski, S V Buldyrev, S Havlin, F Liljeros, and H A Makse. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences of USA*, 106(31):612640–12645, 2009.

[105] C-K Peng, S Havlin, H E Stanley, and A L Goldberger. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos*, 5(82), 1995.

[106] B F de Blasio, Å Svensson, and F Liljeros. Preferential attachment in sexual networks. *Proceedings of the National Academy of Sciences of USA*, 104(26):10762 – 10767, 2007.

[107] P L Krapivsky, S Redner, and F Leyvraz. Connectivity of growing random network. *Physical Review Letters*, 85:4629–4632, 2000.

[108] R Gaillard, S Dehaene, C Adam, S Clémenceau, D Hasboun, M Baulac, L Cohen, and L Naccache. Converging intracranial markers of conscious access. *Public Library of Science Biology*, 7(3):e1000061, 2009.

[109] D Liben-Nowell and J Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences of USA*, 105(12):4633–4638, 2008.

[110] World health statistics 2011. Technical report, World Health Organization, 2011.

[111] D Bernoulli and S Blower. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in Medical Virology*, 14:275–288, 2004.

[112] W H Hamer. Epidemic disease in England. *The Lancet I*, pages 733–739, 1906.

[113] W O Kermack and A G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 115:700–721, 1927.

[114] H W Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[115] S M Blower and P Farmer. Predicting the public health impact of antiretrovirals: Preventing HIV in developing countries. *AIDScience*, 3(11), 2003.

[116] D J Watts, R Muhamad, D C Medina, and P S Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences of USA*, 102(32):11157–11162, 2005.

[117] V Colizza, A Barrat, M Barthelemy, A-J Valleron, and A Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *Public Library of Science Medicine*, 4(1):e13, 2007.

[118] T Ueno and N Masuda. Controlling nosocomial infection based on structure of hospital social networks. *Journal of Theoretical Biology*, 254:655–666, 2008.

[119] M E J Newman and D J Watts. Scaling and percolation in the small-world network mode. *Physical Review E*, 60:7332–7342, 1999.

[120] R Pastor-Satorras and A Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, 2001.

[121] M E J Newman. Spread of epidemic disease on networks. *Physical Review E*, 66:016128, 2002.

[122] E M Volz. Dynamics of infectious disease in clustered networks with arbitrary degree distributions. *arXiv:1006.0970v1*, 2010.

[123] S Bansal, J Read, B Pourbohloul, and L A Meyers. The dynamic nature of contact networks in infectious disease epidemiology. *Journal of Biological Dynamics*, 4(5):478–489, 2010.

[124] M C González, C A Hidalgo, and A-L Barabási. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[125] J M Read, K T D Eames, and W J Edmunds. Dynamic social networks and the implications for the spread of infectious disease. *Journal of the Royal Society Interface*, 5:1001–1007, 2008.

[126] J Mossong, N Hens, M Jit, P Beutels, K Auranen, R Mikolajczyk, M Massari, S Salmaso, G S Tomba, J Wallinga, and et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *Public Library of Science Medicine*, 5(3):e74, 2008.

[127] M C Vernon and M J Keeling. Representing the UK's cattle herd as static and dynamic networks. *Proceedings of the Royal Society of London B*, 276(1656):469–476, 2009.

[128] M Morris and M Kretzschmar. Concurrent partnerships and the spread of HIV. *AIDS*, 11(5):641–648, 1997.

[129] L Sawers and E Stillwaggon. Concurrent sexual partnerships do not explain the HIV epidemics in Africa: A systematic review of the evidence. *Journal of the International AIDS Society*, 13(34):1–23, 2010.

[130] L Stone, R Olinky, and A Huppert. Seasonal dynamics of recurrent epidemics. *Nature*, 446(29):533–536, 2007.

[131] M F Heath, M C Vernon, and C R Webb. Construction of networks with intrinsic temporal structure from UK cattle movement data. *BMC Veterinary Research*, 4(11), 2008.

[132] C Kamp. Untangling the interplay between epidemic spread and transmission network dynamics. *Public Library of Science Computational Biology*, 6(11), 2010.

[133] T House and M J Keeling. Insights from unifying modern approximations to infections on networks. *Journal of the Royal Society Interface*, 8:67–73, 2011.

[134] M S Green, T Swartz, E Mayshar, B Lev, A Leventhal, P E Slater, and J Shemer. When is an epidemic an epidemic? *The Israel Medical Association Journal*, 4:3–6, 2002.

[135] R Cohen, K Erez, D ben Avraham, and S Havlin. Breakdown of the Internet under intentional attack. *Physical Review Letters*, 86(16):3682–3685, 2001.

[136] D Stauffer and A Aharony. *Introduction to Percolation Theory*. CRC Press, second edition, 1994.

[137] R Albert, H Jeong, and A-L Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[138] M K Nordvik, F Liljeros, A Österlund, and B Herrman. Spatial bridges and the spread of chlamydia: The case of a county in Sweden. *Sexually Transmitted Diseases*, 34(1):47–53, 2007.

[139] R Cohen, S Havlin, and D ben Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24):247901, 2003.

[140] M E J Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25(1):83–95, 2002.

[141] P Holme. Efficient local strategies for vaccination and network attack. *Europhysics Letters*, 68(6):908, 2004.

[142] M Kitsak, L K. Gallos, S Havlin, F Liljeros, L Muchnik, H E Stanley, and H A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6:888–893, 2010.

[143] C Song, Z Qu, N Blumm, and A-L Barabási. Limits of predictability in human mobility. *Science*, 327:1018–1021, 2010.

[144] S Lee, L E C Rocha, F Liljeros, and P Holme. Exploiting temporal network structures of human interaction to effectively immunize population. *arXiv:q-bio/1011.3928*, 2010.

[145] A-L Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[146] S Boccaletti, V Latora, Y Moreno, M Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[147] A Barrat, M Barthélemy, and A Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, Cambridge, first edition, 2008.

[148] A L Hill, D G Rand, M A Nowak, and N A Christakis. Infectious disease modeling of social contagion in networks. *Public Library of Science Computational Biology*, 6(11):e1000968, 2010.

[149] S Sahasrabudhe and A E Motter. Rescuing ecosystems from extinction cascades through compensatory perturbations. *Nature Communications*, 2(170), 2011.

[150] A T Bernardes, D Stauffer, and J Kertész. Election results and the Sznajd model on Barabasi network. *The European Physical Journal B*, 25:123–127, 2002.

[151] P Sobkowicz. Modelling opinion formation with physics tools: Call for closer link with reality. *Journal of Artificial Societies and Social Simulation*, 12(1):11, 2009.

[152] M E J Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

[153] I Simonsen, L Buzna, K Peters, S Bornholdt, and D Helbing. Transient dynamics increasing network vulnerability to cascading failures. *Physical Review Letters*, 100(21):218701, 2008.

[154] A Vazquez, B Rácz, A Lukács, and A-L Barabási. Impact of non-Poissonian activity patterns on spreading processes. *Physical Review Letters*, 98:158702, 2007.

[155] G Miritello, E Moro, and R Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83:045102(R), 2011.

[156] T Takaguchi and N Masuda. Voter model with non-Poissonian inter-event intervals. *arXiv:1011.4445v1*, 2010.