Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2006 REPORT B55

COMPLEX NETWORKS IN THE STUDY OF FINANCIAL AND SOCIAL SYSTEMS

Jukka-Pekka Onnela

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2006 REPORT B55

COMPLEX NETWORKS IN THE STUDY OF FINANCIAL AND SOCIAL SYSTEMS

Jukka-Pekka Onnela

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Electrical and Communications Engineering, Helsinki University of Technology, for public examination and debate in Auditorium E at Helsinki University of Technology (Espoo, Finland) on the 10^{th} of July, 2006, at 12 noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio

Distribution:
Helsinki University of Technology
Laboratory of Computational Engineering
P. O. Box 9203
FIN-02015 HUT
FINLAND
Tel. +358-9-451 4826
Fax. +358-9-451 4830
http://www.lce.hut.fi

Online in PDF format: http://lib.tkk.fi/Diss/2006/isbn9512282704

E-mail: Jukka-Pekka.Onnela@tkk.fi

ISBN 951-22-8269-0 (printed)
ISBN 951-22-8270-4 (PDF)
ISSN 1455-0474
PicaSet Oy
Helsinki 2006

HELSINKI UNIVERSITY OF TECHNOLOGY
P. O. BOX 1000, FI-02015 TKK
http://www.tkk.fi

ABSTRACT OF DOCTORAL DISSERTATION

Author Jukka-Pekka Onnela

Name of the dissertation
COMPLEX NETWORKS IN THE STUDY OF FINANCIAL AND SOCIAL SYSTEMS

Date of manuscript 25.4.2006 | Date of the dissertation 10.7.2006

☐ Monograph ☒ Article dissertation (summary + original articles)

Department Department of Electrical and Communications Engineering
Laboratory Laboratory of Computational Engineering
Field of research Complex systems
Opponent(s) Prof. Alessandro Vespignani
Supervisor Acad. Prof. Kimmo Kaski
(Instructor)

Abstract

Complex systems consist of a large number of interacting elements, giving rise to the emergence of organisation without any external organising principle being applied. Consequently, decomposing the system and studying its subparts in isolation does not contribute to our understanding of how it works. Fortunately, complex systems can be described, analysed, and modelled using complex networks. Here one focuses only on the elements and topology of interactions between them, providing a system-level perspective to the system under study.

This thesis contributes to the network approach to complex systems in two ways. First, using empirical data on the financial market, we show how a general problem dealing with correlated actors can be recast as a network problem. In principle, this simple method is applicable to any complex system with temporal correlations of quantities attached to each element. In the context of social systems we demonstrate how a network model of social actors can be constructed to capture the hypothesised structure of interactions. Using the rate equation approach, we develop a simple phenomenological model intended for future study of processes unfolding on social networks. Second, we develop generalisable methods and measures for network characterisation. These measures are not too application specific so they can be transferred to disparate complex systems. For example, we augment the motif framework to incorporate interaction strengths, enabling us to go beyond topology and account for the heterogeneity of interactions.

The studied complex systems can be classified broadly as financial and social systems. As data rich systems they enable a thorough testing of the developed concepts. In addition, they have numerous applications within the complex systems paradigm, including developing better risk management schemes and enabling *in silico* testing of different disease immunisation scenarios. The focus of this thesis lies in the universal aspects of these systems; the line of enquiry is to motivate a question in one system, then step back and see if the machinery developed to tackle the problem has relevance to other systems. In addition to conceptual developments, we provide tools for characterising complex networks in practice.

Keywords complex systems, complex networks, financial systems, social systems, network models

ISBN (printed) 951-22-8269-0 | ISSN (printed) 1455-0474
ISBN (pdf) 951-22-8270-4 | ISSN (pdf)
ISBN (others) | Number of pages 75 p. + app. 72 p.

Publisher Laboratory of Computational Engineering

Print distribution Laboratory of Computational Engineering

☒ The dissertation can be read at http://lib.tkk.fi/Diss/2006/isbn9512282704

TEKNILLINEN KORKEAKOULU PL 1000, 02015 TKK http://www.tkk.fi	VÄITÖSKIRJAN TIIVISTELMÄ

Tekijä Jukka-Pekka Onnela

Väitöskirjan nimi
COMPLEX NETWORKS IN THE STUDY OF FINANCIAL AND SOCIAL SYSTEMS

Käsikirjoituksen jättämispäivämäärä 25.4.2006	Väitöstilaisuuden ajankohta 10.7.2006

☐ Monografia ☒ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)

Osasto	Sähkö- ja tietoliikennetekniikan osasto
Laboratorio	Laskennallisen tekniikan laboratorio
Tutkimusala	Kompleksiset systeemit
Vastaväittäjä(t)	prof. Alessandro Vespignani
Työn valvoja	akat. prof. Kimmo Kaski
(Työn ohjaaja)	

Tiivistelmä

Kompleksisessa systeemissä on suuri määrä keskenään vuorovaikuttavia elementtejä, joiden ansiosta siihen syntyy rakennetta ilman ulkopuolista ohjausta. Tämän vuoksi osien erillinen tarkastelu ei anna kuvaa systeemin käyttäytymisestä. Kompleksisia systeemejä voidaan kuitenkin kuvata, analysoida ja mallintaa kompleksisten verkkojen avulla. Tässä lähestymistavassa keskitytään vain elementteihin ja niiden välisten vuorovaikutusten topologiaan.

Väitöskirja pyrkii kahdella tavalla edistämään verkkoihin pohjautuvaa lähestymistapaa kompleksisiin systeemeihin. Yhtäältä osoitamme kuinka tietyt probleemat voidaan kuvata verkkoina. Esimerkkinä tästä käytämme rahoitusmarkkinoilta peräisin olevaa dataa ja osoitamme, että tällainen korreloituneita toimijoita käsittelevä ongelma voidaan esittää verkkona. Periaatteessa tämä yksinkertainen menetelmä soveltuu useisiin systeemeihin, missä verkon solmuihin liittyvät suureet ovat korreloituneita. Sosiaalisten systeemien yhteydessä esitämme, kuinka toimijoiden välille oletetut vuorovaikutussuhteet voidaan toteuttaa malliin. Rate-yhtälöiden avulla kehittämämme fenomenologisen mallin tarkoituksena on jatkossa tutkia erilaisten prosessien käyttäytymistä sosiaalisissa verkoissa. Toisaalta kehitämme yleiskäyttöisiä menetelmiä ja mittoja verkkojen karakterisointiin. Mitat ovat varsin yleisluontoisia, jotta niitä voitaisiin hyödyntää eri systeemien tutkimuksessa. Esimerkkinä tästä laajennamme motiivi-viitekehystä huomioimalla topologian lisäksi vuorovaikutusten voimakkuuden.

Väitöskirjassa tutkitut kompleksiset systeemit voidaan väljästi jakaa taloudellisiin ja sosiaalisiin. Nämä soveltuvat hyvin esitettyjen menetelmien ja mittojen tutkimiseen, koska niistä on dataa runsaasti saatavilla. Lisäksi molemmille on lukuisia sovelluksia, esimerkiksi rahoitusmarkkinoiden riskien kehittyneempi hallinta ja tarttuvien tautien erilaisten immunisointiskenaarioiden laskennallinen mallinnus. Väitöskirja keskittyy tarkasteltujen systeemien universaaleihin ominaisuuksiin. Kantavana ajatuksena on ollut kehittää koneistoa jonkin tietyn systeemin tarkasteluun ja hahmottaa, missä määrin menetelmät ovat sovellettavissa johonkin toiseen systeemiin liittyvän ongelman tarkasteluun. Lisäksi tarkoituksena on ollut kehittää käytännön työkaluja erilaisten kompleksisten verkkojen karakterisointiin.

Asiasanat kompleksiset systeemit, kompleksiset verkostot, taloudelliset ja sosiaaliset systeemit, verkostomallit

ISBN (painettu) 951-22-8269-0	ISSN (painettu) 1455-0474
ISBN (pdf) 951-22-8270-4	ISSN (pdf)
ISBN (muut)	Sivumäärä 75 s. + liit. 72 s.

Julkaisija Laskennallisen tekniikan laboratorio

Painetun väitöskirjan jakelu Laskennallisen tekniikan laboratorio

☒ Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/2006/isbn9512282704

Preface

This dissertation was completed in the Complex Networks and Agent-Based Models group, Laboratory of Computational Engineering, Helsinki University of Technology, and it concludes my Doctor of Science degree. I had a good time doing this work, and hope that some of my enthusiasm has been transferred to these pages.

I am grateful to my supervisor Acad. Prof. Kimmo Kaski for his continual encouragement and support, and for providing an excellent environment for this work to take place. I have been privileged to have Prof. János Kertész as my co-supervisor, and I am thankful for his abundant and always insightful suggestions. I have enjoyed and benefited from numerous lively discussions with Dr. Jari Saramäki, which are probably my best experiences of science in progress. I am grateful to Dr. Anirban Chakraborti for his input in the early stages of this work. I would also like to thank my collaborators Riitta Toivonen and Jörkki Hyvönen for their effort in our joint projects.

I would also like to thank the other people in our group, namely, Riku Linna, Ville Mäkinen, Jussi Kumpula, Tapio Heimo, and Antti Niinikoski. I have enjoyed many interesting discussions and debates on and off the topic with Marko Sysi-Aho, Teemu Leppänen, and Markus Miettinen. I am thankful to Eeva Lampinen for patiently taking care of so many practical matters that have come up throughout the years. I would also like to express my gratitude to my other collaborators. Although our common work is not included in this thesis, they have significantly influenced my doctoral studies. These people are Dr. Juuso Töyli, Dr. Gábor Szabó, Prof. David Lazer, and Prof. Albert-László Barabási.

Finally, I am indebted to my parents Merja and Seppo and my grandmother Liisi in so many ways, and grateful to my brother Juha-Matti for being there. Last, I would like to thank you, Kulsoom, for all the things you are.

Espoo, 19th June 2006

Jukka-Pekka Onnela

List of publications

This thesis consists of an introduction and the following publications:

(1) J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, Dynamic asset trees and portfolio analysis, *Eur. Phys. J. B*, **30**, 285 (2002).

(2) J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész, Dynamic asset trees and Black Monday, *Physica A*, **324**, 247 (2003).

(3) J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto, Asset trees and asset graphs in financial markets, *Physica Scripta*, **106**, 48 (2003).

(4) J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto, Dynamics of market correlations: Taxonomy and portfolio analysis, *Phys. Rev. E*, **68**, 056110 (2003).

(5) J.-P. Onnela, K. Kaski, and J. Kertész, Clustering and information in correlation based financial networks, *Eur. Phys. J. B*, **38**, 353 (2004).

(6) J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, Intensity and coherence of motifs in weighted complex networks, *Phys. Rev. E*, **71**, 065103 (2005).

(7) J. Saramäki, J.-P. Onnela, J. Kertész, and K. Kaski, Characterizing Motifs in Weighted Complex Networks in: J.F.F. Mendes, *et al.* (Eds.), Science of Complex Networks, *AIP Conference Proceedings*, **776**, 108 (2005).

(8) J.-P. Onnela, J. Saramäki, K. Kaski, and J. Kertész, Financial market - a network perspective in: H. Takayasu (Ed.), Practical Fruits of Econophysics, *Nikkei Econophysics III Proceedings*, Springer, Tokyo, 302 (2006).

(9) R. Toivonen, J.-P. Onnela, J. Saramäki, J. Hyvönen, and K. Kaski, A Model for Social Networks, *Physica A, in press* (2006).

Author's contribution

The research reported in this thesis is a result of collaboration between Jukka-Pekka Onnela, the author of this thesis, and the other authors of the included publications. He has initiated the research lines leading to Publications (1; 2; 3; 4; 5), and had a central role in developing the ideas for Publications (6; 7; 8). He has implemented all computer programs used in Publications (1; 2; 3; 4; 5; 6; 7; 8). He has analysed all data and performed numerical calculations under the supervision of more senior scientists in all Publications except (9). As the primary author of publications (1; 2; 3; 4; 5; 6; 8) he is responsible for their written material, and as the second author of publications (7; 9) has significantly contributed to their writing. He has also been actively involved in developing the model presented in Publication (9) and verified its analytical results.

Contents

Chapter 1

Introduction

This Chapter reviews the background of the research presented in this thesis and summarises its purpose.

1.1 Background of research

The conventional view of science holds knowledge to be specific and, thus, more knowledge may be gained by providing more details of the system under study. This view has lead to professional specialisation, with the outcome that the individual disciplines have progressively become more and more isolated from one another. This conception of science has recently been challenged by a paradigmatic shift taking place in the natural sciences. As a consequence, scientific enquiry is becoming increasingly holistic, more focused on system-level behaviour rather than system constituents, but also increasingly multidisciplinary as the existence of potential synergies across different fields is being acknowledged (10).

The study of complex systems in a unified framework has recently become recognised as a new discipline, breaking barriers not only between physics, chemistry and biology, but also between these disciplines and some of the so-called softer sciences, including economics, sociology, and psychology. There are three primary reasons for the emergence of the complex systems approach: (1) large electronically recorded sets of data on these systems have become available, (2) advances in computing power enable simulation and analysis of these systems, and (3) enough mathematical machinery exists for tackling these problems. The increasing availability of data and computer power have enabled scientists to search for regularities and patterns in different complex systems, which can be seen as manifestations of the underlying laws governing their dynamics (11). Computational capacity is crucial because theoretical research typically involves large-scale numerical simulations, and computers are also needed in the analysis and interpretation of data, whether obtained from empirical observations or from

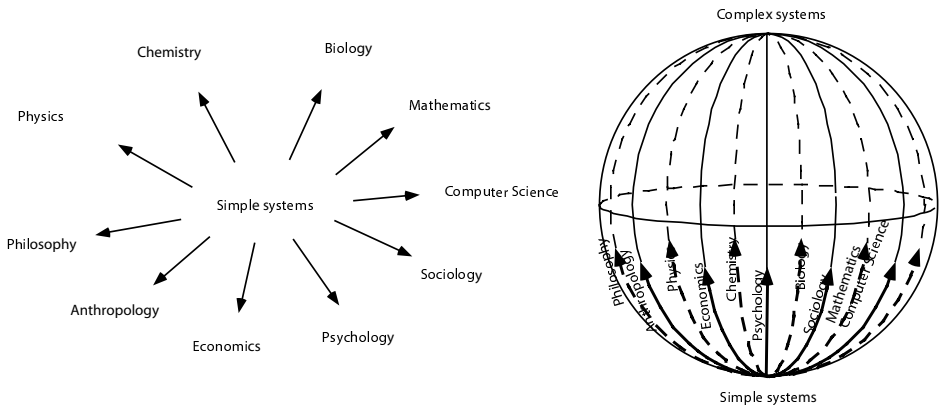

Figure 1.1: Different conceptions of science. According to the conventional view of science (left panel), different disciplines are progressively separating in order to gain knowledge. The complex systems view suggests that all complex systems have universal properties (right panel), and by considering them one can approach the specifics of particular complex systems both from the top of the sphere as well as from the bottom. Figure adapted from (10).

simulations. The relevant mathematical developments have mostly taken place in statistical mechanics, non-linear dynamics, and in the physics of complex networks (12; 13). Indeed, many complex systems are amenable to the complex network paradigm in which spectacular developments have been made in the past few years (Section 3.1). In this framework we learn about a complex system by studying its network representation. The approach is possible because the interaction topology of the underlying system, captured by the network, is related to its function and dynamics.

This thesis deals with two different complex systems, broadly classified as financial and social systems, from the perspective of complex networks. Loosely defined, complex systems consist of a large number of elements capable of interacting with each other and with their environment, resulting in *the system becoming organised without any external organising principle being applied*. The consequence of this, which is also a characteristic of complex systems, is that decomposing the system and studying its subparts in isolation does not contribute to our understanding of how it works. Putting the two concepts together, we can identify complex systems by what they do (organisation or emergence), and also by how they may or may not be analysed (decomposition) (11; 14). It may seem that almost any system could be classified as a complex system. While a more precise definition of complex systems reveals that this is not the case, it is also

true that numerous systems can be studied from the complex systems perspective. The complex systems approach is, therefore, better distinguished from alternative approaches by the type of questions asked, rather than the specific systems studied.

It is undeniable that there exists a dichotomy between universality and specificity. Put differently, there is always the danger that the universalist will step on the specifist's toes. However, a study of universal principles does not and should not replace a detailed description of a particular complex system. Indeed, there is no reason to view universality and specificity as mutually exclusive approaches (Figure1.1). The view of the field of complex systems suggests that instead of science separating into disparate disciplines, it may also be fruitful to consider the common properties of complex systems. This system level understanding may then guide us and simplify our enquiries into the study of the specifics (10). While one should always view every mode of scientific enquiry with some scepticism, it appears that at least physicist Stephen Hawking has faith in the approach: "I think the next century will be the century of complexity" (15). Let us hope that he is right or, at the very least, that in pursuing this approach we will not step on his toes.

1.2 Purpose of research

One of the key success factors of the complex networks approach is the wide applicability of relatively simple techniques across various fields. The purpose of the research presented in this thesis is to contribute to this overall approach in two ways by showing how a problem may be formulated as a network problem and how the resulting network may be analysed. More specifically, on the one hand we show how a general problem dealing with correlated actors can be recast as a network problem (financial systems), and also how a model can be developed to capture the imposed or hypothesised structure of actors (social systems). On the other hand, we show how to develop generalisable methods and measures for network characterisation. These measures should not be too application specific, such that they can be transferred to disparate complex systems not studied here, hopefully leading to cross-fertilisation of ideas.

Financial and social systems were chosen because they are data rich systems and thus enable the concepts developed to be thoroughly tested, but also because there are a wealth of potential applications for them. Further, financial systems arise from interactions between people, firms, cities, or nations, and these can be seen as coarse-grained social actors embedded in their own social network. This is to say that financial and social systems are not completely separate systems, but that every financial system also reflects an underlying social system (16).

In the context of the financial market, the purpose is to construct a network

representation of the system aimed at characterising the top-level structure and dynamics of a large number of interacting elements. Instead of looking at the interactions directly, our perspective is a little more abstract, focusing on correlations of interactions. The methods are applicable, at least in principle, to any complex system in which interactions are reflected in temporal correlations of some relevant quantity attached to each node, and serve the general need of identifying the core or backbone of the system. The resulting simplified system is more amenable to analysis and a number of different measures may be developed to characterise it further.

Another important line of research in this thesis has to do with augmenting the motif framework to incorporate link weights. The motivation for this has to do with the nature of the coupling between interaction strengths and local network topology. It not only appears that the nature of this coupling is not universal, but also that the coupling might be related to the function of the network (17). Consequently, augmenting the motif framework can improve our understanding of networked systems, enabling us to go beyond topology. In addition to conceptual developments, the aim has been to provide practical tools for making measurements of weighted motifs.

While the study of complex systems has to start from empirical observations using real life data, the natural next step is to build network models. Using the rate equation approach, we develop a simple phenomenological model that mimics the structure of various kinds of social networks. We have followed the principle of parsimony in developing the model, as inclusion of too many parameters and processes will only obscure qualitative understanding. Also, the purpose of the model is not to build perfectly realistic social networks per se. Instead, it is intended for future system-level study of processes unfolding on social networks and how they are affected by the network's structural properties, such as the existence of communities, and the coupling between network topology and interaction strengths.

Chapter 2

Overview of complex systems

This Chapter provides a brief overview of complex systems. It then introduces the two complex systems studied in this thesis, namely, the financial market and the network of social acquaintances, reviewing very briefly some of the elementary background of these systems.

2.1 Complex systems

The dictionary definition of complex is (1) consisting of interconnected or interwoven parts, and (2) not easy to understand. Of these, definition (1) is closely related to the definition of a complex system meant here, although it has to be admitted that no succinct definition exists. Some of the prominent characteristics of complex systems are: (a) they contain a large number of interacting elements (e.g., molecules, neurons, individuals, stocks), (b) the interactions are stochastic, and (c) the topology of interactions is sparse. Based on the number of elements these systems contain, they are considered to be within the mesoscopic or mesoscale domain, although theoretical approaches often assume the size of the system to be infinite. The sparsity of interactions indicates that the system is not homogeneous in structure, suggesting that complex systems can be characterised by the topology of their interactions.

Some often cited examples of complex systems are the economy, social systems, food webs, ecosystems, the Internet, traffic, animal flocks, intracellular systems, and the brain. As this diverse list makes clear, the complex systems paradigm has more to do with how something is studied than what is studied. Here lies the key to understanding complex systems: we must understand not only the behaviour of the parts but how they act together to form the behaviour of the whole. In practice, it is often not possible to treat these systems analytically and, therefore, computer simulations play a key role in their study (14; 11).

It is instructive to distinguish between simple, complex, and complicated sys-

tems (11; 14). *Simple systems* typically have a small number of components which act according to well-understood laws. For example, the pendulum consists of just one part, and its behaviour can be described using Newton's equations of motion. An oscillator, spinning wheel, and orbiting planet serve as other examples of simple systems. *Complicated systems*, on the other hand, have a large number of components which have well-defined roles and are governed by well-understood rules. Importantly, the pieces in complicated systems can be understood well in isolation, and the whole can be assembled from its parts. For example, a Boeing 747-400 has some 3×10^6 parts, which all have to work in unison to accomplish a function, to make the plane fly. One defect in a key part can bring the system to a halt (18).

Complex systems typically also have a large number of components. In contrast to complicated systems, the common characteristic is that they display organisation without any external organising principle being applied. The complexity of the system emerges from the behaviours of the numerous interacting simple parts, and the behaviour of the small part is different in isolation from when it is part of the larger system. This is usually referred to as *emergence of complexity*. Because of emergence, decomposing the system and studying its subparts in isolation does not contribute to our understanding of how it works. This separates it from a complicated system. Indeed, it appears that the emergent properties are unpredictable from the properties of the component parts. This insight questions some of the commonly held expectations about, for example, economics. Is it realistic to assume that we can derive macroeconomics from microeconomics? If not, then we must surely fail to predict the behaviour of global economy from the behaviour of nations. Applied to social systems, we can ask if the notion of emergent properties imply that societal behaviour cannot be adequately described by any practically achievable integration across the behaviour of individuals (19)? If this is the case, even if we knew everything that we feasibly can know about the behaviour of individuals in a group, we still would not be able to describe the behaviour of the individuals as a group. Based on the above, it is clear that the categories of simple and complex systems differ greatly in size and diversity, and that there are, indeed, many complex systems around us. The late mathematician Stanislaus Ulam put this into perspective by commenting that research on complex systems might be compared to the study of non-elephants (19).

To study a complex system one must focus on the right level of description to catch the phenomena of interest. This means that one needs to course-grain the system, omitting the details, until one arrives at the desired level of abstraction. For example, if we try to simulate protein dynamics by following each small part of the molecule, nothing *biologically interesting* can happen on the time scale allowed for by this level of detail. This is because a successful description of the small-scale structure may be irrelevant for the large-scale structure. This is even more obvious if we consider a higher level system, such as a social system. While

it is true that every system is fundamentally a physical system, it does not mean that every system should be modelled as such (20).

One of the core assumptions in the study of complex systems is the existence of universality across different complex systems. The idea is, as examples from statistical mechanics demonstrate, that the fine details do not matter for the behaviour of the whole. Consequently, some aspects of complex behaviour are anticipated in many different systems consisting of a large number of interacting agents. However, unlike for simple systems, it may be that there are no general laws for complex systems valid for all times and places. Instead, one might need to reach for lessons that may be learnt in one system and applied in another (20). Also, what has traditionally constituted a scientific result may need to be re-evaluated. For example, network analysis has revealed the structure of metabolic networks, and this fact alone gives rise to a number of questions and avenues of enquiry that otherwise may not have occurred (18).

2.2 Financial systems

Financial markets are often characterised as evolving complex systems (16; 21; 22). The evolution is a reflection of the changing power structure in the market and it manifests the passing of different products and product generations, new technologies, management teams, alliances and partnerships, among many other factors (4). Financial markets are possibly the largest and the most data-rich real-world complex system and, as such, are an excellent empirical test-bed for tools and theories. Indeed, one of the major challenges facing complex systems research is the real-world validation of any theoretical concept. Fortunately, the benefits of an improved understanding of financial markets are not limited to the academia, but have potential for several applications, including better risk hedging schemes.

Choosing the appropriate variables for studying the financial market is not a trivial task. Mantegna and Stanley characterise this difficulty in the following way (23): "The scales used are often given in units (currencies) that are themselves fluctuating in time and transactions occur at random times with random intensities." While price and volume can be called the *fundamental observables* in the system, there are some other quantities that often are of interest.

Price $P_i(\tau)$. For publicly traded stocks the stock (share) price is a security's last reported sale price on an exchange and it is determined on the market by buyers and sellers. Closing price is the price of the last transaction for a given security at the end of a trading session, which for stock i at time τ is denoted by $P_i(\tau)$.

Logarithmic return $r_i(\tau)$. Absolute prices are seldom used in studies since the investors obviously work in terms of relative returns, i.e., in terms of potential

gain or loss in proportion to the invested sum. Logarithmic returns are most commonly used, since they incorporate an average correction of scale changes and simply lump inflation together with other sources of steady compound growth. The daily logarithmic return of stock i is given by $r_i(\tau) = \ln P_i(\tau) - \ln P_i(\tau - 1)$.

Volume $V_i(\tau)$. Trading volume is the number of shares traded during a given time period for a security or an entire exchange. Sometimes the word *activity* is used interchangeably with volume. Aggregate daily volume, i.e., the cumulative volume over a trading day for stock i at time τ, is denoted with $V_i(\tau)$.

Relative change in volume $\dot{V}_i(\tau)$. Volume is not always a suitable variable because, just like price, its absolute level is arbitrary.[1] Therefore, in some cases it is more informative to use relative changes in volume, defined as $\dot{V}_i(\tau) = V_i(\tau)/V_i(\tau - 1)$. Relative volume changes can be used, for example, to study whether trading activities of different stocks are correlated. Absolute volumes may be misleading, for the aforementioned reasons, but also because large (valuable) companies have a larger absolute number of outstanding shares than small ones and, therefore, are expected to have larger absolute trading volumes as well.

Flow of capital $C_i(\tau)$. Sometimes the term *dollar volume* is used to indicate the dollar amount of shares traded during a given period, but it can be more generally called *monetary volume*. It can be used to characterise either an individual security or an entire stock exchange. For a stock i it becomes $C_i(\tau) = P_i(\tau)V_i(\tau)$. It is important to note that actual closing prices must be used together with actual volumes, or split-adjusted prices with split-adjusted volumes, but raw and split-adjusted data obviously cannot be combined.

Relative change in flow of capital $\dot{C}_i(\tau)$. It is also natural to make the capital flows independent of scale, and the most obvious way to do this is to study its relative change. Analogously to the relative change in volume, this can be defined as $\dot{C}_i(\tau) = C_i(\tau)/C_i(\tau - 1) = P_i(\tau)V_i(\tau)/[P_i(\tau - 1)V_i(\tau - 1)]$.

In general, the processes that influence the performance of stocks can be divided roughly into endogenous and exogenous processes. Endogenous processes are those that result from the internal dynamics of the financial market, such as competition amongst investors, while exogenous processes are those that influence the performance of the market or individual stocks from outside the system, such as changes in the global economic environment. Different stocks are prone to changes in different external factors, e.g., a company in the energy business

[1] When a company is incorporated, it is formed into a legal corporation by completing the required procedures. The articles of incorporation state how many shares are authorised and can be issued. When some proportion of the shares is sold to the public, for example via an initial public offering (IPO), the initial IPO price range is adjusted to match the volume of outstanding shares, i.e., the shares of a corporation's stock that have been issued and are in the hands of the public (24). Another way to look at this is through market capitalisation, which refers to a company's market price, calculated by multiplying the number of shares outstanding by the price of a share. Thus price and volume are related through market capitalisation but, apart from this constraint, their absolute values are arbitrary.

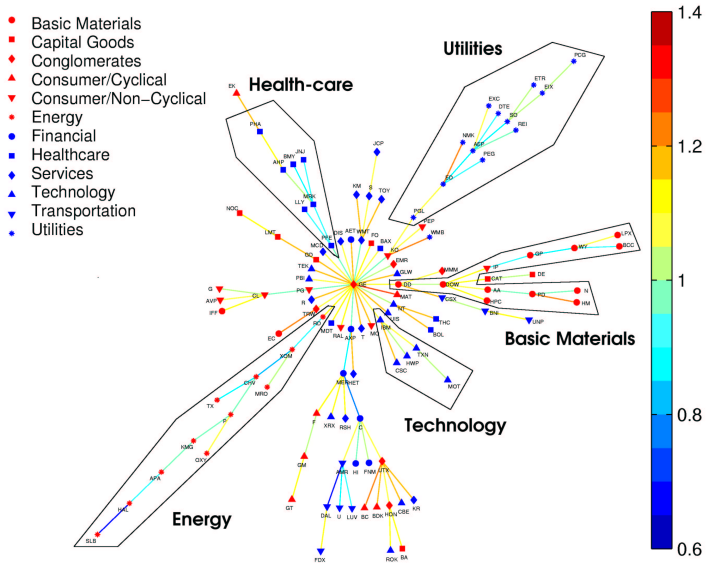

Figure 2.1: Example of a small financial network (asset tree) from (4). The markers correspond to stocks, and the business sectors assigned by Forbes are indicated by the type of markers used. The colour of the link reflects a distance between the stocks and it is scaled on the [0, 2] interval.

is likely to be affected by changes in the world market price of oil regardless of whether it is actually directly involved with the oil business, as its competitors still might be. This is one of the reasons why there is structure in the return correlations of the market (Figure 2.1), contributing to *non-systematic* or stock-specific risk. This is the part of the risk that may be diversified by holding a portfolio of stocks, subject to the assumption that the stock prices of different stocks are not fully correlated, an assumption that is well justified in practice.[2]

The financial market can also be viewed as a multi-agent game, in which traders repeatedly compete for limited resources. The resulting interactions between different market participants influence the performance of stocks, which is compactly characterised by a single number, the stock price. Although the exact nature of these interactions is not known, their manifestations in the market are visible. This is indeed the underlying rationale for the complex systems approach to financial markets: the fluctuations observed in financial time-series should reflect the

[2]The part of the risk that cannot be diversified away and needs to be accepted by the investor as the price of potential future returns is called the *systematic risk* associated with the market as a whole (24).

interactions, feedback, and adaptation of the market participants. Constructing a portfolio is an active market event as the stocks in the portfolio need to be purchased from the market. If a large fraction of investors decides to acquire the same portfolio, the demand for the portfolio stocks will increase and, consequently, so will their prices. This sequence of events will induce some correlations in the stock prices, resulting from correlated market actions. More generally, variations in the supply and demand of stocks, regardless of their origin, tend to introduce some correlated structure in the market. This endogenous structure can be seen to be superimposed with the exogenous structure.

So what does the structure of the market look like? How does it evolve? How robust is it? To answer these and other questions we consider a network representation of the market in which the nodes correspond to stocks and links to return correlation based distances between them (Section 3.2). The purpose is to grasp the essence of the market without drowning in the abundance of information, which requires pruning the system. In the complex networks framework, interactions have typically been considered to be binary in nature, meaning that two nodes either interact (are connected) or they do not (are not connected). In many cases imposing a binary interaction requires setting a threshold value for interaction strengths, such that interactions falling below the threshold are discarded. Although suitable as a first approach, thresholding can lead to a loss of information. Consequently, a natural step forward is to assign correlation based weights to the links, to reflect the strength of interaction, an approach pursued in this thesis.

The problem of portfolio optimisation is central to financial theory. The task is to find portfolio weights w_i for stocks, corresponding to the value invested in stock i out of the overall value of the portfolio, such that the portfolio risk is minimised for a given portfolio return. In the classic *Markowitz portfolio optimisation scheme* risk is quantified by the standard deviation of returns (25), but other alternatives also exist (26). The expected mean and variance of portfolio return r_P can be written as

$$\begin{aligned} \mu_P = \quad E(r_P) \quad &= \sum_{i=1}^{N} w_i \mu_i \\ \sigma_p^2 = \quad \mathrm{Var}(r_p) \quad &= \mathrm{Var}(w_1 r_1 + \cdots + w_N r_N) \\ &= \sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{j=1}^{N} \sum_{i=1, i\neq j}^{N} w_i w_j \sigma_i \sigma_j \rho_{ij}, \end{aligned} \tag{2.1}$$

where μ_i is the expected return of stock i, r_i and σ_i^2 denote the return and standard deviation of return of stock i, respectively, and ρ_{ij} denotes the correlation of re-

turns r_i and r_j. In fact, the above holds as an identity for any random variables r_i, but the Markowitz framework gives this expression an interpretation in the portfolio context: since variance can be identified with risk, the higher the variance of portfolio returns, the riskier the portfolio. We see in the above equation that the variance of portfolio returns depends on the correlation structure of the market, making understanding this structure extremely important. Even if one does not use variance as a measure of risk and ignores the Markowitz framework, one may still wish to characterise the portfolio return distribution by its variance, or characterise interdependencies on the market using correlations. These considerations underline the importance of understanding market correlations. Against this background one might expect that networks constructed from correlation based interactions might have a relation to Markowitz portfolio optimisation. This is indeed the case, as the results of Publications (1; 4) show (Section 4).

2.3 Social systems

Social science attempts to explore and analyse the relation between the characteristics of individuals and the characteristics of the social systems they comprise. Its aim is both to understand how individual behaviours translate into large-scale social systems as well as how large-scale observation can be used to understand the behaviour of individuals (27). Social structures can be studied using social networks analysis, which is an important technique in modern sociology, anthropology, social psychology, and organisational studies. According to the network paradigm, the social world consists of a web of interactions and relationships channelling information and resources of various kinds among social actors. Thus, social life consists of the flow and exchange of norms, values, ideas, and other social and cultural resources (28). In this paradigm, social action and outcomes are affected by the structure of the relationship network (29).

Social network analysis has been used since the mid-1930s to advance research in the social and behavioural sciences, although their progress was relatively slow (30). The coining of the social networks concept is credited to J. Barnes in the 1950s (31). Activity in this field increased among social scientists in about 1990 when the use of social network methodology began to grow at a more rapid rate. This was related to the understanding in behavioural sciences that the social contexts of actions matter. For example, epidemiologists realised that epidemics do not progress uniformly through populations, sex researchers started considering sexual networks, and organisational studies were recognised as being at the heart of management research (30). Perhaps most importantly, network models have helped to understand how individual behaviours and interactions translate into macro-level social systems. Some of the classic notions of sociology in network theory include the *six-degrees of separation* and the *strength of weak*

ties (32; 33). Nowadays there are several extensive reference texts available on social network theory (34; 35; 30).

Towards the end of the 1990s physicists also started taking an interest in social networks. Much of the interest was sparked by the so-called small-world network model which captured clustering and short path lengths, considered to be the two key structural characteristics of social networks (36). The key insight behind the small-world model was that empirically found networks interpolate between two extremes, order and randomness. This seminal paper was released at the time when a solid theoretical framework for dealing with very large networks in general, and very large social networks in particular, was rapidly being developed. More recently, network scientists have have studied the structure of different social networks based on scientific co-citation and collaboration (37; 38; 39; 40), email traffic (41; 42), and human sexual activity (43), and there have been efforts to uncover the underlying community structure of social networks (44). Some simple models have been developed to mimic the structure of social networks (45; 46; 47; 48; 9), address network searchability in terms of identities (49), and find the mechanisms responsible for the structure of collaboration networks in creative enterprises (50).

Social networks can be viewed at many different levels of coarse-graining, from individuals to nations (Figure 2.2). Here we consider networks in which nodes correspond to individuals and links correspond to social familiarities or social interactions between them. These are characterised by a set of properties, covered in Section 3.1: (1) the connectivity distribution shows a broad, power-law like tail, (2) clustering is high, (3) assortativity (highly connected individuals tend to be connected to other highly connected individuals and vice versa) at least up to some saturation connectivity, and (4) the average shortest paths connecting individuals in the networks are relatively short.

While the study of social networks has to start from empirical observations using real life data, the natural next step is to model both processes responsible for constructing the networks themselves and the processes taking place on them. A simple phenomenological model that mimics the structure of different kinds of social networks is presented in Publication (9). While this is a good starting point, it is necessary in the future to go beyond phenomenology and consider the actual microscopic mechanisms that give rise to the kinds of social networks we see in reality. Although there is variation between the behaviour and social preferences of individuals, it should, in principle, be possible to capture the essential elements which determine the connectivity patterns of individuals and their social groups in mathematical terms.

While the last few years have witnessed a surge of activity in studies related

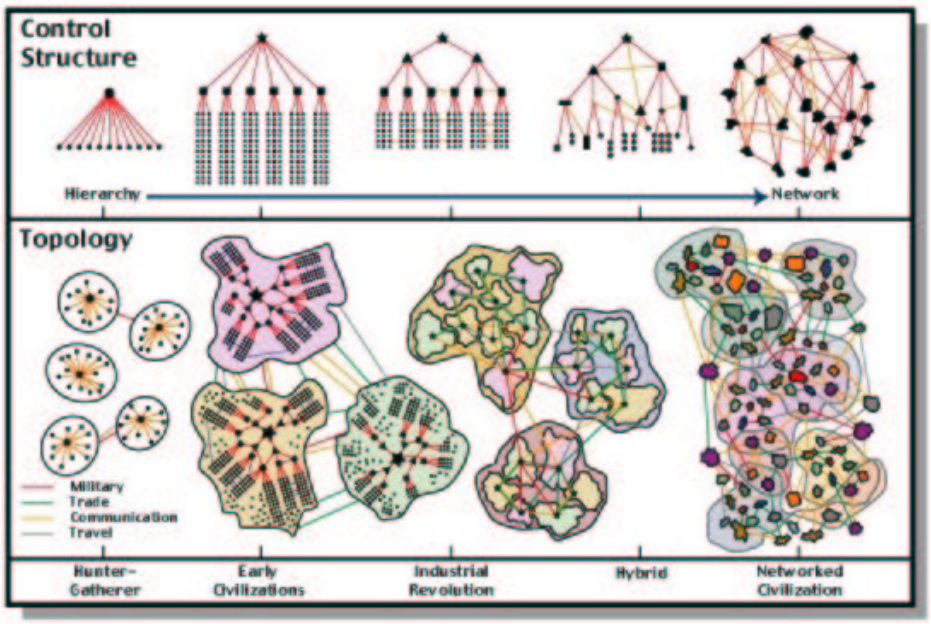

Figure 2.2: Human civilisation as a complex system. Figure from the website of New England Complex Systems Institute (NECSI) at `www.necsi.org`.

to the structure of different networks, relatively little work has been done on processes unfolding on networks. This is understandable as we simply have not had large-scale data available, nor the computational power required for studying these systems. Examples of dynamic processes taking place on social networks are social diffusion, i.e., the spreading of rumours and information, opinion formation models, and models of cultural invasion (51). In general, the structural properties of the network strongly influence the dynamics of processes unfolding on it. The role of topology on network dynamics may be further enhanced if the dynamic process is sensitive to interaction strengths, which in the case of social networks may be partially driven by the network structure around the tie (17).

An excellent popular level account of the application of tools and concepts from modern physics to understanding collective human behaviour is given in the acclaimed book by Ball (52). He demonstrates how the application of physics models can yield very reliable predictions for large-group outcomes without negating the individual's free will.

Chapter 3

Analysis and modelling methods

Physics knows many phenomena in which the whole can be described by describing the parts. For example, we understand from statistical mechanics how magnetism emerges from the collective behaviour of millions of microscopic magnets, or spins. One of the reasons why this approach has been so successful lies in the simplicity of interactions. In physical systems, there is no ambiguity as to what interacts with what, and the interaction strength is uniquely determined by the physical distance (53). In complex systems this is not the case, and this ambiguity makes them more difficult to analyse than physical systems. This Chapter first gives an overview of the complex network approach to complex systems, and then reviews some of the methods developed and used in the Publications.

3.1 Complex networks

Background

Complex networks are a powerful fairly recent framework for describing, analysing, and modelling several complex systems found in nature and society. In the network representation of a complex system one focuses only on the essentials, namely, the elements represented by nodes (vertices), and the interactions between the elements represented by links (edges). For example, the cell can be described as a complex network of chemicals connected by chemical reactions, the Internet as a complex network of routers and computers linked together by physical connections, and society as a network of individuals connected by various kinds of social relationships. Some networks, such as highways and neural networks, are embedded in a metric space, whereas others, such as collaboration, social, and correlation networks, are defined in an abstract space.

Perhaps the key reason for the success of the complex networks approach lies in its simplicity; one has to omit the complicated details and focus on an abstract

system-level description by identifying the parts and describing the interactions between them. This often entails coarse-graining the system and forces one to adopt a perspective that is likely to be fruitful in the study of such systems. However, what is the greatest strength of the complex networks approach to complex systems might also be its greatest weakness: once one adopts the complex networks approach, almost everything starts to look like a network. This may be a problem because, although many systems look like networks, the network approach will not necessarily be fruitful in studying them. This is exacerbated by the fact that some ideas in the complex network literature seem deceptively simple, although looking below the surface proves otherwise. But this is not perhaps so much a problem of the paradigm itself but that of the practitioner who decides to apply the paradigm to a certain problem. Therefore, adopting the network view calls for a critical evaluation of the problem and the ability of the framework to contribute to it. The reason for vocalising this criticism is not to downplay the complex network paradigm nor the remarkable results that have been obtained with it. Instead, it is intended to promote critical discussion on the scope of complex networks in the study of complex systems.

This thesis studies financial and social systems from the perspective of complex networks. Since we are interested in universal aspects of these systems, the specific questions covered do not have to be the issues of current importance in the study of these specific systems[1]. Rather, the purpose is to motivate a question of interest in the context of a particular complex system, and then step back and see if the method of study has relevance to other complex systems (10). It turns out that this is indeed the case. For example, the machinery for dealing with weighted motifs can be developed in the context of biological networks and then brought to the context of social networks.

Complex networks have traditionally been the domain of graph theory, a branch of discrete mathematics. The birth of graph theory is usually accredited to the Swiss mathematician Leonhard Euler who in 1736 published the solution to the famous Koenigsberg bridge problem. Initially, graph theory focused on the properties of regular (non-random) graphs and had contributions from several well-known mathematicians. An important development in the history of graph theory took place in the 1950s when the Hungarian mathematicians Paul Erdös and Alfred Rényi initiated the study of random graphs, giving rise to the Erdös-Rényi (ER) model (54; 55). In the model one starts with N nodes, connecting every pair of nodes with probability p, which creates a graph that has on average $pN(N-1)/2$ edges distributed randomly. This deceptively simple model was studied analytically, first by Erdös and Rényi with the main goal of determining the connection probability p at which a particular property of the graph will most likely

[1] Having said that, as this work has progressed practitioners from both financial and social systems have taken an interest in it, and also found it to be of contemporary importance.

arise. Since Erdös and Rényi, the model has been most notably studied by Bollobás (56), and the basic results of the model are included in many graph theory textbooks (56; 57; 58; 59).

The increased computing power and computerisation of data acquisition lead to increased availability of data on complex systems, giving rise to a revived interest in networks in the late 1990s. This surge of interest was driven predominantly by physicists, leading to the paradigm of complex networks. Although social scientists had used networks for describing social systems for a long time (34), physicists provided a significant conceptual and theoretical augmentation and formalisation of the framework, relying on tools from statistical mechanics, and started using networks not only for characterising social networks but for studying complex systems in general. The flurry of activity from the physics community was triggered by two seminal papers. The first one, on small-world networks by Watts and Strogatz, was published in Nature in 1998 (36), followed by another on scale-free networks by Barabási and Albert, published in Science a year later (60) [2]. Since then, interest in complex networks has grown phenomenally and has produced a large body of knowledge on both empirical and model networks and has made significant theoretical and conceptual advances. It is out of the scope of this thesis to cover these developments and that would hardly be meaningful since there are some excellent review articles on complex networks that, in combination, more or less cover the field (53; 61; 62; 63; 64), as well as some books intended for the professional (65; 66; 67) and nonprofessional audience (68; 69; 70; 71). The remaining part of this Section will provide an elementary level introduction to some of the most important concepts and characteristics used in the study of complex networks, but will only touch the surface of this incredibly interesting field.

Basic characteristics

Comparative analysis of networks from different fields has yielded unexpected and, in some cases, dramatic results. The conclusion always seemed to be the same: the networks found in nature are not ER random networks. The empirical networks were characterised using some elementary concepts from traditional graph theory but, as the need for differentiating between different networks became evident, additional characteristics were developed to better capture some of the more subtle structural properties of networks. The following discussion outlines some of the main structural properties for studying real networks. The position of the ER model with respect to these characteristics is addressed and, even though the model does not capture the properties of real networks, it is still

[2] According to ISI Web of Knowledge, at the time of writing on April 7th, 2006, these papers had 1466 and 1274 citations, respectively.

often used as a basic reference, or null hypothesis, in judging how far a particular network is from randomness. The list below can also be seen as a "wish list" or motivation for the properties that could be incorporated in more realistic and advanced networks models. In a sense the list also entails one of the main results of complex network theory, namely, that using the framework it is possible to identify some unifying principles and structural properties common to most real networks, but it also gives an indication of what may be missed in using a network description of complex systems.

Degree distribution $P(k)$. Node degree k_i is defined as the number of links adjacent to node v_i. For an ER graph with N nodes and n links, the degree distribution is a Poisson distribution of the form $P(k) = e^{-\langle k \rangle}\langle k \rangle^k / k!$, where the average degree is given by $\langle k \rangle = 2n/N = p(N-1) \approx pN$. In contrast, in many empirical networks the degree distribution either follows a power-law or has a power-law tail of the form $P(k) \sim k^{-\gamma}$, thus very different from the Poisson distribution (53). The average degree is now given by $\langle k \rangle \sim k_{\max}^{2-\gamma}$, where the maximum degree $k_{\max} < N$. Perhaps the best known power-law degree distribution is that of the World Wide Web, in which the nodes are the documents and the links are the hyperlinks between them. Here the scaling holds for several orders of magnitude with two different exponents $\gamma_{\text{in}} \approx 2.1$ and $\gamma_{\text{out}} \approx 2.45$ for incoming and outgoing links, respectively (72).

Degree-degree correlations $P(k, k')$. Degree correlations refer to the idea that the probability that a node of degree k is connected to another node of degree k' depends on the value of k or, put differently, the degrees of two adjacent nodes are not independent. In principle, this phenomenon is completely described by the joint probability distribution $P(k, k')$, giving the probability that a node of degree k is connected to a node of degree k'. However, the empirical evaluation of $P(k, k')$ is cumbersome due to limited data and often yields noisy results. Hence, it is more practical to define the average nearest neighbours degree of a node v_i as $k_{nn,i} = \frac{1}{k_i} \sum_{j \in \mathcal{N}(v_i)} k_j$, where $\mathcal{N}(v_i)$ denotes the neighbourhood of v_i. From this one can calculate the average degree of nearest neighbours with degree k denoted with $\bar{k}_{nn}(k)$, which corresponds to $\sum_{k'} k' P(k'|k)$ (73). ER graphs have no degree-degree correlations, but many real networks do. The network is said to exhibit *assortative mixing* if $\bar{k}_{nn}(k)$ increases as a function of k and *disassortative mixing* if it decreases as a function of k. Social networks are usually found to be assortatively mixed (74).

Clustering coefficient C. The clustering coefficient for a node v_i quantifies the local cliquishness of its immediate (nearest) neighbourhood, defined as $C(v_i) = 2t_i/[k_i(k_i - 1)]$, where t_i is the number of triangles attached to the node v_i and k_i is its degree (36). In words, the clustering coefficient is the number of existing connections between the k_i neighbours of v_i divided by the maximum possible number of connections that could exist between them, normalising the clustering coefficient between 0 and 1. In an ER graph the probability for two

neighbours of a node to be connected is the same as the probability of any two arbitrarily chosen nodes to be connected and, consequently, $C_{\mathrm{ER}} = p = \langle k \rangle / N$. If we keep the average degree $\langle k \rangle$ fixed, we see that in ER networks $C/\langle k \rangle$ scales as $1/N$ so that the ratio decreases as N increases. If we plot this ratio for empirical networks, it appears to be independent of N, which clearly deviates from the ER graph prediction. Most complex networks have higher average clustering coefficient than ER graphs. A good example of clustering is found in social networks, which is the context in which the coefficient was initially proposed (36; 34).

Network motifs. Network motifs were introduced as elementary structures that are repeated frequently throughout the network, and they are considered to have some elementary function for the system. The standard approach to motifs considers them merely as topological building blocks (75; 76; 77). The author of the thesis, together with collaborators, were the first to augment the motif framework to incorporate interaction strengths instead of considering all interactions homogeneous. The motif framework can be seen as a generalisation of the concept of clustering, whereas communities can be seen as a generalisation of motifs. Communities are important for the function of systems and identifying them may help us understand the function of networks. For example, communities in the social network correspond to circles of friends, in the WWW communities correspond to pages on the same or closely related topic, and communities in cellular and genetic networks are related to functional modules. The way motifs are defined in (6) allows a generalisation of the clustering coefficient into a weighted clustering coefficient, and the concepts of intensity and coherence are immediately applicable to communities as well. Motifs will be covered in Section 3.3. For now it suffices to say that some motifs are also created by chance in ER graphs, but in real networks some motifs can be several orders of magnitude more frequent than in a comparable ER graph.

Path length ℓ. Most complex networks exhibit the small-world property, i.e., the shortest path between any two nodes is relatively short. The path length can be seen as characterising the global topology of the network or, more specifically, how spread out the network is. In ER graphs the typical distance between any two nodes scales as the logarithm of the number of nodes so that $\ell_{\mathrm{ER}} \sim \ln(N)/\ln(\langle k \rangle)$ and, in fact, path lengths in real networks are close to that of ER graphs of the same size. Thus, most empirical networks seem to exhibit the small-world property. Regular lattices are, in a sense, the opposite of small-world networks since, for example, for a d-dimensional hypercubic lattice the average node-node distance scales as $N^{1/d}$, which increases much faster than $\ln(N)$. However, introducing a small number of shortcuts, either by rewiring part of the lattice or by adding some links to it, dramatically decreases path lengths while leaving clustering close to its high initial value. This has the implication that a transition from a regular lattice to a small-world is not visible locally to the nodes of the network in terms of the clustering coefficient; this is one of beautiful ideas introduced in (36).

Role of network models

Measurements of the above characteristics in empirical networks convinced network scientists time after time that ER graphs were not adequate for capturing the prominent structural properties of real networks. This lead to a revitalisation of network modelling with the aim of mimicking the formation mechanisms of empirical networks in order to reproduce their structural properties. The underlying philosophy of this approach is based on the assumption that the observed network structure is the result of some evolutionary mechanisms. Therefore, modelling networks might reveal what these mechanisms are and lead to a better understanding of not only their structure but also their function.

It is illustrative to consider the modelling philosophy through a simple example. It is possible to construct scale-free networks or, for that matter, networks with completely arbitrary degree distributions purely phenomenologically using a suitable algorithm or a generalised random graph model (78). We then obtain a realisation of the network with the desired degree distribution, but this does not get us any closer to understanding how scale-free networks might emerge in nature. The success of the Barabási-Albert (BA) model (60) in producing scale-free networks was rooted in the realisation that two generic mechanisms are required simultaneously for these networks to emerge, and they are generally considered to be the principal mechanisms responsible for the ubiquity of scale-free networks. The mechanisms are (1) *growth* and (2) *preferential attachment*. Here (1) refers to the idea that a new node is added at each time step to the network, and (2) states that the probability for a new node to connect to an existing node v_i depends on the degree k_i as $\Pi(k_i) = k_i / \sum_j k_j$. An inspection of the latter mechanism reveals that the preferential attachment contains two assumptions. First, that $\Pi(k)$ depends on k unlike for, say, ER graphs that have $\Pi(k) = p$ and, second, that the functional form of $\Pi(k)$ is linear in k. We can write the probability more generally as $\Pi(k) \sim k^\alpha$ and it turns out that non-linear preferential attachment ($\alpha \neq 1$) destroys the scale-free nature of the topology, and the resulting topology is different for the sublinear ($\alpha < 1$) and superlinear ($\alpha > 1$) cases. In the case of scale-free networks identification of these two evolutionary mechanisms (growth and preferential attachment) increases our understanding of how such networks might be generated in nature. However, the mechanisms may not be universally valid as the assumption of global preferential attachment is questionable in some cases, such as in the context of social networks. In this case a more plausible mechanism might be one based on short local random walks. This process does not explicitly incorporate linear global preferential attachment but it follows implicitly from the random walk itself (79; 80). This beautiful idea is really just a formalisation of the old saying “all roads lead to Rome”. In this context all links eventually lead to hubs and, therefore, walking on a network leads to preferential attachment.

Beyond topology

The above discussion has focused on topological characterisation of networks. Topology is relevant because it is related to the function and dynamics of the network and affects the ability of the network to perform its function. For example, topology is crucial for network robustness and how it responds to external perturbations, such as random failures and targeted attacks on nodes (81; 82; 83; 84; 85; 86; 87). Network topology also plays an important role in determining the emergence and nature of collective dynamics, such as the behaviour of different spreading processes that use the network as a platform, whether it be the spread of epidemics, information, or rumours (88; 89; 90; 91; 92; 93; 94). Network structure matters also from the point of view of synchronisation. For instance, there is evidence supporting the notion that some brain diseases result from abnormal synchronisation of a large number of neural populations, placing investigation of the network mechanisms involved in the generation, maintenance, and propagation of epileptic disorders at the forefront of neuroscience (64). Synchronisation is also relevant, in sociology and social psychology, in understanding the emergence of collective social behaviour such as the emergence of new habits, fashions, or leading opinions (64).

As the complex network paradigm has matured in the past few years, the view that the framework should be augmented beyond topological considerations has been increasingly voiced. One such augmentation is the movement from binary to weighted networks that allow the heterogeneity of links to be taken into account. One of the motivations for this development comes from social networks, where it is believed that weak and strong ties have functionally different roles in the network (33; 17). Some other motivations are discussed in Section 3.3. All the Publications included in this thesis except for (9) deal with weighted networks. The tools developed in Publications (6; 7) augment the framework of clustering and motifs to weighted networks, and may be applied to the study of communities as well, representing another major and fairly new focal point of research in the field. The role of communities is also addressed in Publication (9) in which a model mimicking social networks is introduced. The mechanism for generating communities also suggests how they could be identified in real complex networks using manageable and fast algorithms, giving rise to a line of research that is pursued further elsewhere (17; 95). Fortuitously, the measures introduced in Publication (6) are immediately generalisable to weighted communities.

Navigation in the absence of information on the global network structure is currently also an important topic. The social network model introduced in Publication (9) paves the way for this type of research by establishing a platform on which different social navigation processes may easily be studied. Also, the effect of inaccurate or even erroneous information on navigation can be studied using the model. Further, the model enables the study of synchronisation and

spreading processes with a realistic network topology, and it can be complemented with insights gained from studies dealing with empirical social networks (17). Finally, there is growing interest in the network community in evolving and adapting networks, such that the topology is not fixed but evolves in time. Examples of such networks are genetic regulatory networks, ecosystems and so-called ad-hoc networks. Networks describing the structural evolution of the financial market are yet another example of adapting networks, and are the topic of Publications (1; 2; 3; 4; 5).

3.2 Exploratory analysis using trees and graphs

This Section deals with methods aimed at characterising the top-level structure and dynamics of interacting systems consisting of a large number of elements. Instead of looking at the interactions directly, our perspective is a bit more abstract in focusing on the correlation patterns of interactions. The methods presented here were developed in the context of the financial market to study the hierarchy of stocks, to classify them into categories, and to monitor the evolution of the market. However, the methods are applicable, at least in principle, to any complex system in which the interactions are reflected in temporal correlations of some relevant quantity attached to each node. In a sense, the process of constructing tree or graph representations can be seen as a non-linear transformation of the underlying correlations that filters out the irrelevant parts. The resulting simplified system is more amenable to analysis, and a number of different measures may be developed to characterise it.

This type of filtering is especially important for correlation-based networks, which use correlation coefficients as a simple measure of linear coupling. Unfortunately, as studies based on random matrix theory (RMT) have shown, the information content of such matrices is limited due to noise (96; 97). The main finding is that 94% of eigenvalues of correlation matrices computed from logarithmic returns of stocks lie within the predictions of the RMT and, thus, would be reproduced by a random matrix without any structure embedded in it. This leaves only 6% of eigenvalues standing out from randomness, suggesting that only 6% of eigenvalues may carry information. This example, based on one particular application, demonstrates the need for developing methods for separating the important from the unimportant in such systems.

Assume that vertex v_i has some quantity $x_i(t)$ attached to it, and that this quantity fluctuates as a function of time. To characterise the synchronous time evolution of the system, we use the equal time correlation coefficients between

vertices v_i and v_j defined as

$$\rho_{ij}^t = \frac{\langle \boldsymbol{x}_i^t \boldsymbol{x}_j^t \rangle - \langle \boldsymbol{x}_i^t \rangle \langle \boldsymbol{x}_j^t \rangle}{\sqrt{[\langle \boldsymbol{x}_i^{t\,2} \rangle - \langle \boldsymbol{x}_i^t \rangle^2][\langle \boldsymbol{x}_j^{t\,2} \rangle - \langle \boldsymbol{x}_j^t \rangle^2]}}, \tag{3.1}$$

where the vector $\boldsymbol{x}_i^t = [x_i(1), \ldots, x_i(T)]$ contains the values of the observables for vertex v_i in a time window of width T, the superscript in $\boldsymbol{x}_i^t$ corresponds to the index of the window, and $\langle \cdot \rangle$ indicates a time average over the elements in the vectors. For the purpose of constructing trees and graphs, we define a correlation-based distance d_{ij} that is associated with the edge connecting vertices v_i and v_j and reflects the level at which the quantities $x_i(t)$ and $x_j(t)$ are correlated. We use a simple non-linear transformation $d_{ij}^t = \sqrt{2(1 - \rho_{ij}^t)}$ to obtain distances with the property $2 \geq d_{ij} \geq 0$, forming an $N \times N$ symmetric distance matrix $\mathbf{D}^t$. Now two alternative approaches may be adopted. The first one leads to trees and the second one to graphs. In both approaches the trees (or graphs) for different time windows may not be independent of each other, but form a sequence of trees (graphs) through time. Consequently, this multitude of trees (graphs) is interpreted as a series of evolutionary steps of a single tree (graph).

Tree approach. In this approach we construct a tree according to the method introduced by Mantegna in studying the taxonomy of the financial market (98). The approach requires an additional hypothesis about the topology of the metric space, the so-called ultrametricity hypothesis. In practice, it leads to determining the minimum spanning tree (MST) of the distances, denoted by $\mathbf{T}^t$. The spanning tree is a simply connected acyclic (no cycles) graph that connects all N nodes (stocks) with $N - 1$ edges such that the sum of the included edge weights $\sum_{d_{ij}^t \in \mathbf{T}^t} d_{ij}^t$ is minimum. We refer to the minimum spanning tree at time t by the notation $\mathbf{T}^t = (V, E^t)$, where V is a set of vertices and E^t is a corresponding set of edges. Since the spanning tree criterion requires all N nodes to be always present, the set of vertices V is time independent, which is why the time superscript has been dropped from the notation. The set of edges E^t, however, does depend on time, as it is expected that edge lengths in the matrix $\mathbf{D}^t$ evolve over time and, thus, different edges are included in the tree at different times. The resulting tree is called an *asset tree* in the Publications included in this thesis.

Graph approach. In this approach we construct graphs by extracting the $N(N-1)/2$ distinct distance elements from the upper (or lower) triangular part of the distance matrix $\mathbf{D}^t$ and obtain a sequence of edges $d_1^t, d_2^t, \ldots, d_{N(N-1)/2}^t$, where we have used a single index notation. The edges are then sorted in a non-decreasing order and form an ordered sequence $d_{(1)}^t, d_{(2)}^t, \ldots, d_{(N(N-1)/2)}^t$. Since we require the graph to be representative of the market, a natural heuristic is to build the graph by including only the strongest connections. The number of edges to be included is arbitrary. Here we include only $N - 1$ shortest edges in the

graph, thus giving $E^t = \{d^t_{(1)}, d^t_{(2)}, \ldots, d^t_{(N-1)}\}$. This is motivated by the fact that the tree also consists of $N-1$ edges, and this choice renders the two methodologies comparable, and possibly even similar to one another. This method of constructing graphs defines them uniquely and, consequently, no additional hypotheses about a metric is required. It is important to note that both the set of vertices V^t and the set of edges E^t are time dependent, and thus we denote the graph with $\mathbf{G}^t = (V^t, E^t)$. Now even a small set of vertices may be strongly inter-connected and, thus, may use up many of the available edges, which may lead to the formation of cycles in the graph. The resulting graph is called an *asset graph* in the Publications included in this thesis.

Instead of fixing the number of links n at $N-1$, we can consider it a parameter and increase it all the way up to $n = N(N-1)/2$, resulting in a fully connected graph. If $d_{(n)}$ is the last added edge, where $n = 1, 2, \ldots, N(N-1)/2$, we can quantify the degree of graph completeness by $p = n/[N(N-1)/2]$, where $p \in [0, 1]$. Contrasting this with ER graphs with n edges chosen out of the possible $N(N-1)/2$, we can view p as the probability of connecting two randomly chosen vertices or, alternatively, as the probability of a given link being present in the network. Therefore, a natural reference for the graph approach is to consider the corresponding ER graph with $p = n/[N(N-1)/2]$, where n is, again, the number of links in the empirical graph. Obviously, the value of p is arbitrary and varying its value will yield different kinds of networks. This mapping between graph completeness and the ER connection probability suggests that one could compare not just two graphs at a fixed value of p, but instead vary p over a range of values from 0 to 1 and compare the empirical graph to the corresponding ER graph (or an ensemble of such graphs) with the same p-value (5). In the mathematical literature, constructing a random graph from a set of isolated vertices by successive additions of links is called *graph evolution*. The same process done for a weighted network, following the graph approach as described above, is called *thresholding*. Section 3.4 demonstrates how differences in growth patterns as $p = 0 \to 1$ are informative about the structure of the system.

The method for constructing trees and graphs is presented schematically in Figure 3.1 and examples based on financial data from S&P 500 are shown in Figure 3.2, for which we used $x_i(t) = r_i(t)$, i.e., the daily logarithmic return. The tree criterion suppresses the formation of loops and clustering and, based on these plots, it seems that important structural properties may be lost. In this sense the tree approach can be seen as a very drastic filtering of correlation-based networks. Indeed, the tree approach has been expanded by incorporating the idea that graphs with different degrees of complexity can be constructed by iteratively linking the most strongly connected nodes under the constraint that the resulting graph can

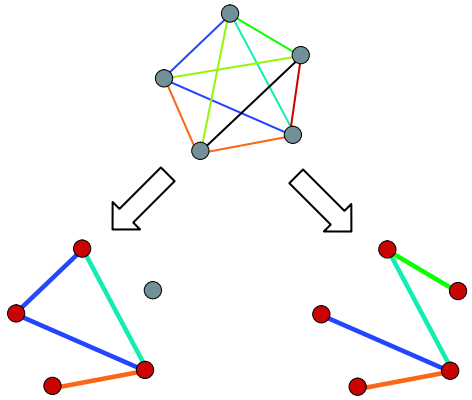

Figure 3.1: Schematic presentation of the tree and graph approach. Both use the fully connected graph (top) as a starting point, from which a large fraction of links are typically pruned. The graph approach (bottom left) may leave some nodes isolated and allows the formation of loops. The tree approach (bottom right) guarantees that all nodes are connected but, by virtue of being a tree, does not allow loops.

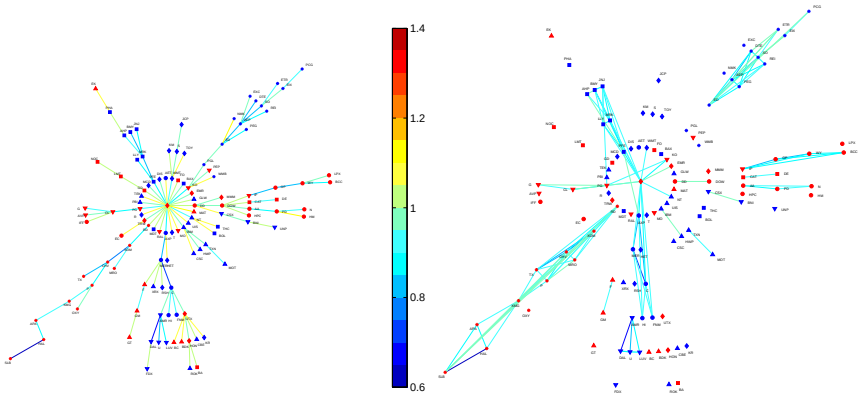

Figure 3.2: Example of the tree and graph approach (left and right, respectively). Both contain $N = 116$ nodes and $N - 1 = 115$ links, but the configuration of links is clearly very different in the two.

be embedded on a surface of a given genus (99). This method is therefore similar to the graph approach above, except that it produces a connected network, which rarely happens with the graph approach. Whether this is a desirable property or not is unclear. Indeed, the graph approach, in its extreme simplicity, is very transparent, whereas it is not so transparent what mapping a network on the surface of a genus really does. As a final remark, the methods presented here can also be viewed as exploratory data analysis tools for visualising and uncovering the structure of high dimensional data and, in this sense, serve the same purpose as principal component analysis (PCA) and self-organising maps (SOM) (100; 101).

3.3 Network characterisation with weighted motifs

Unweighted motifs

The study of structural properties of networks usually starts from global network characteristics, such as degree distribution $p(k)$ and average shortest path length $\langle \ell \rangle$. Global characteristics are useful because they can be employed for classifying networks in broad terms. If it turns out, for example, that a network has scale-free topology, we immediately have some idea about its error and attack tolerance. Similarly, if the network exhibits the small-world property, we know that spreading and diffusion processes will cover ground easily.

In some cases it is desirable to go beyond these global characteristics towards a more detailed and local analysis of the network. This can be accomplished using motifs. The concept of a motif was originally introduced to denote "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomised networks" (75). Put differently, motifs are elementary structures that are repeated frequently throughout the network and are considered to have some some elementary function for the system. An example of this is found in sensory transcription networks that control gene expression in bacteria and yeast in response to external stimuli. Here the nodes represent genes or operons and the edges represent direct transcriptional regulation. In such networks a motif termed "feed-forward loop" is repeated frequently, and it has been shown, theoretically and experimentally, to perform signal-processing tasks, such as pulse generation (75).

Not all subgraphs are informative, as some will be created just by chance (Section 3.4). Consequently, a measure that takes this into account is needed. To study the frequency of subgraphs, one counts the number of times a given subgraph, such as a triangle, appears in the network. The absolute counts are meaningless as such, however, since they will depend, among other things, on the size of the network. To draw statistical conclusions about the appearance frequency of subgraphs, one needs to specify a random reference, which, in the language of statistical hypothesis testing, can be seen as setting up a null hypothesis H_0. One

can now compare the appearance frequency of a given subgraph in the empirical and the random reference. In practice, one will need to consider an ensemble of random networks, not just a single realisation. A suitable test statistic for studying the statistical significance of unweighted motifs was defined in (77) as

$$z_M = \frac{n_M^E - \langle n_M^R \rangle}{\sigma_M^R} \tag{3.2}$$

where n_M^E is the number of subgraphs M in the empirical network, and $\langle n_M^R \rangle$ and σ_M^R are the expectation and standard deviation, respectively, of the number of subgraphs M in the reference ensemble.

The results will, of course, depend on what kind of reference system is chosen. Usually, in the case of unweighted motifs, one carries out a random topological rewiring of the network, which removes all structural correlations while conserving the degree distribution. The rationale of this approach is to keep the global connectivity unchanged and see what type of local configurations will be produced in the random ensemble. In principle, one could set additional global criteria for the reference system and require, for example, that the clustering spectrum $c(k)$ be conserved. As one sets tighter conditions for different global characteristics, the number of possible local configurations gets smaller and smaller. In practice, however, it is difficult to implement reference ensembles that meet a number of different conditions. One approach is to define a Hamiltonian, an energy function, for the system and use Monte Carlo methods by iteratively generating a large number of trial configurations and accepting only those that decrease the energy of the system (75). Although this would be computationally heavy for large systems, it is a feasible alternative for small systems.

Irrespective of the practical implementation of the reference ensemble, this line of thought presents a conceptual advance and an analogy to statistical physics. The ensemble theory of statistical mechanics considers microcanonical, canonical, and grand canonical ensembles. In the microcanonical ensemble all members of the ensemble have the same energy, to within a small tolerance, and the number of particles is fixed. By relaxing the requirement that the energy be fixed and allowing the system to exchange energy with a heat reservoir, we arrive at the canonical ensemble. The grand canonical ensemble is built upon the canonical ensemble by relaxing the restriction to a definite number of particles (102). This means that in moving from microcanonical to grand canonical ensemble we are increasing the degrees of freedom in the system and thus allowing a larger number of microscopic states. Expressed differently, the system occupies a greater volume in its phase space. In the case of complex networks, by fixing the degree distribution we are imposing a limit on the possible "microstates" accessible to the system and, by introducing additional restrictions, we further decrease the number of possible states.

How the reference ensemble is specified is not just a theoretical curiosity, because it affects the strength of conclusions that may be drawn from the study. For example, if one were to set the null hypothesis that the network is an ER random graph, then one could, at most, conclude that the empirical network is or is not an ER graph. This may serve a purpose in cases in which there is no *a priori* information about the topological organisation of the system. However, if we know, to begin with, that the empirical network is far from an ER graph, the results of such comparisons have little scientific value.

Weighted motifs

The motifs in the above discussion were considered topological building blocks of networks, without any regard for the link weights within them. In what follows we augment the motif framework to incorporate weights reflecting interaction strengths between the constituent elements. The motivation for this has to do with the nature of the coupling between links weights and local network topology. If there existed a universal coupling between weights and topology, including weights in the analysis of networked systems would not provide us with any additional insight. For example, if every triangle in every network was always made up of strong links, adding weights to the analysis would make little difference. Fortunately, the increasing availability of weighted data sets enables us to study the coupling between interaction topology and interaction strengths, and this picture can be complemented with some insights emerging from theoretical work. It appears not only that the nature of this coupling is not universal, but also that the coupling is related to the function of the network. Consequently, augmenting the motif framework by incorporating interaction strengths can improve our understanding of networked systems, and is a worthwhile pursuit. In addition to developing the theoretical framework, the aim has been to provide practical tools for characterising weighted motifs (6).

Let us make the discussion more concrete with an example. There exists some empirical and theoretical work supporting the notion that link weights in transportation networks, reflecting their capacity for transmission, are proportional to the *betweenness centrality* of links, defined as the number (or fraction) of shortest paths in the network passing through a given link (103; 87). In such a global transportation network we would expect few triangles and the existing ones should have low weights. At the other extreme we have systems like social networks, in which the link weights are inversely proportional to the betweenness centrality of links, and the system appears to be better suited to efficient local information processing than global information transmission (17). This is confirmed by some studies in sociology (104; 105; 33) as well as common experience: if A has two very close friends B and C, then B and C are also likely to know each other, and are perhaps also close. In a network description of a social system we would,

therefore, expect triangles to be abundant, and many of them should consist of strong weights. A network based on mobile phone calls might have a dual role: on the one hand it serves as a social network and, on the other, as a conduit for information transfer. In this case, inclusion of link weights can provide substantial new information about the system, most importantly about its function, and also places some constraints on the dynamical processes that may unfold on the network (17).

The following repeats some parts of the material in Publications (6; 7) in order to present it in a broader context and to augment some of the introduced ideas. We suggested the following definition for a motif in Publication (6): *a motif is a set of topologically equivalent subgraphs in a network.* Now consider a weighted (directed or undirected) network where the weight w_{ij} between v_i and v_j is assumed non-negative. We introduced in (6) the *intensity* i_g of subgraph g with vertices v_g and links ℓ_g as the *geometric mean* of its weights:

$$i_g = \left(\prod_{(ij) \in \ell_g} w_{ij} \right)^{1/|\ell_g|} , \tag{3.3}$$

where $|\ell_g|$ is the number of links in g. Note that the subgraph g can be, but does not have to be, a small fully connected graph. For example, one could take g as a fully connected graph of order 3, equivalent to a "closed triangle" (leftmost graph in Figure 3.4), or one could take g to be an "open triangle" (rightmost graph in Figure 3.4). With the above definition the units of intensity are the same as the units of the weights.

Due to the nature of the geometric mean, the subgraph intensity i_g may be low because one of the weights is very low, or it may result from all of the weights being low. In order to distinguish between these two extremes, we introduce subgraph *coherence* q_g as the ratio of the geometric to the arithmetic mean of the weights as

$$q_g = I\,|\ell_g| / \sum_{(ij) \in \ell_g} w_{ij} . \tag{3.4}$$

Here $q \in [0, 1]$ and it is close to unity only if the subgraph weights do not differ much, i.e., are internally coherent.

The *total intensity* I_M of a motif M in the network is the sum of its subgraph intensities $I_M = \sum_{g \in M} i_g$. For certain weighted directed motifs, the total intensities can be computed using matrix operations. Let the $N \times N$ weight matrix $\mathbf{W}$ describe the network weights. Analogously, let $\mathbf{A}$ represent the underlying $N \times N$ adjacency matrix such that $a_{ij} = 1$ if $w_{ij} > 0$, and $a_{ij} = 0$ if $w_{ij} = 0$. In an

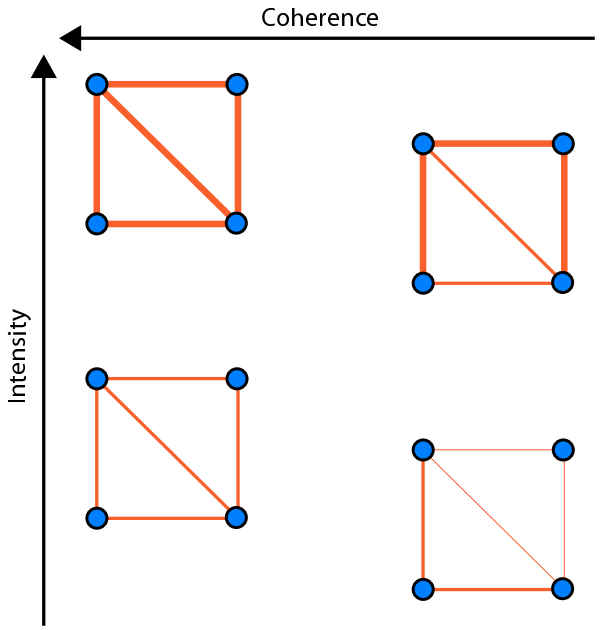

Figure 3.3: Schematic illustration of intensity and coherence for a subgraph with 4 nodes and 5 links, demonstrating why both intensity and coherence are needed for describing the distribution of weights within a subgraph. Here intensity increases to the top and coherence increases to the left. In the top row, the subgraph on the left has both high intensity and high coherence, whereas the one on the right has somewhat lower intensity and considerably lower coherence. The low coherence value suggests that the significantly weaker diagonal link may be due to noise. In the case of this high intensity–low coherence subgraph the issue of double counting (see text) is particularly relevant, because *within this subgraph* one should perhaps be looking for the subgraph with higher coherence. So instead of forcing a decision on, say, whether to consider this subgraph a loop or, alternatively, a loop with a diagonal link, one can do both, i.e., double count the nodes and links making up the loop. In practice, this means computing the intensity and coherence for both the loop and the loop with the diagonal link. If the subgraph is regarded a loop, its coherence will be considerably higher than if it is regarded a loop with a diagonal link. Although the subgraph in the bottom left corner has low intensity, its coherence is still high because all the links are consistently weak as quantified by the high coherence value.

unweighted network, the number of directed paths returning to the starting node after k steps can be written as

$$N(k) = \sum_{i_1,\ldots,i_k} \prod_{x=1}^{k} a_{i_x,i_{x+1}} = \sum_{i_1,\ldots,i_k} a_{i_1,i_2} a_{i_2,i_3} \cdots a_{i_k,i_1} = \mathrm{Tr}\{\mathbf{A}^k\}, \tag{3.5}$$

where the summation goes over all possible sites and $i_{k+1} = i_1$ (53). Let $\mathbf{W}^{(1/k)}$ represent a matrix obtained from $\mathbf{W} = [w_{ij}]$ by taking the k-th root of its individual elements such that $\mathbf{W}^{(1/k)} = [w_{ij}^{1/k}]$. The total intensity of motif M in the network is

$$\begin{aligned} I_M &= a_M \sum_{i_1,\ldots,i_k} \left(\prod_{x=1}^{k} w_{i_x,i_{x+1}} \right)^{1/k} \\ &= a_M \sum_{i_1,\ldots,i_k} w_{i_1,i_2}^{1/k} w_{i_2,i_3}^{1/k} \cdots w_{i_k,i_1}^{1/k} = a_M \mathrm{Tr}\{(\mathbf{W}^{(1/k)})^k\}, \end{aligned} \tag{3.6}$$

where a_M is a combinatorial factor ensuring that each subgraph is counted only once. A change in the direction of a link can be taken into account using the matrix transpose. In some cases additional block matrices are needed when the motifs are not cyclical (6). The benefit of the matrix formulation is that it enables very straight-forward computation of intensities for certain motifs. For example, for the non-frustrated triangle the total intensity becomes $I_\Delta = \frac{1}{3}\mathrm{Tr}\{(\mathbf{W}^{(1/3)})^3\}$, which can be computed in Matlab with `1/3*trace((W.^(1/3))^3)`.

In the case of unweighted motifs, the reference system is established by rewiring the network while conserving its degree distribution as explained above. The idea behind this *topological reference* is to remove local structural correlations present in the original network. In the framework of weighted motifs we have the additional issue of weight correlations, and the easiest way to remove them is simply to shuffle the weights. This *weight permuted reference* removes weight correlations while leaving the network topology unaltered. In principle, one can use either topological, weight permuted, or *mixed reference*, in which one both rewires the topology and shuffles the weights. However, since the motivation behind the weighted motif framework is to study the nature of the coupling between link weights and local network topology, it is meaningful to compare the empirical network to a weight permuted reference. Using this reference, any deviation in motif intensities between the empirical and reference system has a straightforward interpretation: the local organisation of weights in the empirical network is not random. In practice it may be difficult to establish a topological reference for a weighted network without distorting the *node strength* distribution, where node

strength s_i is defined as $s_i = \sum_{j,\, j \in \mathcal{N}(v_i)} w_{ij}$ (106; 4; 107).

Equation (3.2) can be generalised to weighted motifs. Replacing the number of subgraphs by their intensities gives the *motif intensity score* as

$$\tilde{z}_M = \frac{I_M^E - \langle I_M^R \rangle}{\sigma_M^R}, \tag{3.7}$$

where I_M^R is the total intensity of motif M in one realisation of the reference system and $\sigma_M^R = \sqrt{\langle (I_M^R)^2 \rangle - \langle I_M^R \rangle^2}$. It is clear that Eqs. (3.2) and (3.7) coincide for binary weights, implying that $\tilde{z} \to z$ in the limit, so that we obtain the unweighted z-score as a special case in this limit. Overall, the intensity-coherence framework suggests a shift in perspective from regarding subgraphs as discrete objects that either exist or not to a continuum of subgraph intensities, where zero or very low intensity values imply that the subgraph in question does not exist or exists at a practically insignificant intensity level. The motifs showing statistically significant deviation from some reference system can then be called high or low intensity (coherence) motifs.

A shortcoming of the z-score framework, whether unweighted or weighted, is that it it based on just one number from the empirical network (either n_M^E or I_M^E) and two numbers characterising the reference distribution ($\langle n_M^R \rangle$ or $\langle I_M^R \rangle$ and σ_M^R). Especially in the case of weighted motifs, a lot of information is lost by adopting this approach. A better way is to compare the intensity distribution $P^E(i_g)$ for subgraphs g in the empirical network to the intensity distribution $P^R(i_g)$ in the random ensemble. Now the problem becomes one of comparing two distributions with one another, for which several tools are available, such as the standard Kolmogorov-Smirnov test, for which the null hypothesis is that the two distributions, or samples, $P^E(i_g)$ and $P^R(i_g)$ are drawn from the same continuous distribution. A more information theoretic approach can be based on the Kullback-Leibler divergence, which measures the distance from a probability distribution P representing the empirical intensity distribution, the observations, to an arbitrary probability distribution Q, representing the reference distribution generated under the chosen null hypothesis model (108).

An important technical detail related to both the unweighted and weighted motif framework has to do with what is called *double counting* in Publication (6). For example, one could be interested in studying the motif statistics for "open triangles" (a node and its adjacent links) and "closed triangles" in the network, corresponding to subgraphs g_1 and g_2, respectively. In this case the open triangle g_1 is a subgraph of the closed triangle g_2. The double counting dilemma is now the following: should every instance of g_2 be also counted as an instance (or, actually, three instances) of g_1? Earlier studies have indeed counted an open triangle if and only if it is not a subgraph of a closed triangle. It is precisely for this reason, i.e., compatibility with earlier work, that we introduced the block matrix in Publica-

tion (6), which prevents us from double counting subgraphs. Applying the block matrix "blocks" or prevents us from counting an open triangle if the open triangle is part of a closed triangle.

However, double counting is not a true problem as long as the same method of counting is applied to both the empirical and random networks. If we allow for double counting in the above example of triangles, we end up counting three open triangles every time we count a closed triangle. This increases the number of open triangles, but does it exactly in the same way for the empirical and random networks. If we still wish to avoid double counting after we have obtained the results, we can always subtract the counts in the end. However, disallowing double counting does become a true problem as one needs to specify an arbitrary upper limit as to what is counted. For example, one could decide to count closed triangles but not open triangles, or equally well count only squares with a diagonal (equivalent to two triangles sharing an edge), but not closed triangles and so on. This is particularly problematic if we have measurement noise in the system. Then, in the worst case, we can have a fully connected graph where most of the edges have a very small weight ϵ corresponding to the noise. We could now end up with closed triangles where one or more edges have weight ϵ. If we disallow double counting and specify an arbitrary upper limit, we will only count the closed triangles that may have resulted purely from noise, but would not count the open triangles that could have considerably higher intensity. This argument shows that although double counting can be avoided, perhaps it should not be.

Triangles are among the most important nontrivial motifs and they play an important role as one of the basic quantities of network characterisation in defining the *clustering coefficient* C_i (Section 3.1) at node i as

$$C_i = \frac{2t_i}{k_i(k_i-1)}, \tag{3.8}$$

where t_i is the number of triangles attached to the node v_i and k_i is its degree (36). This quantity is normalised between 0 and 1, and it characterises the tendency of the nearest neighbours of node v_i to be interconnected. Replacing the number of triangles t_i in Eq. (3.8) with the sum of triangle intensities yields the *weighted clustering coefficient* as

$$\widetilde{C}_i = \frac{2}{k_i(k_i-1)} \sum_{j,k} (\tilde{w}_{i,j}\tilde{w}_{j,k}\tilde{w}_{k,i})^{1/3}, \tag{3.9}$$

where we use weights scaled by the largest weight in the network given by $\tilde{w}_{ij} = w_{ij}/\max_{ij}(w_{ij})$. This definition fulfils the requirement that $\widetilde{C}_i \to C_i$ as the weights become binary. This, and several other requirements that lead to the above formulation of weighted clustering coefficient, are elaborated in Publication (7). We can relate the unweighted and weighted clustering coefficients through the

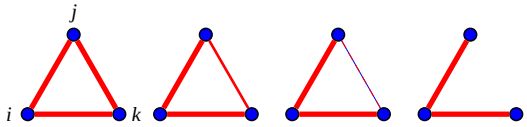

Figure 3.4: A schematic illustration of the weighted clustering coefficient $\widetilde{C}_i$ for vertex v_i. The weight w_{jk} is gradually decreased from left to right. The value of $\widetilde{C}_i$ decreases as $C_i \sim w_{jk}^{1/3}$, tending smoothly to zero in the limit.

average intensity of triangles at node i defined as $\bar{I}_i = \frac{1}{t_i}\sum_{g\in\mathcal{N}(v_i)} I(g)$, where $\mathcal{N}(v_i)$ denotes the neighbourhood of v_i, and this allows us to write the weighted clustering coefficient as

$$\widetilde{C}_i = \bar{I}_i C_i. \tag{3.10}$$

This equation gives a plausible interpretation of the weighted clustering coefficient: it is the unweighted (topological) clustering coefficient renormalised by the average intensity. Alternative formulations of weighted clustering coefficient are given in (107; 109).

3.4 Percolation on networks

Percolation theory addresses some of the same questions as random graph theory does, offering a different perspective on the same problem. Therefore, understanding some of its basic predictions related to networks is important for a more complete understanding of networks. The standard introduction to percolation theory is the book by Stauffer and Aharony (110), and some background in the theory of phase transition is also helpful for understanding it (111).

According to random graph theory, it is known that there exist critical, system size dependent, connectivity probabilities $p_c(N)$ marking the appearance of certain kinds of subgraphs (56). The probability that a graph with N nodes and connection probability $p(N)$ has property Q satisfies

$$\lim_{N\to\infty} P_{N,p}(Q) = \begin{cases} 0 & \text{if } \frac{p(N)}{p_c(N)} \to 0 \\ 1 & \text{if } \frac{p(N)}{p_c(N)} \to \infty \end{cases} \tag{3.11}$$

In other words, if the connectivity probability $p = p(N)$ grows faster than the critical probability $p_c(N)$, the considered subgraph will be present, otherwise it will not. The critical probability at which *almost every graph* contains a subgraph

with k nodes and l links is $p_c(N) \sim N^{-k/l}$, where the definition of almost every graph containing Q means that the probability of having property Q approaches 1 as $N \to \infty$. As an important special case, for cycles the number of nodes and links is the same, thus $k = l$, and the critical probability of having a cycle of any order is $p_c(N) = cN^{-1}$, where c is a constant.

Assume that the connectivity probability scales as $p(N) \sim N^z$, with the tunable parameter $z \in (-\infty, 0]$. When z passes through -1, the asymptotic probability of cycles of all orders jumps from 0 to 1. This is to say, for example, that in an ER network with N nodes and $n = N$, we have $p(N) = n/[N(N-1)/2] = 2/(N-1) \sim N^{-1}$ and, consequently, would expect to see cycles. In fact, for the case of $z = -1$ the topological properties of the graph depend on the value of the average degree $\langle k \rangle$. If $0 < \langle k \rangle < 1$, almost surely all clusters are either trees or clusters containing exactly one cycle. The largest cluster is a tree and its size is proportional to $\ln N$. As $\langle k \rangle = \langle k \rangle_c = 1$, the structure of the graph changes abruptly. Now the largest structure takes on a complicated structure, its size scales as $N^{2/3}$, and a finite fraction of nodes belong to it. Except for this giant cluster, all other clusters are small. As $\langle k \rangle$ increases further, the small clusters coalesce and join the giant cluster (53).

The formation of a giant cluster at a critical probability is very similar to a percolation transition. Consider a regular d-dimensional (triangular, honeycomb, hypercubic) lattice. In percolation the nodes of the lattice are usually called *sites* and the links of the lattice are called *bonds*. There are two different basic types of percolation. In *site percolation* all the bonds are present and each site is randomly occupied with probability p and empty with probability $1-p$. A cluster is defined as a group of occupied neighbouring sites, and the size of the cluster is simply the number of occupied sites belonging to the cluster. In *bond percolation* each site is considered occupied and a bond is open (link present) with probability p and closed (link not present) with probability $1-p$. A cluster is a group of sites connected by open bonds, and when referring to its size one has to define whether one counts the number of sites or the number of bonds. In short, percolation theory deals with the properties of these clusters (110).

The basic concepts of percolation are easiest to understand through site percolation. Here the probability of an arbitrary site being occupied, p, is sometimes called the *concentration*. At the percolation threshold p_c a percolating cluster connecting two opposing sides of the lattice emerges. The percolation transition gives the position of a phase transition at which the system changes its qualitative behaviour for one particular value of a continuously varying control parameter. In the case of percolation, the control parameter is the concentration p, and there is no percolating cluster for $p < p_c$ but at least one percolating cluster for $p > p_c$.

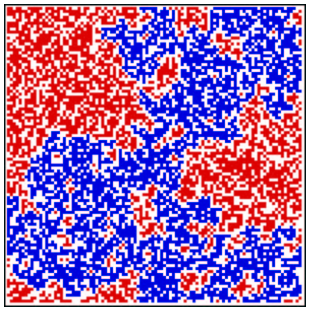

Figure 3.5: Illustration of a modified percolation problem. Here the blue cluster percolates through the system. Figure from `www.math.cornell.edu`.

To determine the phase the system is in, one uses an order parameter, which is a quantity that provides a signature of order in the system. It is zero in the unordered phase ($p < p_c$) and non-zero in the ordered ($p > p_c$) phase. Although in general the choice of the order parameter is not unique for a given physical system, one usually uses the *percolation probability* P defined as the probability of an arbitrary site to belong to the infinite network, the giant cluster, and P is also known as the *strength* of the infinite network. For percolation the order parameter goes to zero continuously as one approaches the critical point from above ($p > p_c$), qualifying it as a continuous or second-order phase transition. At precisely $p = p_c$ a path connects the top and bottom of the lattice. Although p_c is only defined in the infinite system size limit, one can use finite size scaling (FSS) to extrapolate a system of finite size N to the limit $N \to \infty$.

The percolation probability P mentioned above is one of the quantities of interest in percolation. Using a general equation valid for all site percolation problems, it can be related to the concentration p through

$$P + \sum_s n_s s = p, \tag{3.12}$$

where the sum runs over all finite clusters s (thus excluding the infinite cluster) and $n_s s$ is the probability that a site belongs to a cluster of size s, where n_s is known as the *normalised cluster number*. This equation simply states that all occupied sites (probability p) either belong to the infinite cluster with probability P or to one of the finite clusters with probability $\sum_s n_s s$. In addition to P, one can characterise a system using the *mean cluster size*

$$S = \sum \frac{n_s s^2}{\sum_s n_s s}, \tag{3.13}$$

which gives the average size of a randomly chosen cluster. Last, the system can be

quantified through the *cluster size distribution* and, according to percolation theory, near the percolation threshold p_c even the most general percolation problem in any dimension obeys a scaling relation of the form

$$n_s(p) = \begin{cases} s^{-\tau} f_-(|p - p_c|^{1/\sigma} s) & \text{if } p \leq p_c \\ s^{-\tau} f_+(|p - p_c|^{1/\sigma} s) & \text{if } p \geq p_c, \end{cases} \tag{3.14}$$

where the precise form of the smooth scaling functions $f_-(z)$ and $f_+(z)$ on $z \in [0, \infty)$ have to be determined, in most cases numerically, and σ and τ are known as the *critical exponents*. Thus, what different percolation problems have in common is that their cluster size distributions follow a power law. As an example, one can show that this expression contains the Cayley tree, a type of lattice, as a special case with $\tau = 5/2$, $\sigma = 1/2$, and $f_+(z) = f_-(z) = e^{-z}$.

Other quantities of interest also diverge near the percolation transition and this asymptotic divergence can be described by simple power-laws and their associated critical exponents can all be derived from the values of τ and σ. For example, the critical exponent of the percolation probability β is defined through the scaling relation $P(p) \sim (p - p_c)^\beta$ with $\beta = (\tau - 2)/\sigma$. Thus for $p = p_c$ the strength $P = 0$ and it increases for $p > p_c$ as β is a positive power. The critical exponent γ of the average cluster size is defined from $\langle s(p) \rangle^x \sim |p - p_c|^{-\gamma}$ with $\gamma = (3 - \tau)/\sigma$. Finally, the *correlation length* ξ, which can be shown to be proportional to a typical cluster diameter, is defined as $\xi(p) \sim |p - p_c|^{-\nu}$ with $\nu = (\tau - 1)/(d\sigma)$, where d is the dimensionality of the system d (110).

Traditionally percolation theory has dealt with regular d-dimensional lattices whereas random graph theory deals with irregular N-dimensional graphs. Since random graph theory investigates the $N \to \infty$ regime, in the limit the graphs are considered to be infinite dimensional as they are not embedded in a metric space. It turns out that percolation and random graph theory meet in the infinite-dimensional limit of percolation as $d \to \infty$. This also demonstrates that, conceptually, percolation theory is applicable in such systems.

The study in Publication (5) is really just bond percolation with the additional ingredient that, in the case of the empirical network, the links are inserted in a certain order (either ascending or descending) based on their weight w_{ij}. One can then compare different characteristics, such as cluster growth types, against some reference system as a function of the control parameter p, the fraction of included links, or the probability for an arbitrary link to be present (i.e., bond to be open). The idea is that the behaviour of these curves carries information about the structure of the network, and one can compare the curves obtained from different networks to learn about their structural similarity. Naturally, one could also determine the critical exponents for different networks, but this might be difficult in practice, especially for small networks.

3.5 Rate equation approach to network models

The historical developments of complex networks covered in Section 3.1 inspired and still inspire the development of parsimonious models able to reproduce and explain the properties exhibited by real complex networks. The ER-model is a static model, meaning that it is defined for a fixed system size, i.e., a fixed number of vertices. A change in the perspective of network modelling took place after the introduction of the Barabási-Albert (BA) model (60), which lead to the idea that complex networks are the result of growth processes, and that the placement of new links is not random but follows preferential attachment. The class of growing network models, to which the BA model also belongs, became important because of its explanatory power with respect to scale-free networks. In addition, networks based on human interactions, including transportation systems, electrical distribution systems, biological systems, and the Internet are also continuously growing (112; 113; 114; 115). Growing networks are also important because, in practice, many networks are generated using a growth algorithm, i.e, an algorithm that increases the size of the network on each time step. Thus, growing networks are an important class of networks, both from the theoretical and practical perspective. This leads to the need for developing methods for solving growing network models analytically with the aim of understanding their structure and time evolution. Standard probabilistic techniques (56) and the generating function formalism (116; 78) are often very difficult and limited in their applicability. Two other methods are the master equation approach, the basic idea of which is explained below, and the rate equation approach, which is covered in more detail and applied to the study of the social network model in Publication (9).

The master equation is one of the most important equations in statistical physics because of its wide applicability. It has been applied to problems in biology, population dynamics, laser physics, fluids, semiconductors, and many other systems. As a system of stochastic variables evolves in time, transitions occur between various realisations of the stochastic variables. In deriving the master equation one assumes that the probability of each transition depends only on the preceding step (117). The master equation is a differential equation describing the time-evolution of the probability of a system to occupy each one of a discrete set of states. If $P(n,t)$ denotes the probability density that the stochastic variable X has value n at time t, i.e., the probability density that the system is in state x, the master equation can be written as

$$\frac{\partial P(n,t)}{\partial t} = \sum_{m=1}^{M} [P(m,t)w_{m,n}(t) - P(n,t)w_{n,m}(t)], \tag{3.15}$$

where $w_{m,n}$ is the transition probability rate, i.e., the probability of a transition from state m to state n during the differential time interval. This equation gives

the rate of change of the probability $P(n,t)$ due to transitions into the state n from all other states (first term) and due to transitions out of state n into all other states (second term). Because of these transitions, the probability of finding the system in a given state changes until the system reaches a final steady state in which transitions cannot cause further changes in the probability distribution (117).

When applied to solving the degree distribution of a growing network model, one writes an equation for the changes in the density $p(k)$ of vertices of degree k as function of time and looks for stationary solutions, i.e., solutions for which the density does not change anymore, corresponding to an infinite network size. For example, for the BA-model the degree distribution is obtained as $P(k) = 2m(m+1)/[k(k+1)(k+2)]$ (118). Other examples of using the master equation approach in solving growing network models include finding the exact form of the stationary distribution of the number of incoming links of sites in the limit of long times for a generalised BA model (118), and the derivation of the exact solution of the time-dependent cluster size distributions for a preferential growth model (119). Although the master equation approach is conceptually very simple, the calculations can become very involved, in practice.

Rate equations are commonly used, for example, in chemical kinetics to link the rate of a reaction to each reactant. In the context of complex networks, the rate equation approach was fist introduced for deriving, in the continuous degree approximation, the degree distribution for the BA model yielding $P(k) \sim 2m^2k^{-3}$, together with the introduction of the mean-field theory for scale-free networks, a prerequisite for the rate equation approach (120; 60). The rate equation formalism was later applied to finding the clustering spectrum of growing network models with preferential attachment (121; 122). The rate equation approach has also been extended to a general model for the evolution of weighted networks that couples the topology and dynamical evolution of weights (123), and augmented by considering two point correlations (degree-degree correlations) for finding the average nearest neighbour degree (124). The basic formulation of the rate equation for finding the degree distribution $P(k)$ and the clustering spectrum $C(k)$ are given below.

Degree distribution $P(k)$. The details of the rate equation approach to solving the degree distribution $P(k)$ depend on the model, but the starting point is to identify the microscopic mechanisms responsible for increasing the degree of a chosen vertex v_i (Figure 3.6). The rate equation for the degree of vertex v_i describes how its degree changes on average during one time step of the network growth. This can be written as

$$\frac{\partial k_i}{\partial t} = R(k_i, \Pi), \tag{3.16}$$

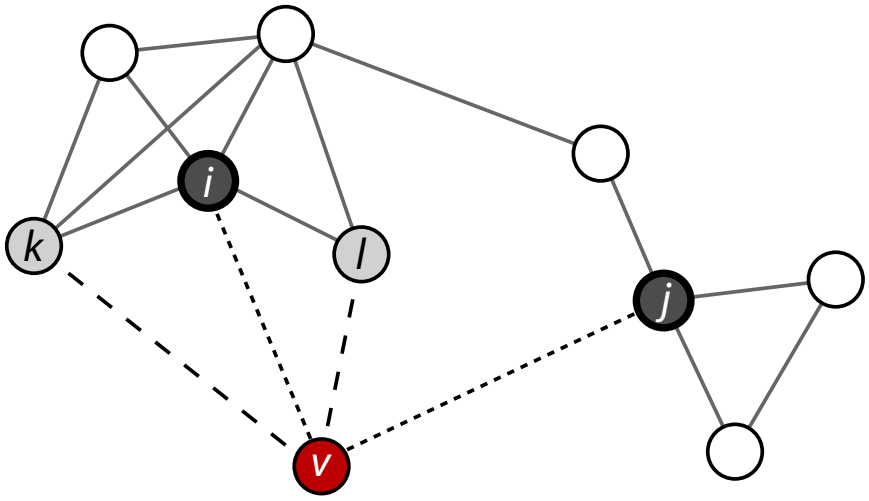

Figure 3.6: Illustration of the microscopic mechanisms in the model introduced in Publication (9). The new vertex v links to one or more randomly chosen initial contacts (here i and j) and possibly to some of their neighbours (here k and l).

where $R(k_i, \Pi)$ is the overall rate at which v_i gets new links (or loses them if rewiring and deletion of links are allowed). The rate $R(k_i, \Pi)$ can be written as a $R(k_i, \Pi) = \sum_l P_l \, \delta k_l$, where the sum is taken over the microscopic mechanisms l, P_l is the probability for mechanism l to take place, and δk_l is the average change in the value of k (mean field approximation) resulting from a realisation of mechanism l. Often it may be necessary to neglect structural correlations, such as the dependence of k_i on the degree of its (nearest) neighbours. Integrating the above rate equation yields an expression for $k_i(t)$, known as the *degree evolution equation*, which describes how the degree of a vertex grows as a function of time t. From $k_i(t)$ one can calculate the degree distribution $p(k)$ by forming the cumulative distribution $F(k)$ and differentiating this with respect to k. Since in the mean field approximation the degree $k_i(t)$ of a vertex v_i increases strictly monotonously from time t_i the vertex is initially added to the network, the fraction of vertices whose degree is less than $k_i(t)$ at time t is equivalent to the fraction of vertices that were introduced after time t_i. Since t is evenly distributed, this fraction is $(t - t_i)/t$. This leads to the cumulative distribution

$$F(k_i) = P(\tilde{k} \leq k_i) = P(\tilde{t} \geq t_i) = \frac{1}{t}(t - t_i). \tag{3.17}$$

Differentiating $F(k_i)$ with respect to k_i gives the probability density distribution $P(k)$ for the degree k.

Clustering spectrum $C(k)$. The rate equation approach to the clustering spectrum $C(k)$ also starts from identifying the microscopic mechanisms respons-

ible for increasing clustering at a vertex v_i. Since the clustering coefficient is a normalised number of triangles around a node, it is more natural to consider the total number of triangles around the node, which is equivalent to the number of links n_i amongst the nodes in the immediate neighbourhood of v_i. We can write the rate equation for number of triangles n_i around vertex v_i as

$$\frac{\partial n_i}{\partial t} = R(k_i, \Pi) \sum_{n \in N(v_i)} R(k_n, \Pi), \tag{3.18}$$

where $R(k_i, \Pi)$ is again the rate at which v_i gets new links. Again, we allow the rate $R(k_i, \Pi)$ to depend on both the degree k_i of the node in question and a set of parameters Π. Expressed in words, the above equation says that the rate of change of the average number of connected neighbours of vertex v_i is the product of two factors: the rate at which v_i gets new links and the rate at which the neighbourhood of v_i gets new links. In general, the complications in solving a rate equation like Eq. (3.18) arise from structural correlations, i.e., the correlations that are embedded between the degree of vertex v_i and the properties (such as degree) of the vertices in its neighbourhood. Depending on the model, it may or may not be possible to ignore some of these correlations and still get a satisfactory solution. Integrating Eq. (3.18) gives the time evolution equation $n_i(t)$ for the number of triangles around v_i. In order to express $n_i(t)$ as a function of vertex degree k as opposed to time t, we can use the degree evolution equation for $k_i(t)$, which relates degree k and time t, yielding $n_i(k)$. Since the clustering coefficient $c_i(k_i) = 2n_i(k_i)/k_i(k_i - 1)$, we obtain the clustering coefficient as a function of degree k, which is the desired clustering spectrum.

Chapter 4

Summary of results

The financial market was studied using the minimum spanning tree (MST) approach in Publications (1; 2; 4). The tree was constructed from correlations of logarithmic returns between stocks and it was referred to as the "asset tree". It was observed to have a hierarchical structure associated with it, such that different branches correspond to companies functioning in the same or closely related businesses. In general, this hierarchy matched well with third party business sector classification of stocks, and the observed deviations were to be expected and could be accounted for (4). Stocks were differentiated from one another also based on their vertex degree, the distribution of which was found to have a power-law tail of the form $P(k) \sim k^{-\gamma}$ with $\gamma = 2.1 \pm 0.1$. The outliers of this distribution are the hubs, or central nodes, of the tree and while the number of companies in Publications (1; 2; 4) is too small for making conclusions about scale invariance of the tree, the existence of hubs from which the hierarchical structure is built up supports this notion. Degree and strength-based criteria for identifying the central node yielded almost identical outcomes and, quite remarkably, it appeared that there is a dominant hub in the network that persists for most of the studied period of 20 years.

The topology of the tree evolves over time reflecting changes in the market. These changes in tree topology were characterised at the level of links using the single- and multi-step survival ratios. The former measure revealed that during a crash, identified as Black Monday, the topology undergoes a significant reconfiguration. The latter measure was used for defining tree half-life, a characteristic time for the network in which half of the links have become rewired. This is almost twice as long for the empirical tree than for the reference obtained by sampling empirical return distributions and disregarding the underlying correlation structure, demonstrating that there is considerable and fairly stable structure in the market correlations ($t^E_{1/2} \approx 0.37$ and $t^S_{1/2} \approx 0.21$ for empirical and reference tree, respectively, using $T = 4$ years and $\delta T = 1/12$ year (125)). The stability

of the central node enabled its use as a reference point for the occupation layer concept, measuring how spread out the tree is at a given time. The value of the mean occupation layer fell at the time of the market crisis, offering another perspective on the change in tree topology during the crash. The position of stocks within the tree was shown to have a connection to portfolio optimisation: the stocks included in the minimum risk Markowitz portfolio tend to be located on the outskirts of the tree.

The graph approach to the financial market was introduced in (3) and studied further, primarily in Publication (5), and the resulting network was referred to as the "asset graph". The graph was used as an example of a weighted complex network in Publications (6; 7). The approach was motivated by the desire to unravel the clustering and interconnected nature of the financial market not captured by the tree approach. It turns out that the trees and graphs share on average 25% of links, and it is the remaining links that cause them to exhibit qualitatively different behaviour. In general, the graph incorporates links corresponding to very high levels of correlation. The graph thus captures well the clusters with their dense internal connections in the market, indicating that some important correlations are missed by the tree. Since the strong links also tend to persist over time, the asset graph exhibits higher single-step survival ratios than the asset tree ($\bar{\sigma}_{\text{graph}} \approx 0.95$ and $\bar{\sigma}_{\text{tree}} \approx 0.80$ for $T = 4$ years and $\delta T = 1/12$ year). The graphs are also more robust in the long run, as witnessed by considerably longer half-lives ($t^E_{1/2} \approx 1.59$ and $t^S_{1/2} \approx 0.25$ for empirical and reference graph, respectively, using $T = 4$ years and $\delta T = 1/12$ year (125)). The functional form of the vertex degree distribution for the graph was left inconclusive.

We studied the formation of the asset graph as a function of the number of included links, sorted by their weight, by dividing the growth processes into four distinct classes, called growth types and, for reference, compared them to an Erdös-Rényi graph. We found that type IV growth, responsible for creating cycles, sets in much earlier for the asset graph, reflecting the strong interconnectedness of the market. Also, the number of clusters was found to be an order of magnitude lower in the asset graph than in the reference, implying that the financial market encompasses relatively few clusters. We also studied the average clustering coefficient as a function of the number of included links and found that the asset graph is essentially clustered after only 10% of the strongest links are included, in contrast to the 60% of links required for the corresponding ER graph. This means that, in the empirical network, the remaining 90% of links probably connect nodes that are not part of the same cluster of stocks. This lead us to conjecture that these 90% of links correspond to sporadic correlations and, as such, may not carry genuine information about the structure of the market.

The concepts of subgraph intensity and coherence were introduced in Publication (6) and reviewed in (7), in which the criteria leading to the chosen formula-

tion of the weighted clustering coefficient $\tilde{C}$ were also presented. To demonstrate the ability of $\tilde{C}$ to capture simultaneous changes in topology and link weights, we applied it to an undirected network known to undergo both types of changes, namely, the asset graph under a market crash. Monitoring the average clustering coefficient as a function of time, we showed that, in entering the market crash, the topological clustering coefficient (36) and an alternative weighted clustering coefficient (107) both increase 5%. In contrast, $\tilde{C}$ increased 39%, demonstrating that it captures also the non-topological aspects of the sudden network adaptation.

Intensity and coherence also allow a natural generalisation of the z-score to the motif intensity score $\tilde{z}$, suggesting a shift in perspective from subgraphs as binary objects, which either exist or not, to a continuum of subgraph intensities and coherences. We studied three simple motifs for the directed metabolic network of *Escherichia coli*: (i) path of order 2, (ii) non-frustrated triangle, and (iii) frustrated triangle. The motif intensity scores for the unweighted networks based on the subgraph counts are $z_i = -5.4$, $z_{ii} = 12.8$, and $z_{iii} = -0.5$, whereas for the weighted networks we have $\tilde{z}_i = 14.8$, $\tilde{z}_{ii} = 33.8$, $\tilde{z}_{iii} = 9.0$. These results showed that a move from unweighted to weighted motif characteristics can cause a change from low to high intensity (shift from under-representation to over-representation), the intensity may become amplified (increase the extent of over-representation), or it may increase from average to high intensity (shift from statistically insignificant to over-representation). Bearing in mind that $\tilde{z} \to z$ as the weights become binary, these results indicate that the inclusion of weights can considerably modify the motif statistics and, hence, the conclusions drawn from motif studies.

Intensity and coherence were applied to study "communities of stocks" in Publication (8), where stocks belonging to the same business sector according to Forbes were considered to form a fully connected community. Our approach indicates the extent to which business sector classifications are visible in market prices, enabling us to gauge the extent of group-specific behaviour. In most cases these stocks were tied together in the sense that intra-cluster interaction strengths were considerably stronger than those of the market on the whole. As expected, the cluster-specific behaviour was temporarily suppressed by the crash but resumed after the market recovered. Business sector clusters were also more coherent than the market with the exception of the Basic Materials cluster, which was accounted for by the diversity of its industry composition. These results suggest that tight clusters (high intensity and high coherence) behave as one entity, like a coarse-grained "super-stock" and, therefore, offer little hope hope for diversification. Instead, a small risk portfolio could be more successfully constructed by focusing on stocks in different clusters.

A growing network model that possesses the typical topological properties of real social networks was introduced in Publication (9). Three requirements were imposed on the model. First, the model should reproduce the salient fea-

tures of empirical social networks. Second, it should be possible to implement the model with a fast algorithm to enable studying very large social systems over large ensembles. Third, the model should be simple enough to allow analytical derivation of the fundamental characteristics. All of these requirements were met. The model incorporates a broad degree distribution, assortative mixing (positive degree-degree correlations), high clustering, short average path lengths (the small-world property) and, most importantly, communities whose structure resembles real social communities. The algorithm of the model is very efficient and simple, consisting of two processes: (i) attachment to random vertices, and (ii) attachment to their neighbourhood. Analytical expressions for the degree distribution and clustering spectrum were derived and compared with simulation. The observed minor discrepancies were shown to result from degree-degree correlations that were ignored in solving the model. We also showed the model to have prominent community structure by studying the scaling of the number of k-clique communities (44; 126) and by visualising small networks.

Chapter 5

Discussion

Financial practitioners have shown some interest in using asset trees as an aid in real life portfolio optimisation, but also in understanding trade dependencies between countries, as indicated by the contacts initiated with the author by a Swiss investment bank and the Bank of England. As for portfolio optimisation, it should be admitted that although the Markowitz portfolio optimisation scheme was considered extremely important at the time of its introduction, so much so that it earned the Nobel Prize in economics, the methods currently used in portfolio allocation are considerably more sophisticated. Some reasons for this are the abundance of financial data and computer power for generating more complex models, but another important factor is the conceptual advances made in portfolio optimisation, of which the Value at Risk (VaR) scheme (26), for example, focuses on extreme losses.

Although Markowitz is no longer a buzzword in portfolio optimisation, it remains the cornerstone of modern portfolio theory and offers a useful reference. As such, one could identify the portfolio stocks based on the Markowitz and some additional schemes and visually compare their location in the tree. The observation that the minimum risk Markowitz portfolio stocks are located on the outskirts of the tree can be used to gauge visually the riskiness of the stocks suggested by different schemes. In addition, portfolio optimisation is rarely based on technical analysis alone, but one usually has some subjective judgement to add, which could lead to a preference to exclude some stocks from the portfolio. If the need for this should arise, one could look as a first approach for suitable candidates in the vicinity of the stock to be replaced in the tree. Once some suitable candidates have been identified, one could approach the problem with more sophisticated techniques.

Visualising the dependencies using correlations between, say, $N = 1000$ companies would require knowledge of $N(N-1)/2 = 499500$ correlation coefficients. Even worse, the number of correlation coefficients scales as N^2, easily

resulting in millions of correlation coefficients for a large stock exchange. With the asset tree approach, the number of links scales as the number of stocks N, and the problem of visualisation is considerably more manageable. Another interesting idea that deserves mention is using the asset tree and asset graph to provide a market based classification of stocks. Although there are different third party classifications available, it might be more relevant from the perspective of portfolio optimisation to define the classes or clusters based on return correlations. Indeed, whenever one wishes to consider the variance of portfolio returns, the correlations enter the equation. These clusters can be considered coarse-grained "super-stocks" and characterised, for example, using the concepts of intensity and coherence. The line of thought followed in Publication (8) is similar to this one, except that the clusters were identified based on an outside classification of stocks. The gist of the present idea is that the asset graph could be used to identify these clusters, and then a similar analysis could be carried out for them as in Publication (8).

It is also notable that the asset graph approach has other potential applications outside the financial market. Perhaps the best current example consists of the correlation networks obtained from brain activity, in which the voxels in the brain correspond to stocks, and their electrical activity to stock returns (127). This is an encouraging example of how a completely different problem can be recast as a network problem, using the asset graph approach. The clusters in this problem represent brain areas that are activated at the same time. One might be interested in studying how intense and coherent these co-activations are, and one possibility for doing this, as suggested by the wording used, is to apply the intensity-coherence framework. This can be easily implemented if one has the network data available and exemplifies the transferability and general nature of the network measures developed in the publications of this thesis. The author is also aware of an ongoing systems biology study that uses the same framework for identifying the overall structure of a large set of clinical data.

Further, the author has become involved in a new multidisciplinary project that includes contributions from cognitive scientists and psychologists, with the aim of applying some ideas used in the study of asset graphs in a completely different context. Using laboratory experiments with human subjects, we aim to create a network consisting of words expressing various emotions. Although linguistic networks have been studied in the past, in an attempt to uncover the structural properties of the vocabularies of different languages, here the point is not to study the lexicon of emotions. Instead, by defining a distance d_{ij} between words i and j describing an emotional state, we aim to go beyond language and use the emerging clusters to define a sort of emotional landscape. It is not unrealistic to expect that the study might be able to identify subjects with the Asperger syndrome, a form of high-functioning autism characterised with atypical or poorly developed social skills and delayed emotional development. This approach is very similar to the

asset graph approach, and some pruning of links will most likely be necessary here too. Although the two problem settings are not identical; for example the emotion graphs are static, so that time average for asset graphs is replaced by ensemble average taken over test subjects, similar measures can still be used in their characterisation.

The social network model introduced in Publication (9) should be seen as a starting point of the some of the research work envisioned in Section 1.2. The greatest strength of the model is probably its ability to produce communities that resemble real-life communities (17), whose details can be fine-tuned by varying the parameters of the model. Since there is heterogeneity in social relations, in the near future the model should be updated to incorporate interaction strengths. This is likely to have important consequences for dynamic processes unfolding on the network, as many processes on social networks depend crucially on interaction strengths. For example, it is reasonable to assume that one's opinions are more strongly affected by those of strongly connected individuals than those of weakly connected ones. Similarly, voting behaviour and information transmission, in general, depend on the closeness of social ties and, therefore, a study using realistic social network structure coupled with realistic interaction strengths would be likely to yield a more accurate picture of social system dynamics (95).

Another suggestion for future work is the incorporation of "uncertainty" in opinion formation models. Current spin-type models have two or more possible states or opinions, but there is no information on how strongly they are held. Yet, common experience indicates that the opinion concerning, say, which item to pick from the lunch menu is more prone to persuasion ("The chicken was great last week!") than the decision on which presidential candidate to vote for. However, augmenting the standard opinion formation paradigm is meaningless unless we first have a functioning model for both the topology and the interaction strengths; there is no point trying to run before you can walk. Fortunately, such a model can now be relatively easily constructed combining the topological model of Publication (9) with lessons learnt on the coupling of topology and interaction strengths in empirical social networks (17). Although speculation is premature, the outcome of this type of study will, in the least, be interesting, but it also has potential for unravelling the dynamics of different social phenomena at a level of realism and scale not possible in the past. The lessons learnt from this endeavour are not limited to understanding human societies, but may be applicable to other systems as well, thereby contributing to the complex systems paradigm itself.

References

[1] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, *Eur. Phys. J. B* **30**, 285 (2002).

[2] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, *Physica A* **324**, 247 (2003).

[3] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, A. Kanto, *Physica Scripta* **T106**, 48 (2003).

[4] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, A. Kanto, *Phys. Rev. E* **68**, 056110 (2003).

[5] J.-P. Onnela, K. Kaski, J. Kertész, *Eur. Phys. J. B* **38**, 353 (2004).

[6] J.-P. Onnela, J. Saramäki, J. Kertész, K. Kaski, *Phys. Rev. E* **71**, 065103 (2005).

[7] J. Saramäki, J.-P. Onnela, J. Kertész, K. Kaski, *Science of Complex Networks, AIP Conference Proceedings*, J. F. F. Mendes, *et al.*, eds. (2005), p. 108.

[8] J.-P. Onnela, J. Saramäki, K. Kaski, J. Kertész, *Practical Fruits of Econophysics, Nikkei Econophysics III Proceedings*, H. Takayasu, ed. (Springer, Tokyo, 2006), p. 302.

[9] R. Toivonen, J.-P. Onnela, J. Saramäki, J. Hyvönen, K. Kaski, *Physica A, in press* (2006).

[10] Y. Bar-Yam, *Dynamics of Complex Systems (Studies in Nonlinearity)* (Perseus Books Group, 1997).

[11] L. A. N. Amaral, J. M. Ottino, *Eur. Phys. J. B* **38**, 147 (2004).

[12] R. Gallagher, T. Appenzeller, *Science* **284**, 79 (1999).

[13] K. Kaski, *et al.*, Center of excellence application for 2006-1022, *Tech. rep.*, Laboratory of Computational Engineering, Helsinki University of Technology (2005).

[14] J. M. Ottino, *Nature* **427**, 399 (2004).

[15] *San Jose Mercury News* (January 23, 2000).

[16] P. W. Anderson, *The Economy As an Evolving Complex System (Santa Fe Institute Studies in the Sciences of Complexity Proceedings)* (Addison Wesley Publishing Company, 1988).

[17] J.-P. Onnela, *et al.* (2006). Work in progress.

[18] J. M. Ottino, *AIChE Journal* **49**, 292 (2003).

[19] G. A. Cowan, D. Pines, D. Meltzer, *Complexity: Metaphors, Models, and Reality* (Perseus Books Group, 1999).

[20] N. Goldenfeld, L. P. Kadanoff, *Science* **284**, 87 (1999).

[21] W. B. Arthur, S. N. Durlauf, D. A. Lane, *The Economy As an Evolving Complex System II: Proceedings (Santa Fe Institute Studies in the Sciences of Complexity Lecture Notes)* (Perseus Books Group, 1997).

[22] L. E. Blume, S. N. Durlauf, *The Economy As an Evolving Complex System III: Current Perspectives and Future Directions (Santa Fe Institute Studies on the Sciences of Complexity)* (Oxford University Press, 2005).

[23] R. N. Mantegna, H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, 1999).

[24] R. A. Brealey, S. Myers, *Principles of Corporate Finance* (McGraw, 2000).

[25] H. M. Markowitz, *Journal of Finance* **7**, 77 (1952).

[26] J.-P. Bouchaud, M. Potters, *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management* (Cambridge University Press, 2003).

[27] T. C. Schelling, *Micromotives and Macrobehavior* (W. W. Norton, 2006).

[28] H. C. White, S. A. Boorman, R. R. Breiger, *American Journal of Sociology* **81**, 730 (1976).

[29] M. Granovetter, *Decision Making: Alternatives to Rational Choice Models* (SAGE Publications, 1992), chap. Economic Action and Social Structure: The Problem of Embeddedness, pp. 304–333.

[30] P. J. Carrington, J. Scott, S. Wasserman, *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)* (Cambridge University Press, 2005).

[31] J. Barnes, *Human Relations* **7**, 39 (1954).

[32] S. Milgram, *Psychology Today* **2**, 60 (1967).

[33] M. Granovetter, *The American Journal of Sociology* **78**, 1360 (1973).

[34] S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)* (Cambridge University Press, 1994).

[35] J. P. Scott, *Social Network Analysis: A Handbook* (Sage Publications Ltd, 2000).

[36] D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).

[37] A. F. J. Vanraan, *Nature* **347**, 626 (1990).

[38] M. E. J. Newman, *Proc. Nat. Acad. Sci. USA* **98**, 404 (2001).

[39] M. E. J. Newman, *Phys. Rev. E* **64**, 016131(R) (2001).

[40] M. E. J. Newman, *Phys. Rev. E* **64**, 016132(R) (2001).

[41] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, *Phys. Rev. E* **68**, 065103(R) (2003).

[42] H. Ebel, L.-I. Mielsch, S. Bornholdt, *Phys. Rev. E* **66**, 35103(R) (2002).

[43] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Åberg, *Nature* **411**, 907 (2001).

[44] G. Palla, I. Derényi, I. Farkas, T. Vicsek, *Nature* **435**, 814 (2005).

[45] E. M. Jin, M. Girvan, M. E. J. Newman, *Phys. Rev. E* **64**, 046132(R) (2001).

[46] J. Davidsen, H. Ebel, S. Bornholdt, *Phys. Rev. Lett.* **88**, 128701(R) (2002).

[47] M. Marsili, F. Vega-Redondo, F. Slanina, *Proc. Nat. Acad. Sci. USA* **101**, 1439 (2004).

[48] M. C. González, P. G. Lind, H. J. Herrmann, *Phys. Rev. Lett.* **96**, 088702 (2006).

[49] P. S. Dodds, R. Muhamad, D. J. Watts, *Science* **301**, 827 (2003).

[50] R. Guimerà, B. Uzzi, J. Spiro, L. A. N. Amaral, *Science* **308**, 697 (2005).

[51] R. Axelrod, *Journal of Conflict Resolution* **41**, 203 (1997).

[52] P. Ball, *Critical Mass: How One Thing Leads to Another* (Farrar, Straus and Giroux, 2004).

[53] R. Albert, A.-L. Barabási, *Reviews of Modern Physics* **74**, 47 (2002).

[54] P. Erdös, A. Rényi, *Publ. Math. (Debrecen)* **6**, 290 (1959).

[55] P. Erdös, A. Rényi, *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17 (1960).

[56] B. Bollobás, *Random Graphs* (Cambridge University Press, 2001), second edn.

[57] B. Bollobás, *Modern Graph Theory, Graduate Texts in Mathematics* (Springer, 1998).

[58] D. West, *Introduction to Graph Theory* (Prentice-Hall, 1995).

[59] F. Harary, *Graph Theory* (Perseus, 1995).

[60] A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).

[61] S. N. Dorogovtsev, J. F. F. Mendes, *Advances in Physics* **51**, 1079 (2002).

[62] M. E. J. Newman, *SIAM Review* **45**, 167 (2003).

[63] T. S. Evans (2004). `cond-mat/0405123`.

[64] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, *Physics Reports* **424**, 175 (2006).

[65] H. S. E. S. Bornholdt, *Handbook of Graphs and Networks: From the Genome to the Internet* (Wiley-VCH, 2003).

[66] S. Dorogovtesev, J. Mendes, *Evolution of Networks* (Oxford University Press, 2003).

[67] R. Pastor-Satorras, A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, 2004).

[68] A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means* (Plume, 2003).

[69] D. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University Press, 1999).

[70] D. J. Watts, *Six Degrees: The Science of a Connected Age* (W. W. Norton & Company, 2004).

[71] M. Buchanan, *Nexus: Small Worlds and the Groundbreaking Theory of Networks* (W. W. Norton & Company, 2003).

[72] R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130 (1999).

[73] R. Pastor-Satorras, A. Vázquez, A. Vespignani, *Phys. Rev. Lett.* **87**, 258701 (2001).

[74] M. E. J. Newman, *Phys. Rev. Lett.* **89**, 208701 (2002).

[75] R. Milo, *et al.*, *Science* **298**, 824 (2002).

[76] S. Wuchty, Z. N. Oltvai, A.-L. Barabási, *Nature Genetics* **35**, 176 (2003).

[77] R. Milo, *et al.*, *Science* **303**, 1538 (2004).

[78] M. E. J. Newman, S. H. Strogatz, D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).

[79] A. Vázquez, *Phys. Rev. E* **67**, 056104 (2003).

[80] J. Saramäki, K. Kaski, *Physica A* **341**, 80 (2004).

[81] A. Broder, *et al.*, *Computer Networks* **33**, 309 (2000).

[82] R. Albert, H. Jeong, A.-L. Barabási, *Nature* **406**, 378 (2000).

[83] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, D. J. Watts, *Phys. Rev. Lett.* **85**, 5468 (2000).

[84] R. Cohen, K. Erez, D. ben Avraham, S. Havlin, *Phys. Rev. Lett.* **85**, 4626 (2000).

[85] R. Cohen, K. Erez, D. ben Avraham, S. Havlin, *Phys. Rev. Lett.* **86**, 3682 (2001).

[86] P. Holme, B. J. Kim, *Phys. Rev. E* **65**, 066109(R) (2002).

[87] P. Holme, *Phys. Rev. E* **66**, 036119(R) (2002).

[88] R. Pastor-Satorras, A. Vespignani, *Phys. Rev. Lett.* **86**, 3200 (2001).

[89] R. Pastor-Satorras, A. Vespignani, *Phys. Rev. E* **63**, 066117 (2001).

[90] D. J. Daley, J. Gani, C. Cannings, *Epidemic Modelling: An Introduction* (Cambridge University Press, 2001).

[91] D. H. Zanette, *Phys. Rev. E* **64**, 050901 (2001).

[92] Z. Liu, Y.-C. Lai, N. Ye, *Phys. Rev. E* **67**, 031911 (2003).

[93] Y. Moreno, M. Nekovee, A. Vespignani, *Phys. Rev. E* **69**, 055101 (2004).

[94] Y. Moreno, M. Nekovee, A. F. Pacheco, *Phys. Rev. E* **69**, 066130 (2004).

[95] J.-P. Onnela, *et al.* (2006). Work in progress.

[96] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, *Phys. Rev. Lett.* **83**, 1467 (1999).

[97] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, H. E. Stanley, *Phys. Rev. Lett.* **83**, 1471 (1999).

[98] R. Mantegna, *Eur. Phys. J. B* **11**, 193 (1999).

[99] M. Tumminello, T. Aste, T. D. Matteo, R. N. Mantegna, *Proc. Nat. Acad. Sci. USA* **102**, 10421 (2005).

[100] I. T. Jollife, *Principal Component Analysis* (Springer-Verlag, 1986).

[101] T. Kohonen, *Self-Organizing Maps* (Springer, 2000).

[102] K. Huang, *Statistical Mechanics* (Wiley, 1987).

[103] L. C. Freeman, *Sociometry* **40**, 35 (1977).

[104] F. Heider, *The Psychology of Interpersonal Relations* (Wiley, New York, 1958).

[105] T. M. Newcomb, *The Acquaintance Process* (Holt, Rinehart & Winston, New York, 1961).

[106] S. H. Yook, H. Jeong, A.-L. Barabási, Y. Tu, *Phys. Rev. Lett.* **86**, 5835 (2001).

[107] A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani, *Proc. Nat. Acad. Sci. USA* **101**, 3747 (2004).

[108] S. Kullback, R. A. Leibler, *Ann. Math. Stat.* **22**, 79 (1951).

[109] P. Holme, S. M. Park, B. J. Kim, C. R. Edling, *Physica A (in press)* (2006).

[110] D. Stauffer, A. Aharony, *Introduction to Percolation Theory* (CRC Press, 1994), second edn.

[111] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Westview Press, 1992).

[112] R. V. Solé, S. C. Manrubia, *Phys. Rev. E* **54**, R42 (1996).

[113] S. Jain, S. Krishna, *Phys. Rev. Lett.* **81**, 5684 (1998).

[114] J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, *Proceedings of the International Conference on Combinatorics and Computing, Lecture Notes in Computer Science* (Springer-Verlag, 1999).

[115] A. Broder, *et al.*, *Computer Networks* **33**, 309 (2000).

[116] H. S. Wilf, *Generatingfunctionology* (Academic Press, 1990).

[117] L. E. Reichl, *A Modern Course in Statistical Physics* (Wiley-Interscience, 1998).

[118] S. N. Dorogovtsev, J. F. F. Mendes, A. N. Samukhin, *Phys. Rev. Lett.* **85**, 4633 (2000).

[119] L. Kullmann, J. Kertész, *Phys. Rev. E* **63**, 051112 (2001).

[120] A.-L. Barabási, R. Albert, H. Jeong, *Physica A* **272**, 173 (1999).

[121] G. Szabó, M. Alava, J. Kertész, *Phys. Rev. E* **67**, 056102 (2003).

[122] P. L. Krapivsky, S. Redner, F. Leyvraz, *Phys. Rev. Lett.* **85**, 4629 (2000).

[123] A. Barrat, M. Barthélemy, A. Vespignani, *Phys. Rev. E* **70**, 066149 (2004).

[124] A. Barrat, R. Pastor-Satorras, *Phys. Rev. E* **71**, 036127 (2005).

[125] J.-P. Onnela, K. Kaski, J. Kertész, *submitted to InterJournal* (2004).

[126] I. Derényi, G. Palla, T. Vicsek, *Phys. Rev. Lett.* **94**, 160202 (2005).

[127] V. M. Eguíluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, A. V. Apkarian, *Phys. Rev. Lett.* **94**, 018102 (2005).

ISBN 951-22-8269-0 (printed)
ISBN 951-22-8270-4 (PDF)
ISSN 1455-0474