Department of Biomedical Engineering and Computational Science Publications
Teknillisen korkeakoulun Lääketieteellisen tekniikan ja laskennallisen tieteen laitoksen julkaisuja

November, 2008 REPORT A06

COMMUNITY STRUCTURES IN COMPLEX NETWORKS: DETECTION AND MODELING

Jussi Kumpula

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

Helsinki University of Technology
Department of Biomedical Engineering and Computational Science Publications
Teknillisen korkeakoulun Lääketieteellisen tekniikan ja laskennallisen tieteen laitoksen julkaisuja

November, 2008 REPORT A06

COMMUNITY STRUCTURES IN COMPLEX NETWORKS: DETECTION AND MODELING

Jussi Kumpula

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences, Helsinki University of Technology, for public examination and debate in Auditorium E at Helsinki University of Technology (Espoo, Finland) on the 29^{th} of November, 2008, at 12 noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Biomedical Engineering and Computational Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Distribution:
Helsinki University of Technology
Department of Biomedical Engineering and Computational Science
http://www.lce.hut.fi
P.O. Box 9203
FI-02015 TKK
FINLAND

Tel. +358-9-451 5726
Fax. +358-9-451 4833
http://www.becs.tkk.fi

Online in PDF format: http://lib.hut.fi/Diss/2008/isbn9789512296569/

E-mail: jkumpula@lce.hut.fi

ISBN 978-951-22-9655-2 (printed)
ISBN 978-951-22-9656-9 (PDF)
ISSN 1797-3996
PicaSet Oy
Helsinki 2008

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

Author Jussi M. Kumpula

Name of the dissertation

Community structures in complex networks: detection and modeling

Manuscript submitted 16.9.2008	Manuscript revised 11.11.2008
Date of the defence 29.11.2008	
☐ Monograph	☒ Article dissertation (summary + original articles)

Faculty	Faculty of Information and Natural Sciences
Department	Department of Biomedical Engineering and Computational Science
Field of research	Complex networks
Opponent(s)	Santo Fortunato, Research Scientist, Ph.D.
Supervisor	Prof. Kimmo Kaski
Instructor	Jari Saramäki, Senior Researcher, Dr.Tech.

Abstract

Complex systems are composed of a large number of interacting elements such that the system as a whole exhibits emergent properties not obvious from the properties of its individual parts. In the network approach, complex systems are represented as networks whose vertices and edges correspond to the elements and their interactions, respectively. Many networks, such as networks of protein interactions or social relationships, contain sets of densely interconnected nodes, communities, which play a concrete functional role in the original system, such as the group of proteins related to cancer metastasis. Detecting such communities in large networks has rapidly become one of the focal topics in the science of complex networks. The challenge in community detection is to define what constitutes a community in such a way that this definition not only yields meaningful communities but also allows for sufficiently fast algorithmic implementation to find them.

This thesis contributes to our understanding of community detection in complex networks in three ways. 1) The limitations of global optimization based community detection methods are analyzed. Here, the focus is on the dependence of the lower size limit of detectable communities on the tuning parameters of the methods. 2) This thesis significantly improves two community detection methods by extending their applicability domain: the Potts method is extended such that it can be applied to dense weighted networks, and a new algorithmic implementation for the k-clique percolation method is presented. The main advantage of the first method is that it allows analysis of dense weighted networks without discarding any of the link weights, whereas the advantage of the second method is its speed especially in the community analysis of weighted networks. 3) This thesis attempts to shed light on the formation of communities in networks. This is done by introducing a weighted model for social networks, whose mechanisms are based on empirical observations of social tie formation as well as observations on the topological role of tie strenghts. In this model, communities emerge only if nodes sufficiently favor their strong connections in the process of establishing new ones. The model is also utilized in studies of the effects of correlations of link weights and community structure on dynamics taking place on networks. Simulations of an opinion formation model show that the dynamics is significantly slowed down due to trapping of opinions in homogenized regions corresponding to communities.

Keywords	Complex networks, community detection, network dynamics, social networks
ISBN (printed) 978-951-22-9655-2	ISSN (printed) 1797-3996
ISBN (pdf) 978-951-22-9656-9	ISSN (pdf)
Language English	Number of pages 54p + app. 56p
Publisher	Department of Biomedical Engineering and Computational Science, TKK
Print distribution	Helsinki Univ. of Technology, Depart. of Biomedical Engineering and Computational Science

☒ The dissertation can be read at http://lib.tkk.fi/Diss/2008/isbn9789512296569/

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

Tekijä Jussi M. Kumpula

Väitöskirjan nimi

Yhteisörakenteet kompleksisissa verkoissa: tunnistaminen ja mallinnus

Käsikirjoituksen päivämäärä 16.9.2008	Korjatun käsikirjoituksen päivämäärä 11.11.2008
Väitöstilaisuuden ajankohta 29.11.2008	
☐ Monografia	☒ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)

Tiedekunta	Informaatio- ja luonnontieteiden tiedekunta
Laitos	Lääketieteellisen tekniikan ja laskennallisen tieteen laitos
Tutkimusala	Kompleksiset verkot
Vastaväittäjä(t)	Santo Fortunato, Research Scientist, Ph.D.
Työn valvoja	Prof. Kimmo Kaski
Työn ohjaaja	Jari Saramäki, Senior Researcher, Dr.Tech.

Tiivistelmä

Kompleksisessa systeemissä on suuri määrä keskenään vuorovaikuttavia elementtejä, jotka muodostavat organisoituja rakenteita ilman ulkopuolista ohjausta. Kompleksisten verkkojen teoriassa tällaiset systeemit kuvataan verkostoiksi, joiden solmut vastaavat elementtejä ja linkit näiden välisiä vuorovaikutuksia. Monet verkot sisältävät ns. yhteisörakenteita, jotka ovat tiheästi kytkettyjä solmujoukkoja. Usein yhteisörakenteet vastaavat systeemille oleellisia toiminnallisia rakenteita, esimerkiksi syövän leviämiseen liittyvät proteiinit voivat erottua omana yhteisönään. Tällaisten yhteisöjen löytäminen suurista verkostoista on viime vuosina muodostunut yhdeksi verkostotieteen avainongelmista. Ongelmana on löytää yhteisöille määritelmä, joka löytää sovelluskohteen kannalta mielekkäitä rakenteita ja lisäksi mahdollistaa tehokkaan algoritmisen toteutuksen.

Tämä väitöskirja edistää ymmärtämystä yhteisörakenteiden etsimisestä kompleksisista verkoista kolmella tavalla: 1) Väitöskirjassa tutkitaan globaaliin optimointiin perustuvien yhteisöjenhakumenetelmien rajoituksia keskittyen erityisesti havaittavien yhteisöjen minimikoon riippuvuuteen mallin parametreistä. 2) Väitöskirjassa esitetään kaksi merkittävää laajennusta yhteisönhakumenetelmiin: ensimmäinen laajentaa Potts-menetelmää soveltuvaksi tiheisiin painotettuihin verkkoihin, ja jälkimmäinen on uusi algoritminen toteutus k-klikkiperkolaatioon perustuvalle yhteisöhaulle. Ensimmäisen menetelmän suurin etu on, että se mahdollistaa verkon kaikkien linkkien huomioimisen yhteisöanalyysissä myös tiheissä painotetuissa verkoissa. Jälkimmäisen menetelmän vahvuus on sen nopeus erityisesti painotettujen verkkojen analysoinnissa. 3) Yhteisöjen tunnistamisen lisäksi väitöskirjassa tutkitaan yhteisöjä tuottavia prosesseja. Tähän käytetään väitöskirjassa esiteltävää painotettua sosiaalisten verkkojen mallia, jonka säännöt perustuvat empiirisiin havaintoihin sosiaalisten siteiden muodostumisprosesseista sekä linkkien vahvuuden topologisesta merkityksestä. Malliverkoissa muodostuu yhteisöjä vain mikäli solmut suosivat vahvoja linkkejään riittävän voimakkaasti uusia kontakteja etsittäessä. Mallia sovelletaan analyysissä, jossa tutkitaan linkkien painojen ja yhteisörakenteiden välisten riippuvuuksien vaikutusta verkoissa tapahtuviin dynaamisiin prosesseihin. Tähän käytetään mielipiteenmuodostusmallia, jonka dynamiikassa havaitaan merkittävää hidastumista johtuen homogeenisten alueiden muodostumisesta yhteisöjen sisälle.

Asiasanat	Kompleksiset verkot, yhteisöjen tunnistaminen, sosiaaliset verkot		
ISBN (painettu)	978-951-22-9655-2	ISSN (painettu)	1797-3996
ISBN (pdf)	978-951-22-9656-9	ISSN (pdf)	
Kieli	Englanti	Sivumäärä	54p + app. 56p
Julkaisija	Lääketieteellisen tekniikan ja laskennallisen tieteen laitos, TKK		
Painetun väitöskirjan jakelu	Teknillinen korkeakoulu, Lääketieteellisen tekniikan ja laskennallisen tieteen laitos		

☒ Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/2008/isbn9789512296569/

Preface

This thesis for the degree of Doctor of Technology has been prepared during the years 2005–2008 in the Department of Biomedical Engineering and Computational Science (former Laboratory of Computational Science) at the Helsinki University of Technology. My research has been funded by the GETA graduate school and by the Academy of Finland Centre of Excellence programs.

I would like to express my gratitude to several people who have helped me to complete this thesis and whose support has been invaluable during the years. First, I am grateful to my instructor Dr. Jari Saramäki for his great insight regarding fruitful lines of research and for his patient guidance over the years. I also wish to thank my supervisor Prof. Kimmo Kaski for supporting and encouraging me in all stages of my studies. Kimmo's faith in my skills has several times encouraged me to continue through difficult moments in research. I am thankful to Dr. Jukka-Pekka Onnela and Prof. Janos Kertész for their insightful ideas and for helping me in numerous ways during the years. I have been fortunate to have fun and exceptionally skilled co-workers, and a special thanks goes to the (not yet mentioned) members of our research group Tapio Heimo, Mikko Kivelä, Jörkki Hyvönen, Riitta Toivonen, and Dr. Riku Linna, as well as to several other people in our laboratory, namely, Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Mikko Viinikainen, and Jari Kätsyri who are excellent workmates and who (almost) never refused to discuss about irrelevant matters.

Beyond work, I am grateful to my parents Helena and Kari for providing me safe, inspirational, and encouraging environment to grow up. Helena brought me tons of books – fiction and popular science – when I was young and I'm sure this had a great effect on arousing my interest in science. My father is by far the best solver of mechanical puzzles I have ever seen, and it appears that I have inherited his interest and persistence in solving complex problems. Pirjo and Eero are the best parents-in-law one could hope for and I appreciate their continuous help in our everyday

life. Finally, my deepest gratitude belongs to my wife Linda and daughter Oona who have brought so much joy in my life. Linda's whole-hearted support has been of great value during all of my studies at TKK and Oona continuously reminds me that there are also other important things in life than work.

Eurajoki, November 2008

Jussi Kumpula

"However, the sum of difficulties is always a constant."
Prof. Juhani von Boehm on a lecture of theoretical mechanics.

Contents

List of publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I J. M. Kumpula, J. Saramäki, K. Kaski, J. Kertész: *Limited resolution in complex network community detection with Potts model approach*, Eur. Phys. J. B 56, 41-45 (2007).

II J. M. Kumpula, J. Saramäki, K. Kaski, J. Kertész: *Limited resolution and multiresolution methods in complex network community detection*, Fluctuations and Noise Lett. 7, 209-214 (2007).

III R. Toivonen, J. M. Kumpula, J.-P. Onnela, J. Saramäki, J. Kertész, K. Kaski: *The role of edge weights in social networks: modelling social networks and social dynamics*, Proc. SPIE 6601, 66010B, (2007).

IV J. M. Kumpula, J.-P. Onnela, J. Saramäki, K. Kaski, J. Kertész: *Emergence of Communities in Weighted Networks*, Phys. Rev. Lett. 99, 228701 (2007).

V J. M. Kumpula, M. Kivelä, K. Kaski, J. Saramäki: *A sequential algorithm for fast clique percolation*, Phys. Rev. E 78, 026109 (2008).

VI T. Heimo, J. M. Kumpula, K. Kaski, J. Saramäki: *Detecting modules in dense weighted networks with the Potts method*, J. Stat. Mech. 2008, P08007 (2008).

Kumpula has also been involved in the following publications, which are beyond the scope of this thesis and not included in it.

VII J. M. Kumpula, A. P. Sutton, K. Kaski: *Using the quantized mean-field method to model Joule heating in a short nano-conductor*, Chem. Phys. 324, 659-666 (2006)

VIII L. S. Kumpula, J. M. Kumpula, M.-R. Taskinen, M. Jauhiainen, K. Kaski, M. Ala-Korpela, *Reconsideration of hydrophobic lipid distributions in lipoprotein particles*, Chemistry and Physics of Lipids 155, 57-62 (2008)

Author's contribution

The research reported in this thesis is a result of collaboration between Jussi Kumpula, the author of this thesis, and the other authors of the included publications. Kumpula was the principal author in Publications I, II, IV, and V, and he performed the theoretical analysis, implemented the computer programs, had the main responsibility of the analysis of the numerical results, and prepared the manuscripts. The other authors in these publications contributed in developing the ideas, in the analysis of the results, and in writing and revising the manuscripts. In Publication III he was responsible for generating the model networks and in Publication VI he implemented the computer programs, performed numerical simulations, and derived the theoretical results. As the second author in Publications III and VI he contributed significantly to the writing of the manuscripts.

List of Abbreviations

AB-model	A two-state opinion formation model
AFG	Arenas, Fernández, and Gómez
CFinder	A program for obtaining k-clique communities
CP	Clique percolation
ER	Erdös and Rényi
RB	Reichardt and Bornholdt
si-model	Social influence -model
SCP	Sequential clique percolation
WSN	Weighted social network

List of Symbols

N	Number of nodes in a network
L	Number of links in a network
ν_i	Node i
W_{ij}	Link weight between nodes ν_m and ν_n
$A = [A_{ij}]$	Adjacency matrix of a network
$W = [W_{ij}]$	Weighted adjacency matrix of a network
k_i	Degree (number of neighbors) of node ν_i
$\langle k \rangle$	Average degree of nodes in a network
$P(k)$	Degree distribution of a network
$\bar{k}_{nn,i}$	Average degree of neighbors of node ν_i
$\bar{k}_{nn}(k)$	Average degree of neighbors of nodes of degree k
$P(k, k')$	Probability that a link exists between nodes of degree k and k'
s_i	Strength (total link weight) of node ν_i
c_i	Clustering coefficient of node ν_i
$c(k)$	Average clustering coeffient of nodes of degree k
$\langle c \rangle$	Average clustering coeffient of nodes in a network
$\mathcal{N}(\nu_i)$	Neighbors of node ν_i
N^i_Δ	Number of links between neighbors of node ν_i
Q	Modularity
Q_w	Weighted modularity
l_{mn}	Number of links between communities m and n
w_{mn}	Sum of link weights between communities m and n
K_m	Degree sum of nodes in community m
$[l_{mn}]_{p_{ij}}$	Expected number of links between communities m and n
γ	Tuning parameter of the RB-method and weighted RB-method
δ	Tuning parameter of the WSN-model
$R_{k=k^*}$	The relative size of the largest k^*-clique community
$\langle n_s \rangle$	Average size of non-giant k-clique communities

Chapter 1

Introduction

This Chapter provides an introduction to complex systems and networks research.

1.1 Background

Let us begin with a statement by the famous French mathematician and astronomer Pierre-Simon Laplace:

> ”We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.”
>
> *Pierre-Simon Laplace, 1814*

This statement summarises the essence of scientific determinism, which dominated the scientific thinking in the 19th century. This rather mechanistic view on nature was challenged in the beginning of the 20th century by the development of quantum physics, which, among other things, states that it is impossible to measure the position and momentum of a particle at arbitrary precision. Without going further into the philosophical debate related to determinism, it is clear that our inherently limited computational resources severely restrict the applicability of such a ”brute-force” approach to understanding nature. Currently, supercomputers are able to simulate certain molecular systems consisting of approximately one million particles for a time period of about 10^{-6} s. However, simulating the behavior of even a single organic cell by integrating the equations of motion for

each contained atom/molecule neither makes much sense nor is it possible. Instead, the key for predicting and understanding the behavior of the world around us is simplifying. This approach relies on the assumption that the simplified system resembles the original one in its relevant properties even though some details have been discarded. Evidently, choosing the proper level of simplification is not always easy. Albert Einstein has said: ”Everything should be made as simple as possible, but not one bit simpler.” This is one of the leading ideas in complex systems research as discussed below.

1.2 Complex systems

In the times of Leonardo da Vinci it was possible for a single person to master practically all branches of science. Since then, the scientific field has been segregating to more and more specialized branches each having their own experts with profound understanding of a specific subject, such as atoms, molecules, neurons, animals, planets, *etc.* This reductionistic approach, which concentrates on the properties of single objects and their interactions, has greatly increased our understanding of these objects. However, there exits numerous systems whose behaviour can not be understood by considering the constituent elements in isolation. In such systems the interactions between elements are usually strong and nonlinear, and the systems exhibit complicated collective behavior. Consider, e.g., climate, cells, human society, economics, *etc.* The approach which attempts to explain the behavior of such systems from the properties and interactions of its parts is known as complex systems research.

Wikipedia[1] defines a complex system as follows:

> A complex system is a system composed of interconnected parts that as a whole exhibit one or more properties [...] not obvious from the properties of the individual parts.

A simple example is traffic, where the flow of cars may suddenly transform to complex patterns of slowly moving jams (1). This behavior can be understood by computer simulations where each driver has simple rules governing his behavior (attempt to maintain constant speed and avoid collisions). In particular, the drivers are not controlled by a supervisor nor do their rules contain instructions related to collective behavior. Instead, *the collective behavior emerges from the individual actions of agents.* Another example is the fluctuations of stock prices, which amplify every now and then to massive bubbles and crashes when stock brokers try to guess what other stock brokers will do. Also, biology is full of complex systems: brain, metabolism, food webs, *etc.* In fact, complex systems showing spontaneous, emergent behaviour can be found in almost all branches of science

[1] An open content encyclopedia, see http://wikipedia.org

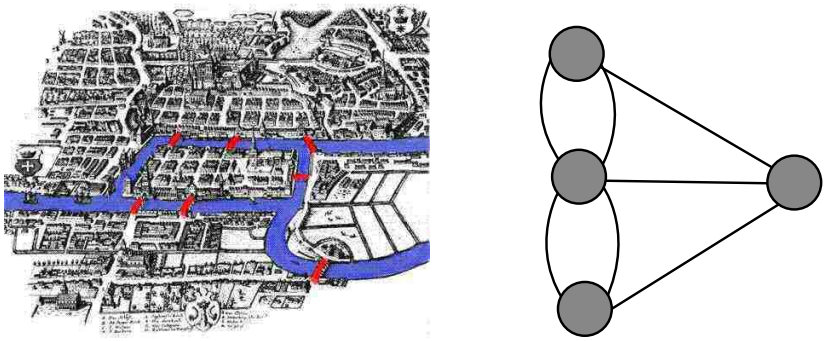
and the multidisciplinary nature of complex systems research often requires that experts of different fields combine their skills for understanding them.

It should be noted that large number of interacting agents does not always lead to complex behavior. An example is a box containing one liter of air. This box contains approximately 10^{23} molecules such that integrating the equations of motion for each particle is far beyond the capacity of any supercomputer. What can we then say about this system? It turns out that it is not necessary to know the precise position and momentum of each molecule in the box to predict physically relevant quantities. Instead, this system is well described by the laws of statistical physics (2), which is a framework for relating the microscopic properties of individual atoms and molecules to the macroscopic properties of materials. Thus, statistical physics connects the microscopic behavior of particles and large-scale properties of the system. However, the systems considered in classical statistical physics, such as the example box of air, are often not complex in the sense defined above: molecules and their interactions may be complicated but the behavior of the whole is simple and can be described by a small number of equations. Of course, this is not always the case. Consider for example Bose-Einstein condensation (3), which is a collective phenomenon and can not be understood by studying the constituent atoms in separation. Phase transitions in general can be regarded as collective behavior, and related concepts and methods are often used in complex systems research. It can be argued that instead of classifying different systems as ”simple” or ”complex”, the complex systems approach is best distinguished from alternative approaches by the questions asked, rather than the specific systems studied.

1.3 Complex networks

Formally, a network is a collection of nodes and links connecting pairs of nodes. The links and nodes may be physical entities like routers and optical fibers of the Internet, or they may represent more abstract relations like networks of word synonyms. The study of networks has a history in mathematics under the name of graph theory. It was initiated by the mathematician Leonard Euler, who solved the Königsberg Bridge Problem in 1736. The question is if it is possible to walk a route that crosses each of the seven bridges of Königsberg exactly once, see Fig. 1.1(a). Euler’s solution was based on simplifying the problem by eliminating all features except the landmasses and bridges connecting them, which were mapped to nodes and links of a network, Fig. 1.1(b). By considering certain general properties of this network Euler was able to prove that such a walk is not possible. Since then, graph theory has become a highly developed field of discrete mathematics (4).

Pure graph theory deals mostly with regular, abstract constructions, which



(a) The seven bridges of Königsberg

(b) Network of bridges and landmasses

Figure 1.1: (a) A drawing showing the Königsberg bridges on the 18th century, and (b) the related network representing the landmasses and the connecting bridges.

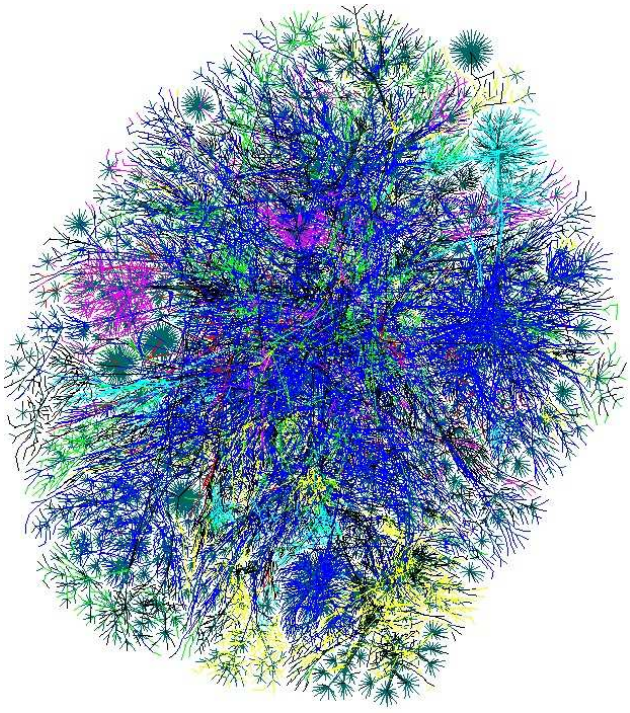
have little in common with real world networks. A significant advancement in graph theory took place in 1950s when Erdös and Rényi (ER) initiated the study of random graphs (5). A particularly famous random graph model is the ER-model, whose properties have been comprehensively analyzed and which served as a model for real world networks for several decades. Sociologists were among the first who noticed that networks provide an efficient way to study empirical data (6). A network could represent for example a system of people and their friendships. Networks were also used in the analysis of technological, engineering, and chemical systems (7; 8), and in computer science (9). Much of this work was limited to rather small data sets and research focused strongly on the properties of single nodes. Sociologists modeled social networks with various statistical models (10) but otherwise network modeling was rare.

The ”new science of networks”, or *complex networks* research, was born in the late 1990s when large electronic data sets and efficient computational resources allowed investigating large scale complex systems. In the network approach to complex systems the interacting elements and their interactions are mapped to network nodes and links, see Fig. 1.2 for examples. This is done in the hope that the essential properties of the system are not in the details of the elements but rather in their interaction patterns. Even though this simplification may appear crude, the simplicity and flexibility of the network approach have made it one of the primary tools in the modern complex system research.

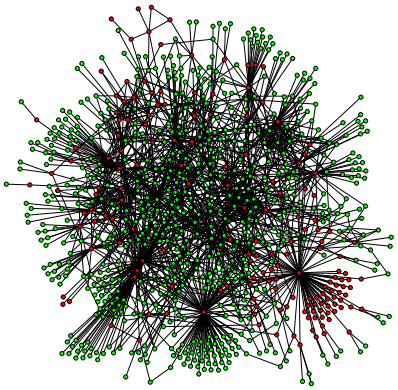
At the end of 1990s Watts and Strogatz (11) and Barabási and Albert (12) demonstrated in their seminal papers that real-world networks from very different

fields have certain universal properties. In particular, they found out that real networks are far from random ER-graphs. Both of these papers are representative of the modern network approach: in addition to analyzing large networks, they represent a model for explaining the results and provide new theoretical measures to augment the analysis. During the last decade network theory has developed extremely rapidly, for reviews see, e.g., (13; 14; 15; 16). Currently, network theory can be regarded as lying between statistical physics and graph theory, but it has also borrowed concepts from other fields of science, such as biology and network sociology.

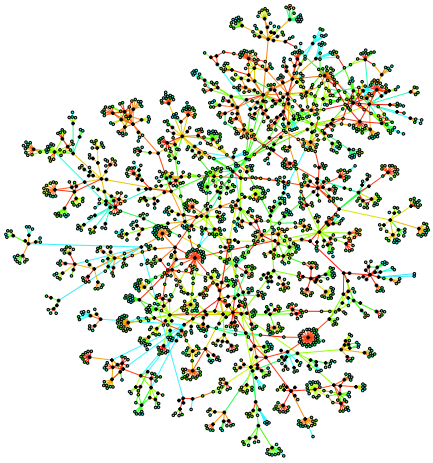
This thesis focuses on communities in complex networks – in particular their detection, emergence, and effects on dynamics. Early studies of networks focused on basic topological properties such as *degree distributions*, *clustering*, and *the small world effect*. These basic characteristics are reviewed in Sec. 2.1. Lately, the focus of network research has been shifting towards functionality and dynamics of networks. Functionality is often related to densely connected sets of nodes, which are called communities, modules, or clusters. The fundamental problem is to exactly define what constitutes a community and how such structures can be efficiently identified in large real networks. These issues were addressed by the author and coworkers in Publications I, II, V, and VI, and they are the subject of Sec. 2.2. Another concurrent focal topic of network research is dynamics – changes of the network structure and/or dynamics of processes taking place on networks. During the last decade, network models have turned out to be a valuable tool in the study of network dynamics, either by providing feasible artificial network structures for simulations of dynamical processes, or by attempting to explain the growth and change of the structure of real networks. Currently, there is a large number of models which attempt to capture the characteristics of various types of real-world networks. Social networks that map the connectivity patters of individuals are an important class of networks and a number of models have been proposed to explain their features. Early models focused on replicating the most salient unweighted characteristics, whereas models incorporating link weights and/or community structure have been rare. In Publication IV the author and coworkers addressed the problem of modelling weighted social networks with communities and Publication III deals with the effects these network characteristics have on a dynamical process taking place on such networks. The background for the social network analysis and results of Publications III and IV are reviewed in Chapter 3. Finally, Chapter 4 summarizes the research of this thesis.



(a) Internet



(b) Yeast



(c) Mobile phone calls

Figure 1.2: Illustrations of various networked systems: (a) Internet at the autonomous systems level (17), (b) part of a protein-protein interaction map of yeast (18), (c) part of a network of mobile phone calls (19).

Chapter 2

Structural properties of networks

Considering the networks in Fig. 1.2 it is immediately clear that networks are not homogeneous: they contain significant structure both at small and large scales. Measuring and characterizing the properties of networks is the first step in the process of understanding the origin and functionality of the underlying complex system. Below, Sec. 2.1 first reviews the most important network characteristics that are frequently referred to in the attached publications as well as in this introductory part. For a comprehensive review of network characteristics see Ref. (20). Having defined the basic network properties, we move on to mesoscopic network structures, i.e., communities, which is the main theme of research in this thesis. Section 2.2.1 deals with the general properties of communities and discusses the various choices one has to make when choosing a suitable community detection method. Section 2.3.1 reviews the properties of popular community detection methods based on global optimization and discusses their recently found limitations. The last part of this Chapter deals with a locally defined community detection method, k-clique percolation, whose properties are reviewed and a new algorithmic implementation of it is presented.

2.1 Basic characteristics

In general, a network may contain several types of nodes, its links can be directed and/or weighted, and there may be self-links and multiple links between a pair of nodes. However, in this work, all links are considered undirected, self-links and multiple links are forbidden, and nodes are of a single type unless otherwise mentioned. These simplifications are common in complex networks theory, with the exception that directed networks have received some attention, especially in the study of WWW and biological systems (21; 22; 23; 24; 25).

Matrix representation. Consider now a network consisting of N nodes and L (weighted) links. Such a network can be represented by a symmetric weight

matrix $W = [W_{ij}]$, where the element W_{ij} is the link weight connecting nodes ν_i and ν_j, and a zero weight indicates the absense of an edge. A related matrix is the adjacency matrix $A = [A_{ij}]$, where $A_{ij} = 1$ if a link exists between nodes ν_i and ν_j, and zero otherwise. Note that for most networks, only a small fraction of possible links is present, and hence the matrix representations are sparse.

Degree. The degree is a basic property characterising a single node, measuring its number of neighbors. The degree of node ν_i is given by $k_i = \sum_{j=1}^{N} A_{ij}$. A related quantity is the degree distribution $P(k)$, which tells the probability that a randomly chosen node has degree k. For the majority of real-world networks, $P(k)$ is fat-tailed, meaning that most nodes have a relatively small number of neighbors but some nodes, "hubs", have huge number of connections (15). It turns out that these hubs are very important in many ways: error and attack tolerance (26; 27), thresholds of spreading processes (28; 29), devising efficient search stragies in networks (30), *etc.* The most commonly encountered type of fat-tailed degree distribution in complex networks research is the power-law (or scale free) distribution where $P(k) \propto k^{-\gamma}$. This distribution has several interesting properties: as the name suggests, it looks similar on all scales and (depending on γ) leads to the existence of hubs of very large degrees. For $\gamma \in [2, 3]$, which is the range commonly associated with real-world networks, its second moment diverges, which is the reason behind many peculiarities related to behaviour of spreading processes and extreme robustness to vertex removal. After the first paper on scale-free networks by Barabási and Albert in 1999 a lot of work has been put to identifying such networks as well as developing models to explain their emergence, see e.g. Refs. (31; 32; 33; 34). Examples of scale-free networks include Internet, WWW, and scientific citation networks to name a few. Scale free networks are most often modelled using variants of the preferential attachment process (12), where high degree nodes are more likely to attract new links than low degree nodes. Interestingly, scale free distributions in nature have been observed earlier many times (35) and the related preferential attachment process has also been described earlier: Pareto's law of incomes and wealth in 1897 (36), Zipf's law of word frequencies in language in 1949 (37), abundance of species in taxonomic trees (38), populations of cities (39), and papers published by scientists (40; 41).

Degree-degree correlations. In social sciences it is well-known that popular people tend to know popular people. Considering the node degree as a measure of popularity, this means that high degree nodes have disproportionally many connections to other high degree nodes. In general, the existence of degree-degree correlations in a network means that the degrees of connected nodes are not independent. These correlations are fully described by the joint probability $P(k, k')$, the probability that a node of degree k is connected to a node of degree k'. However, $P(k, k')$ is difficult to measure empirically and thus it is often favorable to define the average nearest neighbor degree of a node ν_i as $\bar{k}_{nn,i} = \frac{1}{k_i} \sum_{j \in N(\nu_i)} k_j$,

where $\mathcal{N}(\nu_i)$ denotes the neighbors of ν_i. Averaging over nodes of degree k one obtains this distribution as a function of k, $\bar{k}_{nn}(k)$. If $\bar{k}_{nn}(k)$ is increasing, then the network is said to be *assortative*, and if it is decreasing the network is *disassortative*. Social networks are usually found to be assortative whereas many technological networks, like Internet, are disassortative (42).

Clustering. When considering the social relationships we have in our everyday life, it is common that the friends of a person are also friends themselves. This phenomenon is characterised by the clustering coefficient (11)

$$c_i = \frac{N_\Delta^i}{k_i(k_i-1)/2}, \tag{2.1}$$

where N_Δ^i is the number of edges between neighbors of i, and $k_i(k_i-1)/2$ is the maximum number of such edges. This means that c_i is confined to values between 0 and 1, where 1 is obtained only if all possible connections between neighbors are present, and 0 if none of them are present. Related quantities are the average clustering coefficient $\langle c \rangle = 1/N \sum_{i=1}^{N} c_i$ and the clustering spectrum $c(k)$, which is the average clustering coefficient of nodes having degree k. In most observed networks the level of clustering is high, where "high" means higher than in a random reference network of the same size (11). Often, $c(k)$ is proportional to k^{-1}.

Path length. A path in a network is simply a chain of links forming a connection between two nodes. The distance between two nodes is defined as the length of the shortest path connecting the nodes, where the length is defined as the number of links on the path. A basic characteristic of a network is the average distance between all pairs of nodes, or the maximum distance, which is known as the network diameter. A famous property of practically all real-world complex networks is the "small-world" phenomenon, which means that the network diameter grows slowly as a function of the number of nodes in the network, typically as $\log N$. Thus, even in very large networks, any two nodes are connected by a short path (in practice the length is of the order of ten or less). In regular lattices, on the contrary, the path lengths grow relatively fast as a function of system size[1] but even a small number of random links is enough to make the lattice a small-world while still keeping the local structure practically unchanged. This competition between order and randomness in the context of small-world networks was first studied by Watts and Strogatz in Ref. (11) and Watts has also written popular science books of the subject (43; 44).

Weighted characteristics. The above mentioned characteristics are related to purely topological network properties, in other words ignoring the possibility for link weights. Many real networks display a large heterogeneity in the intensity of connections in addition to complex topological structure. Examples include

[1]Diameter scales as $N^{1/d}$, where d is the dimension of the system

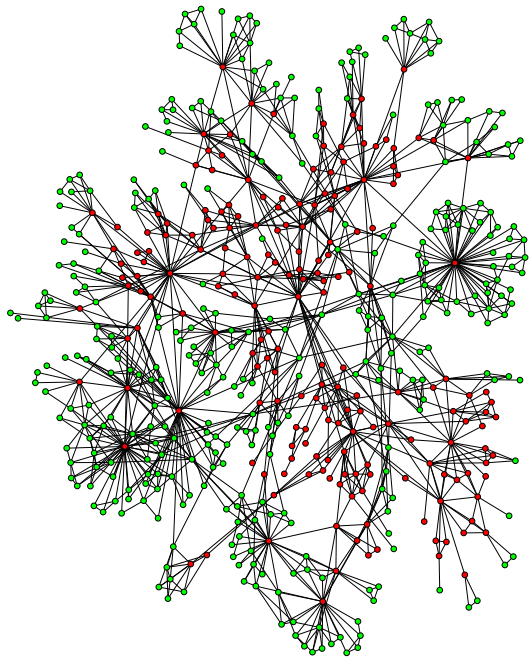


Figure 2.1: A network containing communities (51).

strong and weak ties in social networks (19), fluxes in metabolic reaction pathways (45; 46), unequal traffic on the Internet (47), or the number of passengers in airline networks (48). In general, the weights introduce an additional degree of freedom, which often provides significant new information about the system under study but it also complicates the analysis of observed networks as well as the analytical treatment of weighted network models. Some unweighted properties generalize easily to the weighted case, e.g., degree, which becomes the node strength $s_i = \sum_{j=1}^{N} W_{ij}$ (48). On the other hand, the weighted clustering coefficient can be defined in several ways and it appears that none of the definitions is free of problems (49). Weighted motifs are measures of local network structure that have turned out to be useful tools especially in the analysis of financial and biological systems (50).

2.2 Communities

Above, we discussed measures characterizing networks at the level of single nodes (e.g. the degree) and at the level of the entire network. However, in many networks, there are structural features of high importance and interest, which fall between these two extremes. An important class of such mesoscopic structures are communities, which are sets of nodes with denser connectivity to each other than to the rest of the network. Figure 2.1 shows an example network having a fairly clear community structure. Communities are important because they can often be closely related to functional units of the system, e.g., groups of individuals interacting with each other in a society (52; 53; 54), WWW pages related to similar topics (55), compartments in food webs (56), or proteins related to cancer metastasis (57). Identifying communities can be regarded as coarse graining the network, but it should be noted that community detection may also provide information about the roles of individual nodes. For example, a node at the boundary of a community may work as an important mediator between communities whereas a central node provides control and stability to the community (58).

Identifying communities has received a lot of attention during the last years and the field is still rapidly evolving (59). The problem is twofold: first, there is no unique way to rigorously define what constitutes a community, and second, any definition is useful in practice only if it allows for sufficiently fast algorithmic implementation such that even very large networks can be processed. In the following some general properties of communities are first reviewed. Then, selected recent community detection methods are presented and their properties are analyzed. In particular, the resolution properties of methods based on global modularity optimization are discussed and a new algorithmic implementation of the k-clique percolation method is presented.

2.2.1 Communities: properties and definitions

During the recent years the problem of detecting communities has received a lot of attention and thus it is somewhat surprising that a unique, widely accepted definition of what constitutes a community does not exist. Intuitively, a community can be viewed as a collection of nodes which are densely connected to each other while being only sparsely connected to the rest of the network. Consider, for example, a network of the people of a small village, where the nodes representing the individuals are connected if the corresponding persons have met during a certain time period, say, one week. It is easy to imagine that families appear as densely connected sets in this network whereas connections between families are sparse. Perhaps, if given a visualization of the social network of the village, we would be able to identify the communities in it. However, the intuitive definition has to be formulated in more rigorous terms for computers such that the analysis

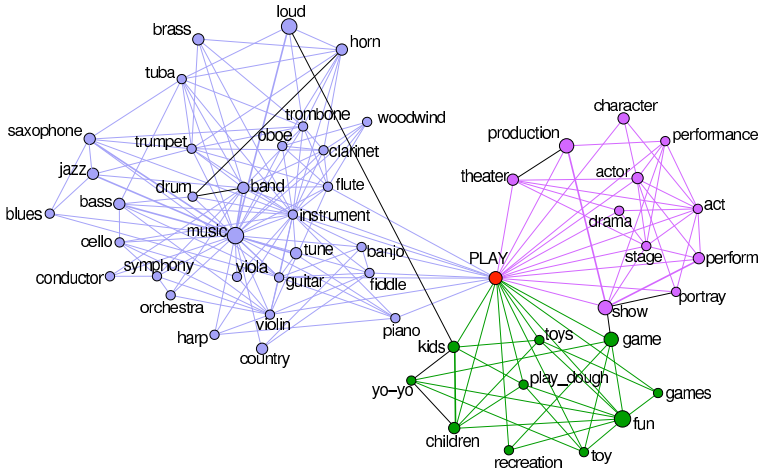


Figure 2.2: The word ”play” belongs to three overlapping communities in a word association network. Figure published in Ref. (66), copyright (2005) Nature Publishing Group.

of networks consisting of millions of nodes becomes possible.

Defining exactly what constitutes a community is unfortunately a tricky task such that currently a wide variety of definitions exists. For every definition several choices have to be made: whether the definition relies on global or local network properties, whether nodes can simultaneously belong to several communities, whether the link weights are utilized, and whether the definition allows for hierarchical community structure. Global methods, as the name suggests, utilise the whole network structure for defining the communities. This can be achieved in several ways, e.g., by global optimization methods (60; 61), by division algorithms based on different global centrality measures (52; 62), by spectral and synchronisation methods (63; 64), or by recently presented information-theory-based methods (65; 25). Local methods, on the other hand, define communities based on purely local network structure. Examples are the k-clique percolation method (66) and the recent method by Lancichetti et al. (67), which is based on finding ”natural” communities around nodes by optimizing a local fitness function. Global methods have been popular during the last decade but it appears that recently local methods have gained more popularity. This may be due to the resolution problems of several popular global methods (Sec. 2.3.2). In addition, local methods are often computationally more effective than global ones.

In general, it is possible that some nodes belong simultaneously to several communities, in which case the communities are said to *overlap*. Overlapping communities are especially abundant in social networks but they also appear, e.g.,

in biological networks. An example of overlapping communities is presented in Fig. 2.2, which shows three communities sharing the word ”play” in a word association network: one comprising drama-related words, one music-related, and one children-related. Currently, only a few methods are able to deal with overlapping communities (66; 67; 68).

Another difficulty in community detection is that networks may contain hierarchical structures, which means that communities may be parts of even larger communities. This leads to the problem of choosing the best partitioning among different alternatives. One solution was suggested by Girvan and Newman, who introduced the concept of modularity as a measure for the goodness of a partitioning. The modularity will be discussed in Sec. 2.3.1. Sometimes it may be better to study the community structure from the point of view of nested hierarchy instead of choosing a single community partitioning (69; 70; 71), which is the motivation for the multiresolution methods discussed in the following sections.

2.3 Community detection methods

Given the various choices in defining a community it is natural that a large number of methods and related algorithms have been proposed over the years. Traditional methods used in graph theory, such as hierarchical- and K-means clustering (72), usually perform poorly for large networks as they strongly restrict the solutions (the number of communities may be fixed *etc.*) (59). Thus, most community detection methods used for complex networks have been developed during the last decade. This section deals with several important globally and locally defined community detection methods. The global methods referenced in this work are modularity optimization, the RB-model, and the AFG-model[2], whereas the k-clique percolation method of Sec. 2.3.4 belongs to local methods.

Global optimization methods have been widely used and they have been shown to perform well for many test networks. However, it was recently shown that the most popular method, modularity optimization, suffers from serious limitations (73). These results were generalized by the present author and coworkers for related, more general global optimisation methods as discussed in Sec. 2.3.2. The k-clique percolation method represents a completely different approach to community detection and does not suffer from the above mentioned shortcomings. Clique percolation has also other advantages compared to most alternatives but it should be remembered that it is not free of problems, e.g., it is not designed for hierarchical community structure and its definition is somewhat rigid such that some nodes may not be assigned to a community at all. The main limitation of clique percolation has been the poor performance of the existing algorithmic implementations for large networks. A solution to this issue was recently presented

[2]See page xiii for the list of abbreviations.

by the present author and co-workers as explained in Sec. 2.3.5.

2.3.1 Global optimization: modularity and the q-state Potts method

The concept of modularity was introduced by Girvan and Newman in 2004 as a measure for the goodness of a given partition of a network into communities is (60). The idea is straightforward: compare the number of links inside communities to the expected number of links in a random reference network which contains no community structure. More precisely, the modularity Q is given by

$$Q = \sum_{m=1}^{q} (l_{mm} - K_m^2/2L), \tag{2.2}$$

where l_{mm} is the number of links inside community m, L is the number of links in the entire network, K_m is the sum of degrees of nodes comprising community m, and the sum is over all communities. The term $K_m^2/2L$ corresponds to the expected number of links inside the community m for a randomized graph of the same size and same degree sequence as the original network. Equation (2.2) shows that Q will obtain a high value, if the density of links within communities is much higher than expected by the random reference. The maximum value of Q is one and it can also obtain negative values, which corresponds to assigning nodes into communities such that communities are more sparse than the random reference. An example is a network having multipartite structure, that is, the network consists of several types of nodes and links exist only between nodes of different types (71; 74).

Originally, Q was used as a criterion for choosing the best community structure from a hierarchy of communities. In particular, Girvan and Newman suggested that the partition of the network which maximizes modularity is the best representation of the community structure of the network. Since Q is independent of the method of obtaining the communities, it soon became the object function to be maximized. One of the first algorithms for modularity maximization was the greedy method by Newman (75), which was later improved by Clauset *et al.* such that the algorithm scales to very large networks (76). Other approaches include simulated annealing (77), extremal optimization (78), and spectral methods (79; 80). Recently it has been shown that modularity optimization is an NP hard problem, which means in practice that only approximate methods can be used since the computational time for obtaining the true global maximum is likely to scale exponentially. Modularity optimization has been shown to perform well for several test networks (81) and it has been used in the analysis of numerous real networks.

A more general framework for community detection containing the modularity optimization as a special case was introduced by Reichardt and Bornholdt (RB)

(61; 82). Their method is based on the q-state Potts model, which was originally developed for modelling a system of interaction spins in a lattice (83). The idea of the RB-method is that the community indices of nodes correspond to spin states ($1 \dots q$) and the interaction between spins is ferromagnetic if the corresponding nodes are connected and anti-ferromagnetic if they are not connected. The goal is then to assign the nodes into communities in such a way that the energy of the system is minimized. This means that, as a result of the above-defined interactions, densely connected groups of nodes should end up having parallel spins and, correspondingly, spins in different groups should be different. The energy of the system can be written in a form which closely resembles Eq. (2.2):

$$\mathcal{H} = -\sum_{m=1}^{q} (l_{mm} - \gamma[l_{mm}]_{p_{ij}}), \tag{2.3}$$

where l_{mm} is again the number of links inside community m, $[l_{mm}]$ is its expected number given the null model p_{ij}, and γ is an adjustable parameter. The null model p_{ij} denotes the probability that a link would exists between nodes ν_i and ν_j if the network was random, i.e., in the absence of community structure. In the case of modularity, the randomized reference network is called the configuration model, which is one possible choice for the null model also for the RB-method. The configuration model assumes that all links are randomly rewired while preserving the degrees of the nodes, which leads to connection probability $p_{ij} = k_i k_j / 2L$ (61). Thus, two high degree nodes are more likely to be connected than two low degree nodes. Another reasonable option for the null model is a constant link probability $p_{ij} = p$, corresponding to the classical Erdös-Rényi random network model (5). The parameter γ allows to explore communities of different sizes such that small values yield large communities and vice versa for large γ. Note that the RB method reduces to modularity optimization when $\gamma = 1$ and the configuration model is used for the null model. Regardless of the null model and choice of γ, the energy (2.3) has to be optimized using approximate methods. A simple way to do this is simulated annealing with single spin flips and block-flipping, but other optimization methods are also available, e.g., extremal optimization (78).

2.3.2 Limited resolution of global optimization methods

Despite of the popularity of the global optimization methods, their properties were until recently not properly studied. For example, it would be natural to assume that the modularity Q would obtain values close to zero for a random network which contains no community structure. However, this is not the case, and it has been shown that the maximum of Q may be close to one even in randomized networks (77; 84). Essentially, this is due to random fluctuations. Authors in Refs. (77; 84) suggest that the obtained maximum value of Q should always be

compared to that of a corresponding randomized network. This is the first sign of the problems of modularity, because in the definition of Q, modules are already compared against a random reference, and the above results indicate that the obtained value of Q still has to be compared to the corresponding value for a random network.

Another perhaps even more surprising property of modularity was revealed in 2007 by Fortunato and Barthélemy, who showed that modularity optimization is unable to detect small communities in large networks (73), a phenomenon known as *limited resolution*. This result holds even if the communities are as clear as possible, i.e., complete cliques connected with single links. By considering several example cases Fortunato and Barthélemy demonstrated that the resolution limit depends on the interconnectedness of the communities, and even considerably large communities may remain unresolved. Analysis of real-world networks showed that a significant fraction of modules may appear merged when they are detected using modularity optimization and should be further investigated.

The analysis and examples presented by Fortunate and Barthélemy efficiently demonstrated the problems of modularity optimization, which immediately raised the question whether the more general RB-method suffers from the same limitations. In particular, what is the role of γ and the null model with respect to the resolution limit? These issues were addressed by the author and coworkers in Publication I as discussed in the following.

Let us now briefly go through the derivation of the resolution limit for the RB-method, which was the main result of Publication I. A similar analysis was performed also in Publications II and VI. Suppose that we have an unweighted network consisting of N nodes and L links, and the nodes have somehow been assigned to communities. Note that the assignment can, but does not have to correspond to the actual community structure of the network. Consider now two communities, m and n, which are connected to each other by l_{mn} links. These communities should be merged by the optimization process if this would lead to a decrease of the energy (2.3). This condition can be shown to be equivalent to

$$l_{mn} > \gamma[l_{mn}]_{p_{ij}}. \tag{2.4}$$

where l_{mn} is the number of links between the communities and $[l_{mn}]_{p_{ij}}$ is its expected value, given the null model. For the configuration model, it is known that $[l_{mn}]_{p_{ij}} = K_m K_n / 2L$, where K_m is the degree sum of nodes in community m. Thus, for constant γ, a sufficiently large network size will always trigger merging because the right hand side of Eq. (2.4) goes to zero when $L \to \infty$. This result is illustrated in the case of a simple example network in Fig. 2.3. It turned out that this property is not restricted to using the configuration model as a null model. Instead, it applies to any reasonable null model and is essentially a consequence of the fact that the average probability of connecting nodes decreases when the

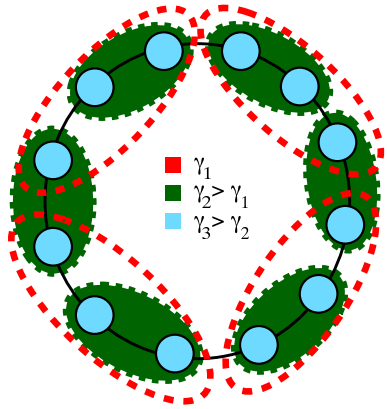


Figure 2.3: A schematic of the effect of the tuning parameter γ on the resolution of the RB method. The ring-like network consists of fully connected sets of nodes (cliques) that are connected by single links. When γ is low, consecutive cliques are combined to communities in the optimization process. Increasing γ increases the resolution and the optimal value of γ depends on the network parameters. When the network structure is more complex there is no simple way to decide the optimal value of γ.

system size grows and $\langle k \rangle$ remains constant[3]. Hence the expected number of links between two communities will decrease when N increases, which triggers their merging. Of course, γ can be used to probe communities of different sizes, but there is no simple way to decide which γ gives the most relevant communities. In addition, the suitable range of γ depends strongly on network size N and the method may also indicate the presence of a hierarchical community structure even when there is none, as it tends to artificially merge communities.

Arenas, Fernández and Gómez (AFG) have recently proposed a method for augmenting modularity optimization, which allows tuning the resolution of the method (85) somewhat similarly to the above RB method. The trick is to add a self-link of weight r to each node. This changes the total weight in the network and thus affects the sizes of the obtained communities. The proposed weighted modularity Q_w resembles Eq. (2.3) and the effect of adjusting the parameter r is qualitatively similar to adjusting γ in the RB-method. The connection between these methods was studied in Publication II. It turned out that the behaviors of the resolution as a function of γ and r are not identical, but essentially all conclusions about the resolution properties of the RB-method apply to AFG-method as well. In numerical tests their performance was seen to be very similar.

In addition to AFG modularity optimization, the paper by Arenas *et al.* (85)

[3]The number of links grows linearly and the number of node pairs quadratically.

introduced the concept of stable regions, which is a useful tool for community detection methods involving continuous tuning parameters. ”Stable regions” denote ranges of tuning parameter values where the obtained community structure essentially remains unchanged. Such regions seem to correspond to ”relevant” community structures of a network. The criterion proposed by AFG for monitoring the stability was to plot the number of communities as a function of the tuning parameter and observe flat regions in such a plot. As discussed later in this work, this criterion may be somewhat crude for real networks. Supposing that the stable regions have somehow been identified, there is still no simple way to judge which of the stable partitions are the most relevant ones. As noted by AFG, if such a measure existed it should be used as the function to be optimized. Nevertheless, sweeping the tuning parameter and observing stable regions allows for studying the community structure at multiple resolutions and gives hints which partitions may be most relevant to the functionality of the network.

2.3.3 The weighted RB-method

The study of networks focuses mostly on systems whose interaction structure is sparse, that is, $\langle k \rangle \ll N$. However, there are also networks whose natural representation is a full or very dense network, such as, correlation and distance based networks (86). In such cases the functionality and structure is encoded mostly in the link weights rather than topology. Currently, however, most community detection methods are unweighted and can thus be applied only to sparse networks. Often dense networks are transformed to sparse, unweighted networks by removing a fraction of the weakest links and considering the remaining links as unweighted (66). This is known as *thresholding*. Setting the proper threshold weight below which links are neglected is a non-trivial task and it is clear that this approach may discard valuable information of the system. Thus, there is a need for weighted community detection methods which utilise the link weights as well as the topology. A number of such methods have already been introduced (67; 85; 87; 88; 89; 25; 74). The various approaches taken in the above mentioned references show that there is not a single obvious (or optimal) way of incorporating the weights in community detection. In the simplest cases the weighted method is obtained by a change of notation, i.e., node degree k_i becomes node strength s_i and so on. While this approach often works fine, the theoretical justification for the obtained weighted methods may not be as solid as in the case of unweighted methods. As an example, in Ref. (85) the authors do not define the used null model. It should be noted that usually the intra-community connections are implicitly assumed to be strong whereas inter-community links are assumed to be weak. If this is not true for a particular network, one is forced to use an unweighted method or to modify the method or the network weights accordingly.

The RB-method was originally introduced in the context of sparse, binary net-

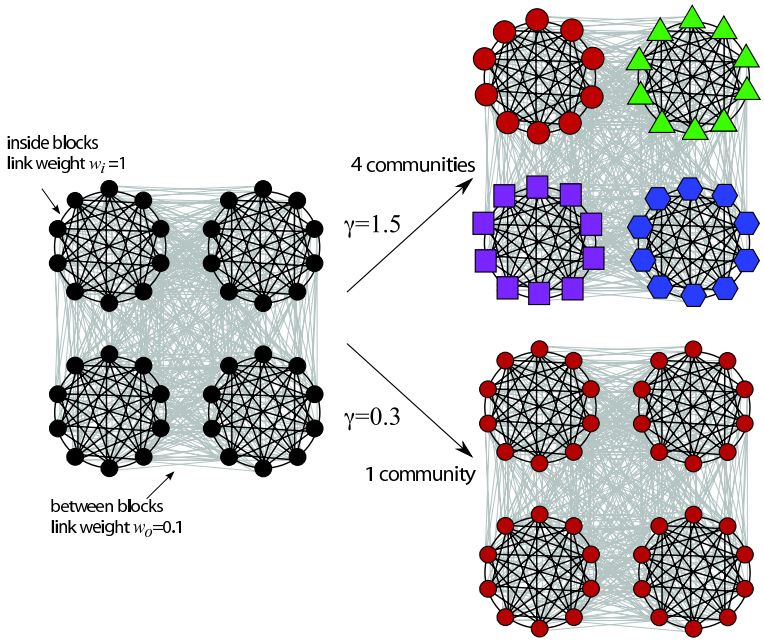


Figure 2.4: *Left:* A network consisting of $N_b = 4$ blocks each having $N_c = 10$ nodes. Links inside blocks have weight $w_i = 1$ and nodes in different blocks are connected with links of weight $w_o = 0.1$. *Right:* An illustration of the effect of γ on the found modular structure. Large values yield the physical communities while for small values the communities appear as one large module. If the number of blocks N_b is large enough, the networks size does not affect the value of γ where merging happens. The figure is originally published in Publication V.

works but edge weights can readily be taken into account as Reichardt and Bornholdt mentioned in Ref. (61), without proposing a null model. In Publication VI, we derived the required weighted null model for the RB-method, discussed its properties, and applied it to a fully connected stock correlation matrix. In short, Eq. (2.3) was modified such that strong links inside communities greatly lower the energy and the expected weight inside a community depends on the strength sum of the nodes in the community.

Before applying the weighted RB-method to real data, its resolution properties were analyzed theoretically. The question was when are connected communities properly resolved, and how does this depend on community sizes, link weights, number of connecting links, γ, and N. It turned out that the resolution behaves differently in sparse and dense networks. For sparse networks the resolution behaves similarly to the unweighted RB-model, namely, increasing the network size

decreases the resolution and weak connections between communities reinforce this behavior, as expected. For the case of dense networks, we considered a simple example network where all links were present and the modular structure was purely encoded in the weights. This network had N_b communities each consisting of N_c nodes and the weights were assigned such that inside the communities the links had weight $w_i = 1$ and links between nodes in different communities had weight w_b ($w_b \leq 1$), see Fig. 2.4. Similarly to the analysis in Sec. 2.3.2 the question was: when does the weighted RB-method merge two built-in communities and how does this depend on the network size? We found out that, similarly to the unweighted RB-method, merging is controlled by γ, i.e., small values yield large communities and *vice versa*. However, in the limit of large network size the resolution does not depend on network size. This happens because the expected weight between communities $[w_{mn}] \approx N_c^2 w_b$ is independent of the number of build-in communities N_b. Further analysis showed that small communities appear merged easier than large ones when the weight between communities is increased by replacing some of the connecting links by stronger ones ($w_{ij} > w_b$), but the network size still does not affect the suitable range of γ. For practical purposes issues such as the distribution of weights both within and between the communities are expected to affect these results. In general, analysis of the example networks in Publication VI suggests that the resolution of the weighted RB-method does not necessarily decrease when dense networks grow in size, unlike in the case of sparse networks. In practise this has the consequence that the range of γ does not depend on N and the stable regions (see below) are best studied as a function of γ instead of $\sqrt{\gamma}$ as is done in the case of sparse networks.

In addition to theoretical considerations the weighted RB-model was applied to a correlation-based network of stock return time series. The structure of this fully connected correlation matrix had earlier been investigated using a wide variety of approaches (91; 92; 86; 93), which allowed us to compare the communities obtained by the weighted RB-method to the known properties of this network. Of particular interest was the question whether the method is able to reveal stable regions in the community structure as a function of the tuning parameter γ and whether these stable communities are relevant with respect to different business sectors. The results of the numerical optimizations showed that the number of communities increases continuously as a function of γ, meaning that stable regions in the sense mentioned at the end of Sec. 2.3.2 were not seen. However, an analysis of the sizes of the two largest communities revealed a stable region in the range $\gamma \in [1.4, 3]$, suggesting that the stability of the components should be investigated using a variety of criteria. Figure 2.5(a) illustrates the classification of the companies according to Forbes (90), Fig. 2.5(b) shows the obtained communities when $\gamma = 1$ (weighted modularity), and Fig. 2.5(c) when $\gamma = 1.4$. There it can be seen that $\gamma = 1$ yields four communities, which appear to be combinations of the Forbes communities. For $\gamma = 1.4$ (corresponding to the onset of the sta-

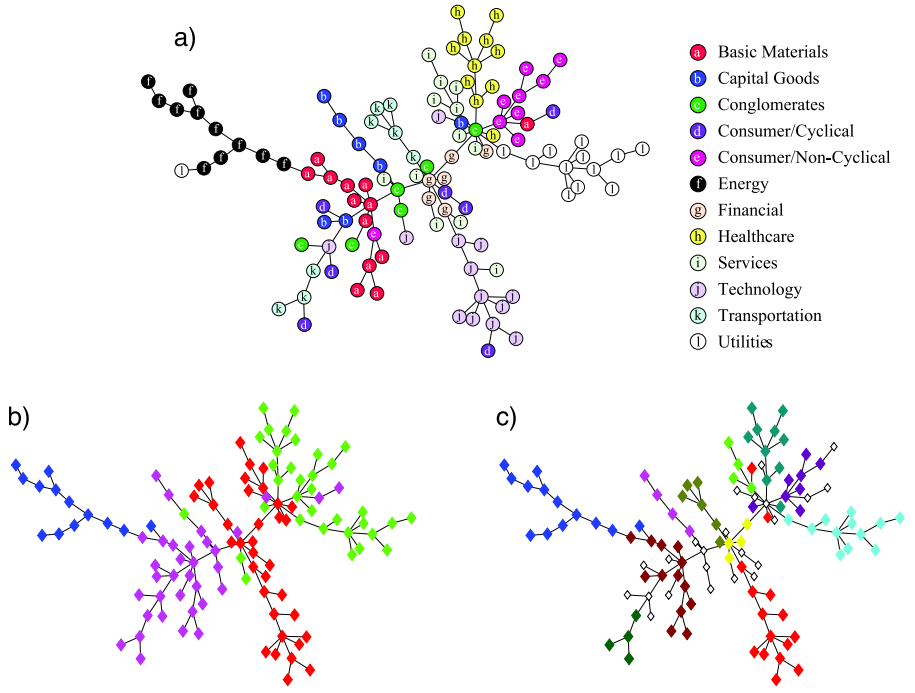


Figure 2.5: (a) The maximal spanning tree and business sectors according to Forbes (90). (b) The maximal spanning tree and the modular structure obtained with the RB method for $\gamma = 1$. Each color corresponds to a module. (c) The maximal spanning tree and the modular structure obtained with the RB method for $\gamma = 1.4$. Modules of size larger than two are depicted by different colors and the rest of the nodes by empty symbols. The figure is originally published in Publication VI.

ble region) the obtained communities correspond remarkably well to the Forbes classification.

According to the above mentioned theoretical and practical results it seems that the weighted RB-method provides a feasible approach for multiresolution analysis of communities in dense weighted networks. Practical limitations are in the problem of optimizing the energy function, which has to be done by some approximate, stochastic method (61; 78). For large systems the energy surface may be complicated and it is difficult to verify if the obtained solution is close to optimal. One of the problems is that structurally very different partitions may result in only a very small change in the energy. On the other hand, this can be interpreted as a sign that both alternatives represent a feasible community structure.

2.3.4 Clique percolation

Let us now move on from global optimization methods to a local community detection method, namely k-clique percolation. A clique, or more precisely a k-clique, is a set of k nodes that are all connected to each other, e.g., a 3-clique is a triangle. As discussed in Sec. 2.1, triangles are abundant in most studied networks which is seen as the high value of clustering. When it comes to community detection, it seems reasonable to argue that most triangles are confined inside communities. In addition, the larger the value of k, the more strictly the k-cliques are restricted to communities because connections between communities are sparse. The clique percolation (CP) method published by Palla et al. in 2005 (66) relies on these observations and on the notion that the members of a community should be reached through densely intra-connected subsets of nodes. A k-clique community is defined as a set of nodes that can be reached by "rolling" a k-clique over adjacent k-cliques, where adjacency means that the k-cliques share $k-1$ nodes, see Fig. 2.6 for illustration. Thus, the clique percolation algorithm defines the communities by considering overlapping chains of small cliques, which are likely to explore a significant fraction of each community, but are not able to "roll" from one community to another.

The choice of k has a significant effect on the found community structure. Typically used values of k are between 3 and 6. High value of k yields tight, internally cohesive communities whereas small values of k yield looser and larger communities. Especially for small k it is possible that adjacent k-cliques percolate through most of the network, which means that there is a single giant k-clique community which can comprise almost the entire network. In such cases the alternatives are either to use a larger value of k, or, if the network is weighted, to remove a fraction of the weakest links similarly to the thresholding in Sec. 2.3.3. Palla et al. suggested that the most representative community structure for a network is obtained by choosing the largest value of k where the giant component may emerge, and then removing all links below a threshold weight w^*, where w^*

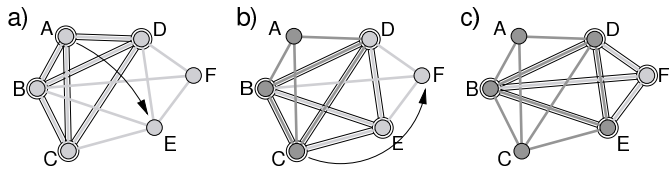


Figure 2.6: Illustration of a k-clique rolling for k=4. Initial clique $ABCD$ first rolls on to $BCDE$ by releasing node A and finally it rolls to $BDEF$ by releasing node C. All nodes reached by this type of rolling belong to the same k-clique community. Figure published in Ref. (66), copyright (2005) Nature Publishing Group.

is chosen such that the giant component is just about to emerge (66). It should be noted that it is impossible to know *a priori* the suitable value for w^*; instead, it has to be found by trial and error.

One of the benefits of the clique percolation method when compared to other community detection methods is that it allows for overlapping communities. As discussed in Sec. 2.2, overlapping communities are abundant in many types of networks, in particular in social and biological networks. Therefore, methods that do not allow overlap are inherently suboptimal for the analysis of such networks. Other benefits of the clique percolation method are that it is based purely on the local topology and it is deterministic meaning that approximative optimization methods are not needed. Successful applications of the CP method include detection of protein communities related to cancer metastasis (57), analysis of communities in co-authorship, word-association and protein-interaction networks (66), and time evolution of social groups (94). In the last of the above examples the authors define an interesting way to track communities in time. In general, this is complicated because communities can appear, merge, grow, break, or disappear as time goes on. The tracking method introduced in Ref. (94) utilizes the properties of k-clique communities in a clever way, which, however, is beyond the scope of this introduction. The main limitations of the CP method are that some nodes may not be assigned to a community at all, hierarchical structure is mostly neglected, the performance of current implementations is limited for large networks, and in general networks whose average degree is very small are problematic because they may contain only a small number of k-cliques and thus most nodes are not assigned to any community.

The popularity of the CP method has been boosted by the publication of a freeware program called CFinder (95), which provides a graphical user interface for performing clique percolation analysis. In the following, the CFinder algorithm is briefly reviewed in order to compare it to a later implementation. The CFinder

algorithm begins by finding all maximal cliques in the network, i.e., cliques which are not subcliques of larger cliques. The algorithm then proceeds by constructing a clique-clique overlap matrix, whose rows and columns correspond to maximal cliques and the matrix elements denote the number of common nodes for each clique pair. Finally, all matrix elements smaller than $k-1$ are set to zero and other elements to 1 with the results that connected components in this thresholded matrix correspond to the k-clique communities. This approach allows finding communities for all values of k from the same overlap matrix. However, if weight thresholding of the original network is needed, then the whole process including construction of the clique-clique overlap matrix has to be done separately for each threshold weight w^*, which is time-consuming. In addition, finding the maximal cliques may be time consuming for large networks. In general, the performance of the algorithm heavily depends on the network structure.

2.3.5 Sequential clique percolation

The limitations of the maximum-clique algorithm of CFinder encouraged the present author and coworkers to develop an alternative clique percolation algorithm. The goal was to design a method which allows rapidly obtaining the community structure corresponding to multiple weight thresholds and which would be capable for analyzing networks consisting of millions of nodes in a reasonable time. The developed sequential clique percolation (SCP) method fulfils these goals and provides also a method for visualizing the hierarchical structure of k-clique communities in weighted networks. In the following, the basic concepts of the SCP method and the related visualization scheme are briefly reviewed, for details see Publication V.

In contrast to the maximum-clique algorithm used in CFinder, which starts from the full network and proceeds by identifying the maximal cliques, the SCP algorithm takes the opposite approach. First all links are removed from the network, and then they are restored one-by-one while detecting k-cliques and the corresponding k-clique communities as they emerge. For unweighted networks the links can be inserted in an arbitrary order, whereas for weighted networks it may be desirable to sort the links by their weights. Note that, unlike CFinder, SCP processes only a single value of k at a time.

The core of the SCP algorithm consists of constructing the k-clique communities continuously when new k-cliques appear. For each new k-clique, there are two possible cases: the k-clique can either form its own community or it can overlap with one or more existing communities. In the latter case, all overlapping communities merge to form a single community. Thus, efficient detection of community overlap is crucial for achieving reasonable performance. Our solution to this is summarized in Fig. 2.7, which shows how we continuously build and update a bipartite network between k-cliques and $(k-1)$-cliques (panels (a) and

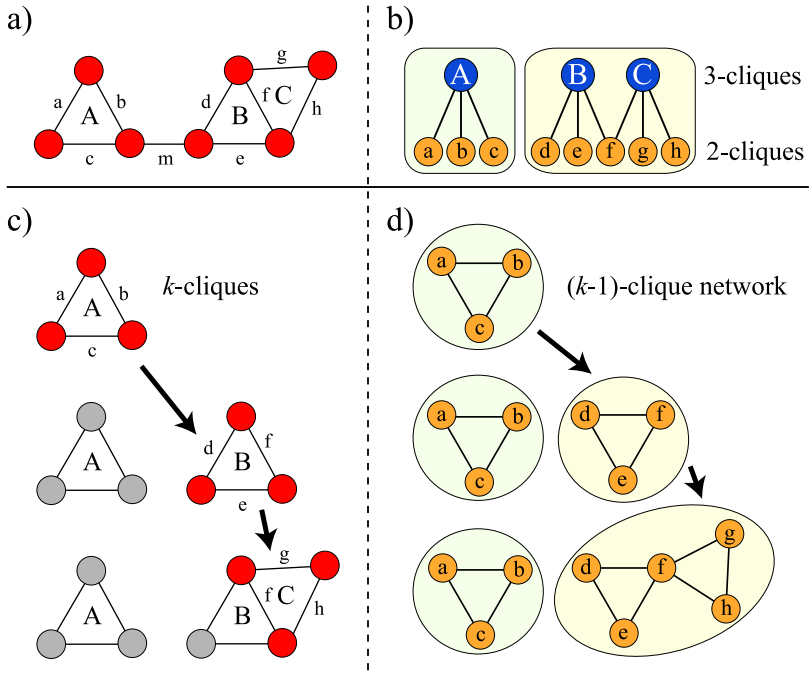


Figure 2.7: Illustration of the algorithm for detecting k-clique communities in a simple example network. Here, $k = 3$. a) The original network consists of three 3-cliques labeled A, B, and C. 2-cliques, *i.e.*, nodes connected by single links, are labeled with lower case letters. b) Bipartite network presentation of the clique structure. Note that in the bipartite network, the 3-cliques B and C, which form a 3-clique community, are connected by the shared 2-clique f. Clique A forms another 3-clique community. c) 3-cliques detected by the first part of the algorithm as links are sequentially inserted into the network. Each new k-clique is denoted by red nodes whereas nodes associated with existing k-cliques appear gray. d) Corresponding updates to the $(k-1)$-clique network as a result of the second part of the algorithm. k-clique communities correspond to connected components of this network (shaded areas). Figure published in Publication VI, copyright (2008) by the American Physical Society.

(b)), how this network is projected to $(k-1)$-cliques, and, how its connected components correspond to the actual k-clique communities (panel (d)). This process is very similar to ordinary link percolation, and actually in the practical implementation neither the bipartite network nor the projected network need to be explicitly constructed. Instead, only book-keeping of the community index of each $(k-1)$ -clique is needed, as well as updating the indices of the merging communities.

One of the benefits of the SCP method is that it allows for obtaining multiple weight thresholds w^* in a fast way. This is achieved by inputting the network links to the SCP algorithm in descending weight order. In this case the k-communities after each added link correspond to those of the original network thresholded with the weight of the last link. Thus, a large number of weight thresholds is obtained in a single run providing a clear advantage over the maximum-clique approach. The community detection process can be stopped at any point, which is useful especially for analyzing dense networks, where the problem is that the k-cliques start to percolate throughout the network sometimes already when only a small fraction of strongest links have been added. This is the case for example in the network of 354 product categories on the Finnish online-auction site Huuto.net, where the network is constructed such that nodes correspond to product categories and link weights correspond to the number of individuals having traded in both connected categories. This network has an average degree $\langle k \rangle \approx 250$, meaning that it is very densely connected. In Publication V this network was analyzed by tracking the formation and merging of k-clique communities during the process of adding links in descending weight order. This allowed us to construct a dendrogram of the communities, shown in Fig. 2.8, where the horizontal axis denotes the threshold weight at which each community forms or merges. It appears that the strongest communities are those where (supposedly) the majority of traders are women. Less than 1% of the strongest links were added before the giant component emerged. Tracking the formation and merging of communities and visualizing this process by dendrograms provides a multiresolution-type approach also for the k-clique percolation method, which had previously been regarded as a single resolution method.

The computational time of the SCP algorithm was shown to scale linearly as a function of the number of k-cliques for all networks, which was expected because each k-clique is processed exactly twice (detection and finding overlapping k-cliques). Scaling is often linear also as a function of N, but this is not the case for certain networks where the number of k-cliques grows faster than linearly when N increases. Such networks are typically projections of bipartite networks, such as coauthor (96) and movie actor networks (54; 11). For large bipartite networks it may thus be preferable to use recently published bipartite variants of the clique percolation method (97). In practice, the SCP algorithm is very fast for most networks and obtaining 4-clique communities for a 4 million node mobile phone call network (19) takes about one minute with a desktop computer.

The basic clique percolation method considers the link weights only indirectly through the network thresholding process. The weighted clique percolation algorithm (WCP) published in 2007 by Farkas et al. (89) improves this situation by assigning a weight I_c to each k-clique, where I_c is defined as the geometric mean of the link weights in the clique (also known as clique intensity (50)). Similarly to the weight thresholding in normal CP method, WCP involves thresholding such

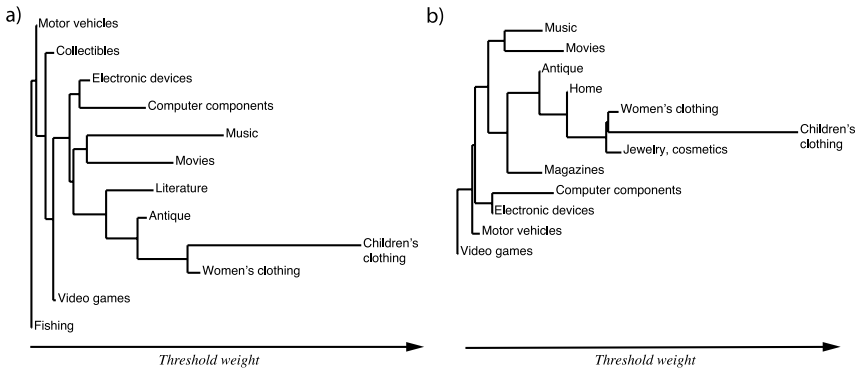


Figure 2.8: Dendrogram visualization of the community structure of the trading categories of the Finnish online auction site Huuto.net detected with the SCP algorithm for $k = 3$ (a) and $k = 4$ (b). Figure published in Publication VI, copyright (2008) by the American Physical Society.

that only cliques having I_c larger than a chosen limit I_c^* are included in the clique percolation process. The WCP method differs from the CP method because it allows the k-cliques to contain links weaker than the threshold I^* (c.f. w^* for the CP method) and can thus be argued to be more tolerant with respect to small noise in the link weights. For networks where strong links prefer to be neighbors of other strong links the methods give similar communities, but for networks where the strong links prefer to have weak links as neighbors the resulting communities can differ strongly. A small modification of the new sequential clique percolation method allows to use it also for weighted clique percolation. This is achieved by first detecting all k-cliques, then sorting them according to their intensities and finally constructing the communities by processing the k-cliques in descending order of intensity. Again, multiple intensity thresholds are obtained in a single run of the algorithm. Note also that the k-clique weight can be an arbitrary function instead of intensity, e.g., coherence or sum of link weights.

In conclusion, the clique percolation method discussed above has many desirable properties: it is deterministic, based on local topology, allows overlapping communities, resolution is tunable by changing k or threshold limit, and the new sequential clique percolation algorithm allows its usage for very large networks. The main limitations are the somewhat restricted treatment of hierarchies and the fact that a large fraction of nodes may not be assigned to any community. Compared to most other methods it is simple, reliable, and fast.

Chapter 3

Social networks

The recent high level of interest in social networks is largely due to their relevance to various processes taking place in society, such as spreading of cultural fads or diseases. In addition, analysis and modelling of social networks helps us to understand how the behavior of individuals and their interactions translate into large-scale social systems. In this Chapter the history of social network analysis is first briefly reviewed together with the most prominent features of social networks. Then, we move on to social network models, where of particular interest is the weighted networks model by the present author and co-workers. This model plays a central role in our attempt to explain the connection between microscopic rules of social tie formation and the formation of communities.

3.1 Background

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." (98) In general, a social network is a set of actors and relationships among them. The actors can vary from individuals to nations, and their relationships can be, e.g., friendship, kinship, sexual, business, or political contacts, or a mixture of those. Here, the focus is on networks between individuals.

Social networks were among the first networked systems to be studied. Sociologists conducted first empirical studies of social networks in mid-1930s, but it was not until the late 1950s that the sociological community started developing an interest in applications of graph theory (6; 99). Social relationships between individuals were typically mapped through questionnaires, which allowed taking a wide spectrum of interactions into account. However, acquiring respondents and processing such questionnaires was laborious and costly, restricting the studies to

small samples consisting of at most hundreds of individuals. In addition to small sample size, this approach has the problem that the interactions are based on individual recollection containing possibly significant individual biases. For example, it has been observed that the reported mutual interactions are often asymmetric, i.e., person A may name person B as a very close friend whereas B considers his relation to A as rather distant (100; 101).

Physicists entered the field of social networks (and complex networks in general) in the end of 1990s when the availability of large electronic data sets and efficient computational resources allowed studying systems whose sizes vary from thousands to millions of individuals (16). Earlier, sociologists had focused mostly on the properties of single nodes, e.g., a typical question being who is the most central or influential individual in a network? However, single-node-level properties become relatively meaningless in large networks, where the interesting questions are related to distributions, statistical properties and averages of various quantities. This approach is familiar for physicists who have plenty of experience in working with statistical mechanics, critical phenomena, *etc.*, and numerous concepts of physics have turned out to be very useful in the network analysis. The role of physicists in the field of social networks has concentrated to analysis of the topological properties of networks, developing new concepts, algorithms, and models, and applying the methods, whereas drawing conclusions about the social concepts has been mostly left for sociologists who have been slow in employing the physics methodology. Also, not all sociologists have been happy with physicist entering their field (10).

3.2 Structure of social networks

Large scale social networks which have been studied so far include e-mail (102; 103; 104), phone calls (19; 105), co-authorship (96; 106; 107), movie actors (11; 108), and instant messaging networks (109) to name a few. It should be noted that networks obtained from electronic data sets involve only a very restricted scope of social interactions, be it phone calls, number of e-mails, *etc.* Thus, it should be understood that the resulting network serves only as a proxy for the underlying ”true” social network. The advantage of electronic data sets is that they are large, relatively easy to process, and accurate in the sense that subjective biases are absent.

Perhaps one of the best-known concepts of network sociology is the ”small-world effect”, which was already discussed in Sec. 2.1. This effect was first described in a fictional story by Karinthy in 1929 (110) and in more scientific terms in an influential paper by Pool and Kochen in 1978 (111)[1]. The first experimental studies of the small-world effect were conducted by Milgram in the late 1960s

[1]This paper was written already in 1958 and circulated twenty years in preprint form.

(112). The idea of these experiments was to choose a target person and a set of random starting individuals, who were each given a package and instructions to pass the package to a person who they considered to be closer to the target person. However, the package was to be passed only to a person who was familiar by first name basis to the current package holder. The packages were supposed to travel by these rules until they reach the target person. The outcome of the study was that 64 of the 296 packages reached the target, the median number of intermediate acquaintances being 5.2. This lead to the famous phrase "six degrees of separation", which appears nowadays often in popular culture, including magazines (113; 114), TV series (115), plays and movies (116), and games (117; 118).

Another famous concept of sociology is the weak link hypothesis by Granowetter, which states that the overlap of the circles of friendship around two individuals increases as a function of the strength of the tie connecting these two individuals (119). Essentially, this means that strong ties bind communities together whereas weak ties function as bridges between communities, thus playing an important role for the flow of information and innovation. This phenomenon is known as "the strength of weak ties".

The recent picture emerging from analysis of large social networks constructed from electronic databases shows the following common (unweighted) features: i) degree distributions are skewed or fat-tailed[2], ii) high-degree nodes are disproportionally often connected to other high-degree nodes (assortative mixing), iii) the average clustering coefficient $\langle c \rangle$ is high compared to random networks, iv) the average shortest path lengths are small (the small-world property), and v) communities are abundant. Weighted characteristics have been studied in large scale systems only recently. One of the first studies was the analysis of a mobile phone call network consisting of millions of individuals by Onnela *et al.* (19; 120). Nodes in this network correspond to phone numbers and edge weights represent the total call time between the involved numbers. This network was shown to obey the above characteristics i)-v) and the analysis of the connection between network topology and edge weights supported the weak-link hypothesis, which, even though having been widely accepted in sociology, had never been empirically verified for a large data set. In particular, it was shown that weak and strong links have different roles in the network such that the weak ties maintain the global integrity of the network while strong ties maintain the communities. This affects, for example, the speed of information spreading.

3.3 Modeling social networks

The above mentioned empirical findings raise the question how such universal properties emerge from the actions of individuals, and whether this emergence

[2]Extremely high degree nodes are absent because no-one can have 10 000 friends.

can be modelled with simple microscopic rules. In the following, the general framework for modelling social networks is briefly reviewed and then a weighted social network model by the present author and coworkers is presented in more detail.

3.3.1 Basic rules

Modeling social networks helps us to understand how social networks form and evolve. In addition, artificial networks are useful in simulations of processes taking place in society, such as spreading of infection or rumor. Over the years, numerous models for social networks have been proposed. These range from the simple small-world model by Watts and Strogatz (11) to more complicated models, where agents may search for new contacts (121; 122; 123; 124), move (125; 126; 127), try to be different from the average (128), or the model may be formulated in a purely abstract mathematical form (129).

Basically all models which bear resemblance to real social networks involve randomness in the network construction rules, which is natural considering the complex and irregular structure of such networks. In the following, we concentrate on two important classes of social network models, namely topological models and spatial models. The difference between these is that the former utilizes only the network structure in its rules, whereas the latter assigns an intrinsic (random) coordinate for each node and close nodes are more likely to be linked to each other than distant nodes. In spatial models the node coordinates can be interpreted as real geographic coordinates or alternatively as coordinates in an abstract "social space", which may represent hobbies, opinions, occupation, *etc.* (130; 131). The reasoning for spatial models can be derived from the sociological concept of homophily (132), meaning the tendency for like to interact with like.

Topological models, on the other hand, try to mimic real networks by basing the network construction rules solely on the network topology. These models can be further divided to dynamic and growing models. Growth models (133; 51) may be a good choice for modeling systems where links are rarely removed and new individuals continuously join the system, e.g., online social networking systems. However, many social systems are essentially constant in size, or grow at such a slow speed that the growth can be neglected. This is the case for many social relationships encountered in our everyday life, e.g., friends, co-workers, people we meet at hobbies *etc.* For such systems the natural representation is a dynamic model, where the network size N is constant but links are continuously created and removed. In dynamic models the networks are constructed from an initial network by repeating the network evolution rules until the network reaches a steady state where its statistical properties do not change anymore.

Each topological model, growing or dynamically evolving, has to specify the rules for addition and removal of links. Network sociology identifies two funda-

mental processes for tie formation: *cyclic closure* and *focal closure* (102). Cyclic closure refers to forming ties with close network neighbors, whereas focal closure refers to forming ties independently of the network distance, and is attributed to forming social ties through shared activities. Focal closure is typically modeled simply by connecting random nodes. Note that usually this automatically guarantees short average path lengths in the network. The other tie formation process, cyclic closure, is responsible for the formation of ”non-random” structures in networks, and the most important form of cyclic closure is triadic closure, i.e., triangle formation (102). In a model by Davidsen *et al.* triadic closure was modeled by choosing two random neighbors of a node and connecting them (121). Marsili *et al.* took another approach, where each node searches for new ”friends” by forming links to nodes reached with two step random walks (123). It should be noted that the concepts of cyclic and focal closures did not appear in physics literature until 2007 (124) and, therefore, it seems that most topological models have been based on intuition and common sense of social tie formation instead of empirical findings.

For link removal there are two basic choices: removing nodes including all its links or removing single links. The node deletion process can interpreted such that the node leaves the visible scope of the network. For example, in a mobile phone call network an individual changes the operator. Since the network size is constant a new node must appear after each node removal. The latter link removal mechanism could correspond to dissolving and eventual disappearance of social ties between individuals (134).

Most likely an ultimate model for social networks should incorporate both spatial and topological mechanisms as well as both link deletion mechanisms. However, these would significantly complicate the model and a proper analysis of such a model would be extremely difficult. Currently, there are several topological models that are able to reproduce the essential topological properties of social networks, i.e., skewed degree distribution, assortative mixing, high clustering, and small path lengths. Some models contain also community structure (122; 51) but in general it appears that there is no single model which is able to simultaneously capture all topological properties (135). In addition, social network models that scale to larger than hundreds of nodes have so far been unweighted, which is a significant drawback considering the important role of link weights in real social networks (19; 119; 120).

3.3.2 Weighted social network model

In Publication IV the present author and coworkers presented a weighted social network (WSN) model, which attempts to be as simple as possible while still capturing the interplay between link weights and topology. The WSN model is a topological model, where links are created by weighted local searches and random

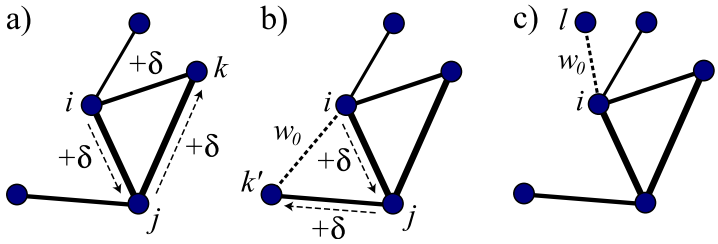


Figure 3.1: The rules of the local search for new acquaintances in the weighted social network model: node i chooses one of its neighbors, node j, with probability w_{ij}/s_i, where w_{ij} is the weight of the link connecting i and j and $s_i = \sum_j w_{ij}$ the strength of i. If the chosen node j has other neighbors apart from i, it chooses one of them, say k, with probability $w_{jk}/(s_j - w_{ij})$. (a): If k is already connected to i, this link is reinforced by δ. (b): If the search ends to a node that is not a neighbor of i, this link is established with probability p_Δ. (c): Node i can also create a link to random node l with probability p_r. In cases a) and b) the weights of involved links are increased by δ. Figure published in Publication IV, copyright (2007) by the American Physical Society.

attachment, and the number of links is controlled by random node removal. The essential difference to previous models is that the local search, which consists of two random steps from the initial node[3], utilizes the link weights such that the search most likely follows strong links. In addition, each time a link is visited in a local search its weight is increased by $\delta \geq 0$. This weight reinforcement process can not be justified by direct empirical findings, but it seems plausible that every interaction makes further interaction even more likely. For detailed explanation of the model rules see Fig. 3.1. Typical simulations were conducted for $N = 5 \cdot 10^4$ networks such that simulations consisted of 25000 iteration steps, which was enough to reach the steady state. Even larger networks could be generated if needed, the upper limit with current desktop computers is somewhere between 10^5 and 10^6 nodes depending on how many days one is willing to wait. Measuring statistical properties requires a large number of network realizations for each set of parameters, which accordingly increases the required computational time.

The most essential parameter in the model is the amount of weight reinforcement δ, while other parameters can be regarded as tuning parameters for achieving desired average degree and controlling the number of random links. The WSN network properties were thus studied as a function of δ. In the unweighted case $\delta = 0$, the WSN model was found to capture the basic properties of social networks except community structure [features i)-iv) of Sec. 3.2], which was ex-

[3]Each node searches for new acquaintances by first choosing (uniformly) randomly one of its neighbors, say node ν_j, and continues by choosing a random neighbor of node ν_j.

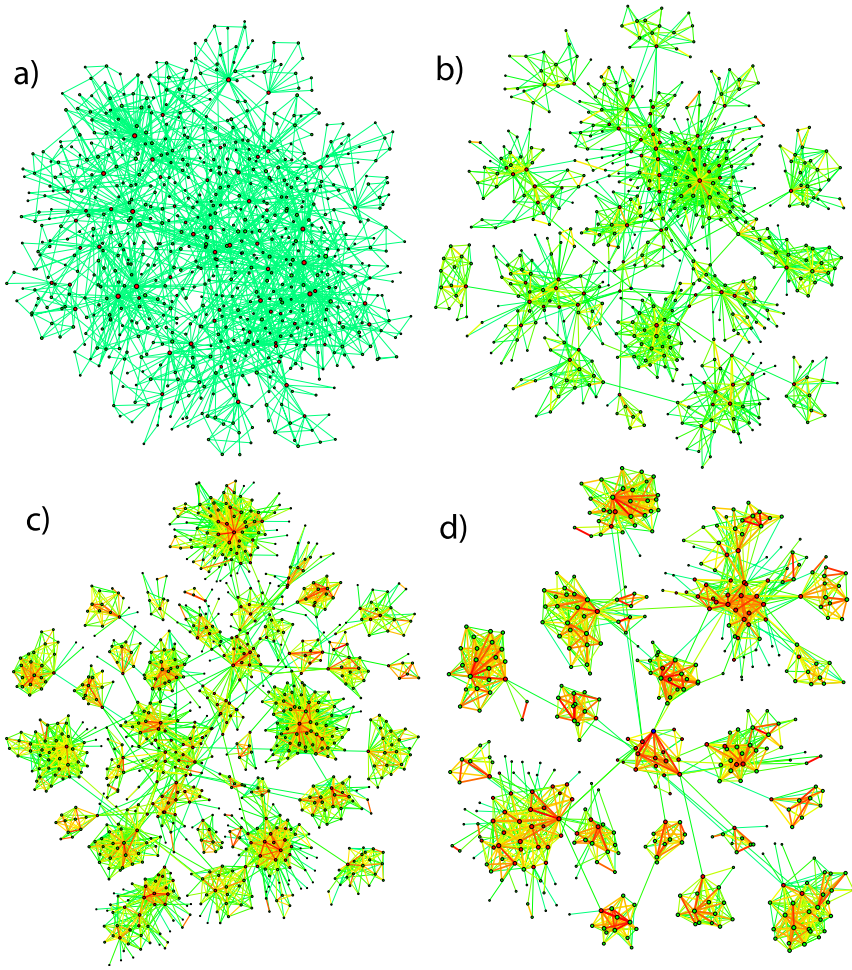


Figure 3.2: The effect of changing δ is seen in the snowball samples (136) of WSN networks with (a) $\delta = 0$, (b) $\delta = 0.1$, (c) $\delta = 0.5$, and (d) $\delta = 1$. Link colors change from green (weak links) to yellow and red (strong links). The average degree is 10 for each network. Figure published in Publication IV, copyright (2007) by the American Physical Society.

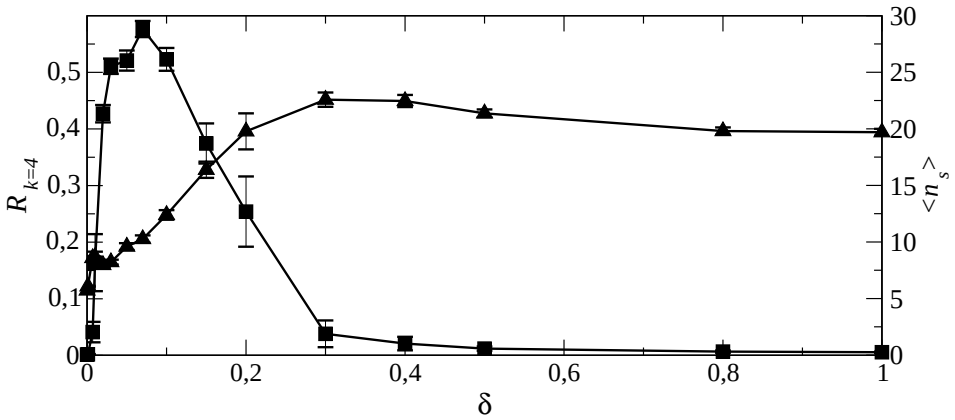


Figure 3.3: $R_{k=4}$ (□) and $\langle n_s \rangle$ (△) as a function of δ for WSN networks of size $N = 5 \times 10^4$. Results are averaged over 100 network realizations and error bars are measured standard deviations. Figure published in Publication IV, copyright (2007) by the American Physical Society.

pected because for $\delta = 0$ the WSN model resembles the unweighted models by Marsili (123) and Davidsen (121). For $\delta > 0$, these basic characteristics were found to remain essentially unaltered, which is a sign that they are robust against small variations in the implementation of the triadic and focal closure processes.

The most important observation in Publication IV was that sufficiently large values of δ lead to the formation of communities in the network. This is illustrated in Fig. 3.2. For large values of δ, the used paths rapidly gain more weight because the local search process favors strong links and reinforce them further. Thus, the local searches have a tendency to end up to familiar neighborhoods, which eventually leads to communities with strong internal links and weak links between them (as seen in Fig. 3.2 and confirmed by link percolation analysis). Figure 3.3 shows the relative size of the largest 4-clique community $R_{k=4}$ and the average size of communities excluding the largest one $\langle n_s \rangle$ as a function of δ. When $\delta = 0$, communities are small and the largest community contains approximately 50 nodes. Increasing δ makes the network first homogeneous in the sense that 4-cliques percolate through most of the network, but as δ becomes larger the nodes begin to condensate in tighter communities. When $\delta > 0.2$ the network contains communities whose average size is about 20 nodes while the largest community consist of several hundred nodes.

The significance of the WSN model is that it demonstrates that the weights are an essential part of the microscopic mechanisms responsible for the formation of communities in social networks. It seems reasonable to assume that similar processes might also be (at least partly) responsible for the formation of communities

in other types of networks.

3.3.3 Dynamics on WSN networks

In addition to demonstrating the importance of weights on community formation, the WSN model is suitable for producing realistic large scale weighted networks to be used in simulations of dynamic processes. Typical dynamical processes related to society include disease spreading (137), spreading of information or rumors (138), opinion formation, and strategic games like Prisoner's dilemma (139). Historically, these processes were first studied under the assumption of homogeneous mixing, which means that every agent in the system has equal probability to interact with any other agent (140). Often this approach allows for analytical treatment of the model through differential and rate equations, but it is clear that neglecting the networked structure of contacts is a severe limitation. Thus, it is not surprising that all of the above mentioned processes have recently been studied on networks (141; 28; 142; 143). As expected, the network structure has significant effect on the behavior of the models, e.g., the existence of hubs in networks decreases the epidemic threshold dramatically meaning that diseases spread and persevere much more easily than in the homogeneous mixing case. On the other hand, vaccinating the hubs provides an efficient way to prevent the outburst of contagious diseases (144; 29; 27). The effect of community structure has also been studied to some extent. For example, it is found that communities cause trapping of information (19) and that in the Prisoner's Dilemma game communities promote cooperation (145).

In Publication III the WSN model was used to study the effects of link weights and community structure on the dynamics of a weighted opinion formation model. The *social influence (si)* model used here is based on an unweighted model , the *AB model*, which concerns the competition of two extreme opinions A and B and also includes a third state AB between these (146). The AB state can be regarded as neutral or ambivalent, and state A can change to B (and vice versa) only through this intermediate state. Each node is in one of the three possible states and nodes change their state depending on the states of their nearest neighbors. In particular, the probability of a node to switch to state A is the larger the larger fraction of its neighbors in state A is (and similarly for state B). The si-model is a weighted modification of the AB model such that the interaction strengths are taken into account in the state switching probabilities. At the beginning of the simulation the states of the nodes are drawn randomly and the simulation proceeds by updating the states in a random order until a consensus is reached. The question of interest is under which conditions consensus is reached and what is the process of emergence and growth of spatial domains of nodes in the same state. This was investigated by numerical simulations for networks of size $N = 3000$ and $\delta = 0.5$, corresponding to fairly pronounced community structure. In addition to

the weighted social network models, we used two types of randomized reference networks: 1) fully randomized networks where the links of the model networks were randomly permuted while preserving the degree sequence, and 2) networks where the link weights of the model networks were randomly permuted while keeping the network topology unchanged. All results were averaged over a large number of network realizations. When comparing the simulation results for each network type we observed that both the distribution of weights within network topology and the community structure have a pronounced effect on the time it takes for the system to reach consensus. In particular, the dynamics turned out to be much slower in the weighted social network model than in the reference networks, which can be attributed to trapping of homogenous domains of opinions inside communities, so that global consus was difficult to reach.

These si-model simulations demonstrate that the correlations of link weights and community structure have significant effect on the dynamics taking place on the network. Thus, in the future there is a growing need for more realistic weighted network models, which can help us to understand various dynamic processes on networks. Possible applications are, for example, predicting and preventing the spread of infectious diseases, or manipulating the spread of political opinions, fashion etc.

Chapter 4

Summary of the results and discussion

Section 4.1 below summarizes the research of this thesis, especially the findings in the attached publications. Then, Sec. 4.2 discusses the future aspects of community detection and related complex network research.

4.1 Summary of the research

The limited resolution of the q-state Potts method was studied in Publication I. This work generalizes the results of Fortunato and Barthélemy, who showed earlier that modularity based community detection methods have a resolution limit such that small communities in a large network are invisible. The q-state Potts method, or RB-method, is a general framework for community detection and it contains the modularity optimization as a special case. In particular, the null model, which describes the random reference network to which the actual network is compared, is arbitrary in the RB-method. Furthermore, it contains a parameter γ which allows probing communities of different sizes. The communities are obtained by assigning nodes to communities such that the energy of the system is minimized. Our goal was to find out how the resolution of the RB-method depends on the null model and γ.

First, we studied the optimal network structure for the RB-method, i.e., the structure that gives the minimum energy. The result was that the optimal network consist of a ring of n^* fully connected communities, where $n^* \approx \sqrt{\gamma L}$ and L is the number of links in the network. This means that, given a ring-like network of n maximally intraconnected cliques ($n > n^*$), the obvious built-in communities are not resolved by optimizing the RB energy; instead, some communities appear merged, which is counter-intuitive and clearly an unwanted property of the method. We continued by considering the resolution threshold for a general null

model, where the main result was that communities of approximately the same size are not properly resolved if their sizes are less than approximately $\sqrt{Nl_{st}/\gamma}$, where l_{st} is the number of links connecting the communities. This order of magnitude estimate does not depend on the chosen null model. The main argument was that because the null model has to be properly normalized, i.e., the expected number of links is L, then the average connection probability between nodes has to decrease as N increases. This, in turn, decreases the number of expected links between any groups of nodes, and eventually triggers merging. We illustrated these results with several example cases. The results of Publication I indicated that when the community structure is not known beforehand, there is no simple way to decide which γ gives the most relevant communities and the "natural" choice $\gamma = 1$ is of no special importance. However, the ability to change γ provides a clear advantage to traditional modularity optimization, which is restricted to a single resolution.

The above mentioned problems of modularity optimization encouraged Arenas, Fernández and Gómes (AFG) to propose a new community detection method, which extends the modularity by introducing a self-link of weight r to each node. By varying r one is able to change the resolution of the method. This method has several similarities with the RB-method, and thus it was worth studying if the AFG-method suffers from the same limitations as the RB-method. We first published the results in a conference proceedings (not included in this thesis) and later in a shortened form in Publication II. There it was shown by theoretical derivations that the resolution of the AFG-method is limited similarly to the RB-method even though the methods are not identical. These results were augmented by numerical simulations, where the methods performed very similarly. In the last example case in Publication II we used a weighted version of the RB-method as introduced in Publication I (the AFG-method was originally formulated for weighted networks). No stable plateaus were seen when plotting the number of communities as a function of γ or r, which suggested that the finding the relevant communities by these methods may not be as straightforward as claimed by AFG in Ref. (85).

In Publication VI we continued the work related to the RB community detection method, this time by considering its application to weighted networks. Even though including the weights in RB method is not particularly difficult, one has to be careful with the null model. In fact, as mentioned above, we introduced a weighted RB-model already in Publication I, but the null model in Publication VI was better justified. We derived the weighted null model and theoretically analyzed the resolution properties of the weighted RB-method by using simple example networks. We found out that for sparse networks the behavior is qualitatively similar to the unweighted model whereas if the network is dense (or full), then the resolution of the method is independent of the network size N. In particular, the range of γ does not depend on N. We then applied the method to a full stock correlation matrix and found out that the method was able to obtain

meaningful communities. By monitoring the sizes of the two largest communities we identified a stable region, which corresponded well to the Forbes company classification. It seems that the RB-method is a reasonable choice for analyzing the community structure of weighted networks, in particular for dense networks where most methods have to discard the majority of links by thresholding them.

Publication V dealt with an another kind of community detection method, namely k-clique percolation. This method has several desirable properties, for example, it is able to deal with overlapping communities. However, the available implementations have been extremely slow for large networks. In Publication V we presented a new sequential clique percolation algorithm (SCP) which is efficient for small values of k allowing the analysis of networks of millions of nodes and links. The scaling of the computational time of the algorithm was investigated by applying it to three types of networks of varying sizes. It was found out that for each network type the computational time scales linearly as a function of the number of k-cliques in the network. This scaling was justified by simple theoretical arguments. The SCP method was found to be particularly well suited to the analysis of weighted networks as discussed in Publication V. In particular, it is able to extract the community structure at multiple weight thresholds in a single run, which is a significant improvement compared to earlier implementations. This feature can be utilized, e.g., in constructing a dendrogram visualization of the hierarchical structure of the communities, which we demonstrated by applying the method to a product category network obtained from a Finnish online auction site. The main limitation of the SCP method is its poor performance for large values of k for networks containing large cliques. However, currently in practical applications of the clique percolation method the best results have been obtained when k has been 5 or less. The SCP method is simple to modify such that it can be used for weighted clique percolation, as described in Publication V.

Publications III and IV consider how to model networks containing communities and how such structure affects the dynamics taking place on the networks. In Publication IV we presented a model for weighted social networks, which was inspired by the actual processes responsible for social tie formation as identified in the social sciences. Our goal was to keep the model as simple as possible while still capturing the most salient topological features and weight-topology correlations of real social networks. The model contains a parameter δ, which controls how strongly the local search processes reinforce visited links. When $\delta = 0$ the model is unweighted, whereas $\delta > 0$ generates weighted networks. We observed that the basic topological properties of the resulting networks corresponded well to those observed in real social networks for all values of δ, but community structure emerged only if δ was large enough, approximately $\delta > 0.1$. This effect was analyzed using the clique percolation method (where the algorithm of Publication V was used because existing methods were too slow) and the connection between link weights and topology was further investigated by link percolation.

The results showed that strong links are restricted to communities whereas links between communities are mainly weak, which is in agreement with empirical findings. The introduced model demonstrated that weights and the related microscopic rules are likely to play an important role in the formation of communities in social networks. It seems that the model provides an excellent way to generate artificial networks resembling real weighted social networks. Unfortunately, because of correlations between weights and topology, an analytical treatment of the model has proven to be difficult and despite of numerous attempts none of the basic characteristics can currently be derived analytically.

The model networks of Publication IV were used in Publication III to study the effect of link weights, community structure, and the correlations of these to the dynamics of a three state voter model. We first reviewed briefly the related results of the Voter model and AB model for unweighted networks containing communities and defined then a weighted extension of the AB model. A question of interest was the time it takes to reach consensus and how homogeneous spatial domains emerge and grow. Simulation of the *social influnence (si)* -model on WSN networks and on randomized reference networks showed that the layout of the weights as well as the community structure both increase the time to reach the consensus significantly. Essentially, nodes in each community were rapidly homogenized to the same opinion, but global consensus was difficult to reach because communities were weakly connected to each other.

4.2 Discussion

During the last decade numerous community detection methods have been proposed. However, it should be noted that so far the practical applications of community detection methods have been few, if one does not count the applications used for demonstrating the performance of each method. Often such applications are somewhat artificial because showing that a method detects the communities properly requires that one more or less knows the communities beforehand. For such networks there is not really a need for community detection. Currently, however, there are several efficient community detection methods and it can be expected (or hoped) that in the future the focus will change from developing new community detection methods to actually using the existing methods in the analysis of various networks.

The most common application of community detection methods is to find out which nodes comprise the communities. Consider, for example, proteins related to cancer metastasis (57). Even greater potential for interesting applications can be achieved by combining meta-data about the nodes with the knowledge of the communities. As an example, in the context of mobile phone call networks such meta-data could be the post code, gender, and age of the owner of each phone

number. The relation of such meta-data to network structure has recently been studied by Lambiotte *et al.* in Ref. (125), Lescovec and Horvitz in Ref. (109), and Blondel *et al.* in Ref. (147), of which the last one included communities in the analysis.

Another example of possible future use of community analysis – and the network approach in general – is augmenting the customer data analysis of companies. This is particularly straightforward for online networking services[1] where the networks naturally form from the interactions of the individuals. By analyzing the resulting networks it could be possible to target marketing, bonus features, new products, rumors, etc. to a small fraction of the most influential users, who would then rapidly spread these to their friends. When it comes to communities it is known that usually new ideas spread fast inside communities but slowly between them. Therefore, it might be advantageous to identify the most influential users in each community and target marketing to them. In addition, some communities are more central than others and one could concentrate only on the most important individuals in the most important communities. By grouping the users into communities it would also be possible to monitor how fast the users in each community adopt new products and how this depends of the community characteristics (average age, gender, etc). This information could provide new insight to the future product development as well as to marketing.

In addition to customer data analysis it would be possible to construct networks of products. A link connecting two products could reflect, for example, how many people own both products or how often they have been purchased at the same time. For obtaining such data the various bonus-card programs would be valuable. Analysis of these product networks could provide information of the buying patterns and interests of the users, which could be used in designing new products and targeting marketing. Undoubtedly some of these ideas are already in use in the industry, but it may take a long time until more advanced concepts will be adopted outside the scientific community.

Multiresolution methods and the problem of limited resolution receive a lot of attention in this thesis. Even though multiresolution methods are able to detect communities of various sizes and their resolution can be continuously tuned, the problem of choosing the ”best” community structure among the numerous alternatives remains. In fact, often there may not be a single proper community representation of a network. Communities can be considered as dense sets of nodes compared to their surroundings, but the problem is to define what exactly is the definition of ”density”, what is the threshold density that suffices to form a community, and how large the included surroundings are. In a way detecting communities in networks is analoguous to asking what are the peaks of a mountain chain. First, at large scales the highest peaks are easily found but when looking

[1] Consider, e.g., FaceBook or Habbo Hotel

closer one starts to see smaller peaks that stand out on the slopes of larger peaks. These smaller peaks are relevant to some extent but ultimately the smallest peaks appear on microscopic scales and this information is hardly useful[2]. This same feature is also present in community detection, namely, the definition of a relevant community structure depends of the specific questions asked.

The weighted social network model introduced in Publication IV has received some attention among network physicists and even among some sociologists. The simple rules of the model make it easy to understand and there are numerous ways one could extend and modify the model, e.g., by tweaking the local search and link reinforcement rules as well as link deletion mechanism. Currently, the nodes in the network are considered homogeneous in the sense that all nodes obey the same rules, but this does not necessarily have to be the case. An example would be a model where some nodes are more active in searching for new acquaintances than others. The effect of requiring reciprocity before a new link is established would also be worth studying. In this case a new link between two nodes would be established only if both nodes would choose each other in the local search process. This would make the life of unpopular nodes hard as gaining new links would be very difficult. Of course, such modifications are most useful only after the properties of the current model are properly understood, which is work in progress.

[2]Consider, e.g., how to name the ”significant” peaks of a mountain chain.

References

[1] **B. S. Kerner and H. Rehborn**, "Experimental features and characteristics of traffic jams", *Phys. Rev. E*, 53(2), R1297 (1996).

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

[3] **M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell**, "Observation of bose-einstein condensation in a dilute atomic vapor", *Science*, 269(5221), 198 (1995).

[4] **D. B. West**, *Introduction to graph theory*, Prentice Hall Upper Saddle River, NJ (1996).

[5] **P. Erdös and A. Rényi**, "On random graphs", *Publ. Math. Debrecen*, 6, 290 (1959).

[6] **J. Scott**, *Social Network Analysis: A Handbook*, Sage (2000).

[7] **N. Deo**, *Graph Theory with Applications to Engineering and Computer Science*, Prentice Hall, Upper Saddle River, NJ, USA (1974).

[8] **A. T. Balaban**, "Applications of graph theory in chemistry", *J. Chem. Inf. Comput. Sci.*, 25(3), 334 (1985).

[9] **N. A. Lynch**, *Distributed Algorithms*, Morgan Kaufmann, San Francisco (1996).

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

[11] **D. J. Watts and S. H. Strogatz**, "Collective dynamics of 'small-world' networks", *Nature*, 393, 440 (1998).

[12] **A. L. Barabási and R. Albert**, "Emergence of scaling in random networks", *Science*, 286(5439), 509 (1999).

[13] **M. E. J. Newman**, "The structure and function of complex networks", *SIAM Rev.*, 45(2), 167 (2003).

[14] **R. Albert and A. L. Barabási**, "Statistical mechanics of complex networks", *Rev. Mod. Phys.*, 74(1), 47 (2002).

[15] **G. Caldarelli**, *Scale-Free Networks: Complex Webs in Nature and Technology*, Oxford University Press, New York (2007).

[16] **M. E. J. Newman, A. L. Barabási, and D. J. Watts**, *The structure and dynamics of networks*, Princeton University Press (2006).

[17] **The Opte Project**, *http://opte.org*.

[18] **H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai**, *"Lethality and centrality in protein networks"*, Nature, *411(6833), 41 (2001).*

[19] **J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási**, *"Structure and tie strengths in mobile communication networks"*, Proc. Natl. Acad. Sci., *104(18), 7332 (2007).*

[20] **L. Costa, F. Rodrigues, G. Travieso, and P. Boas**, "*Characterization of complex networks: A survey of measurements"*, Arxiv preprint cond-mat/0505185 *(2005).*

[21] **S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin**, *"Giant strongly connected component of directed networks"*, Phys. Rev. E, *64(2), 025101 (2001).*

[22] **A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener**, *"Graph structure in the web"*, Computer Networks, *33(1-6), 309 (2000).*

[23] **M. E. J. Newman, S. H. Strogatz, and D. J. Watts**, *"Random graphs with arbitrary degree distributions and their applications"*, Phys. Rev. E, *64(2), 26118 (2001).*

[24] **W. Aiello, F. Chung, and L. Lu**, *"A random graph model for massive graphs"*, Proceedings of the thirty-second annual ACM symposium on Theory of computing, *171–180 (2000).*

[25] **M. Rosvall and C. T. Bergstrom**, *"Maps of random walks on complex networks reveal community structure"*, Proc. Natl. Acad. Sci., *105(4), 1118 (2008).*

[26] **D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts**, *"Network robustness and fragility: Percolation on random graphs"*, Phys. Rev. Lett., *85(25), 5468 (2000).*

[27] **R. Cohen, S. Havlin, and D. ben Avraham**, *"Efficient immunization strategies for computer networks and populations"*, Phys. Rev. Lett., *91(24), 247901 (2003).*

[28] **R. Pastor-Satorras and A. Vespignani**, *"Epidemic spreading in scale-free networks"*, Phys. Rev. Lett., *86(14), 3200 (2001).*

[29] **Z. Dezso and A. L. Barabási**, *"Halting viruses in scale-free networks"*, Phys. Rev. E., *65(5), 055103 (2002).*

[30] **J. Kleinberg**, *"Authoritative sources in a hyperlinked environment"*, Journal of the ACM, *46(5), 604 (1999).*

[31] **S. N. Dorogovtsev, J. F. Mendes, and A. N. Samukhin**, *"Structure of growing networks with preferential linking"*, Phys .Rev. Lett., *85(21), 4633 (2000).*

[32] **P. L. Krapivsky, S. Redner, and F. Leyvraz**, *"Connectivity of growing random networks"*, Phys .Rev. Lett., *85(21), 4629 (2000).*

[33] **G. Bianconi and A. L. Barabási**, *"Competition and multiscaling in evolving networks"*, Europhys. Lett., *54(4), 436 (2001).*

[34] **R. V. Solé, R. Pastor-Satorras, E. Smith, and T. B. Kepler**, *"A model of large-scale proteome evolution"*, Advances in Complex Systems, *5(1), 43 (2002).*

[35] **M. Mitzenmacher**, *"A brief history of generative models for power law and lognormal distributions"*, Internet Math., *1(2), 226 (2004).*

[36] **V. Pareto**, Cours d'Économie politique, *Rouge, Lausanne (1897).*

[37] **G. K. Zipf**, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, *Addison-Wesley Press, Cambridge (1949).*

[38] **G. U. Yule**, *"A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs"*, Philos. Transact. R. Soc. B, *213, 21 (1925).*

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

[40] **D. J. d. S. Price**, *"Networks of scientific papers"*, Science, *149, 510 (1965).*

[41] **R. K. Merton**, *"The matthew effect in science"*, Science, *159(3810), 56 (1968).*

[42] **A. Vázquez, R. Pastor-Satorras, and A. Vespignani**, *"Large-scale topological and dynamical properties of the internet"*, Phys. Rev. E, *65(6), 066130 (2002).*

[43] **D. J. Watts**, Six Degrees: The Science of a Connected Age, *WW Norton & Company (2003).*

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

[45] **E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A. L. Barabási**, *"Global organization of metabolic fluxes in the bacterium escherichia coli"*, Nature, *427(6977), 839 (2004).*

[46] **P. Csermely**, *"Strong links are important, but weak links stabilize them"*, Trends in Biochem. Sci., *29(7), 331 (2004).*

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

[48] **A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani**, *"The architecture of complex weighted networks"*, Proc. Natl. Acad. Sci., *101(11), 3747 (2004).*

[49] **J. Saramäki, M. Kivelä, J. P. Onnela, K. Kaski, and J. Kertész**, *"Generalizations of the clustering coefficient to weighted complex networks"*, Phys. Rev. E., *75(2), 027105 (2007).*

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

[51] **R. Toivonen, J.-P. Onnela, J. Saramäki, J. Hyvönen, and K. Kaski**, *"A model for social networks"*, Physica A, *371(2), 851 (2006).*

[52] **M. Girvan and M. E. J. Newman**, *"Community structure in social and biological networks"*, Proc. Natl. Acad. Sci., *99(12), 7821 (2002).*

[53] **D. Lusseau and M. E. J. Newman**, *"Identifying the role that animals play in their social networks"*, Proc. R. Soc. B, *271, 477 (2004).*

[54] **A. Arenas, L. Danon, A. Díaz-Guilera, P. M. Gleiser, and R. Guimerá**, *"Community analysis in social networks"*, Eur. Phys. J. B, *38(2), 373 (2004).*

[55] **G. Flake, S. Lawrence, C. Giles, and F. Coetzee**, *"Self-organization and identification of web communities"*, Computer, *35(3), 66 (2002).*

[56] **A. E. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor**, *"Compartments revealed in food-web structure"*, Nature, *426(6964), 282 (2003).*

[57] **P. F. Jonsson, T. Cavanna, D. Zicha, and P. A. Bates**, *"Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis"*, BMC Bioinformatics, *7(2) (2006).*

[58] **R. Guimerá and L. A. N. Amaral**, *"Functional cartography of complex metabolic networks"*, Nature, *433(7028), 895 (2005).*

[59] **S. Fortunato and C. Castellano**, *"Community structure in graphs"*, arXiv:0712.2716 *(2007).*

[60] **M. E. J. Newman and M. Girvan**, *"Finding and evaluating community structure in networks"*, Phys. Rev. E., *69(2), 026113 (2004).*

[61] **J. Reichardt and S. Bornholdt**, *"Statistical mechanics of community detection"*, Phys. Rev. E, *74(1), 016110 (2006).*

[62] **S. Fortunato, V. Latora, and M. Marchiori**, *"Method to find community structures based on information centrality"*, Phys. Rev. E., *70(5), 056104 (2004).*

[63] **L. Donetti and M. A. Muñoz**, *"Detecting network communities: a new systematic and efficient algorithm"*, J. Stat. Mech., *10, 10012 (2004).*

[64] **A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente**, *"Synchronization reveals topological scales in complex networks"*, Phys. Rev. Lett., *96(11), 114102 (2006).*

[65] **M. Rosvall and C. T. Bergstrom**, *"An information-theoretic framework for resolving community structure in complex networks"*, Proc. Natl. Acad. Sci., *104(18), 7327 (2007).*

[66] **G. Palla, I. Derényi, I. Farkas, and T. Vicsek**, *"Uncovering the overlapping community structure of complex networks in nature and society"*, Nature, *435(7043), 814 (2005).*

[67] **A. Lancichinetti, S. Fortunato, and J. Kertész**, *"Detecting the overlapping and hierarchical community structure of complex networks"*, arXiv:0802.1218 *(2008).*

[68] **M. E. J. Newman and E. A. Leicht**, *"Mixture models and exploratory analysis in networks"*, Proc. Natl. Acad. Sci., *104(23), 9564 (2007).*

[69] **M. Sales-Pardo, R. Guimerá, A. A. Moreira, and L. A. N. Amaral**, *"Extracting the hierarchical organization of complex systems"*, Proc. Natl. Acad. Sci., *104(39), 15224 (2007).*

[70] **A. Clauset, C. Moore, and M. Newman**, *"Structural inference of hierarchies in networks"*, Lect. Notes Comput. Sc., *4503, 1 (2007).*

[71] **A. Clauset, C. Moore, and M. E. J. Newman**, *"Hierarchical structure and the prediction of missing links in networks"*, Nature, *453(7191), 98 (2008).*

[72] **J. A. Hartigan**, Clustering Algorithms, *John Wiley & Sons, New York (1975).*

[73] **S. Fortunato and M. Barthelemy**, *"Resolution limit in community detection"*, Proc. Natl. Acad. Sci., *104(1), 36 (2007).*

[74] **A. Arenas, A. Fernández, S. Fortunato, and S. Gómez**, *"Motif-based communities in complex networks"*, J.Phys. A, *41(22), 224001 (2008).*

[75] **M. E. J. Newman**, *"Fast algorithm for detecting community structure in networks"*, Phys. Rev. E, *69(6), 066133 (2004).*

[76] **A. Clauset, M. E. J. Newman, and C. Moore**, *"Finding community structure in very large networks"*, Phys. Rev. E, *70(6), 66111 (2004).*

[77] **R. Guimerá, M. Sales-Pardo, and L. A. N. Amaral**, *"Modularity from fluctuations in random graphs and complex networks"*, Phys. Rev. E, *70(2), 025101 (2004).*

[78] **J. Duch and A. Arenas**, *"Community detection in complex networks using extremal optimization"*, Phys. Rev. E, *72(2), 027104 (2005).*

[79] **M. E. J. Newman**, *"Modularity and community structure in networks"*, Proc. Natl. Acad. Sci., *103(23), 8577 (2006).*

[80] **M. E. J. Newman**, *"Finding community structure in networks using the eigenvectors of matrices"*, Phys. Rev. E, *74, 036104 (2006).*

[81] **L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas**, *"Comparing community structure identification"*, J. Stat. Mech., *9 (2005).*

[82] **J. Reichardt and S. Bornholdt**, *"Detecting fuzzy community structures in complex networks with a potts model"*, Phys. Rev. Lett., *93(21), 218701 (2004).*

[83] **F. Y. Wu**, *"The potts model"*, Rev. Mod. Phys., *54(1), 235 (1982).*

[84] **J. Reichardt and S. Bornholdt**, *"Partitioning and modularity of graphs with arbitrary degree distribution"*, Phys. Rev. E, *76(1), 15102 (2007).*

[85] **A. Arenas, A. Fernández, and S. Gómez**, *"Analysis of the structure of complex networks at different resolution levels"*, New J. Phys., *10(5), 053039 (2008).*

[86] **J. P. Onnela, K. Kaski, and J. Kertész**, *"Clustering and information in correlation based financial networks"*, Eur. Phys. J. B, *38(2), 353 (2004).*

[87] **J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész**, *"Limited resolution in complex network community detection with potts model approach"*, Eur. Phys. J. B, *56, 41 (2007).*

[88] **M. E. J. Newman**, *"Analysis of weighted networks"*, Phys. Rev. E, *70(5), 56131 (2004).*

[89] **I. Farkas, D. Ábel, G. Palla, and T. Vicsek**, *"Weighted network modules"*, New Journal of Physics, *9(6), 180 (2007).*

[90] **Forbes**, *http://www.forbes.com*, referenced in march-april 2002.

[91] **R. N. Mantegna**, *"Hierarchical structure in financial markets"*, Eur. Phys. J. B, *11(1), 193 (1999).*

[92] **G. Bonanno, N. Vandewalle, and R. N. Mantegna**, *"Taxonomy of stock market indices"*, Physical Review E, *62(6), 7615 (2000).*

[93] **T. Heimo, G. Tibély, J. Saramäki, K. Kaski, and J. Kertész**, *"Spectral methods and cluster structure in correlation-based networks"*, Physica A, *(in press), arXiv:0708.1862v1 (2008).*

[94] **G. Palla, A.-L. Barabási, and T. Vicsek**, *"Quantifying social group evolution"*, Nature, *446(713), 664 (2007).*

[95] **B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek**, *"Cfinder: locating cliques and overlapping modules in biological networks"*, Bioinformatics, *22(8), 1021 (2006).*

[96] **M. E. J. Newman**, *"The structure of scientific collaboration networks"*, Proc. Natl. Acad. Sci., *98(2), 404 (2001).*

[97] **S. Lehmann, M. Schwartz, and L. K. Hansen**, *"Bi-clique communities"*, Phys. Rev. E, *78, 016108 (2008).*

[98] **H. C. White, S. A. Boorman, and R. L. Breiger**, *"Social structure from multiple networks. i. blockmodels of roles and positions"*, Am. J. Sociology, *81(4), 730 (1976).*

[99] **S. Wasserman and K. Faust**, Social Network Analysis: Methods and Applications, *Cambridge University Press (1994).*

[100] **P. Bearman and P. Parigi**, *"Cloning headless frogs and other important matters: Conversation topics and network structure"*, Soc. Forces, *83, 535 (2004).*

[101] **M. C. González, H. J. Herrmann, J. Kertész, and T. Vicsek**, *"Community structure and ethnic preferences in school friendship networks"*, Physica A, *379(1), 307 (2007).*

[102] **G. Kossinets and D. J. Watts**, *"Empirical analysis of an evolving social network"*, Science, *311(5757), 88 (2006).*

[103] **J. Wang and P. D. Wilde**, *"Properties of evolving e-mail networks"*, Phys. Rev. E, *70(6), 066121 (2004).*

[104] **G. Caldarelli, F. Coccetti, and P. De Los Rios**, *"Preferential exchange: Strengthening connections in complex networks"*, Phys. Rev. E, *70(2), 27102 (2004).*

[105] **W. Aiello, F. Chung, and L. LU**, Random evolution of massive graphs, *Handbook of Massive Data Sets, Kluwer Academic (2002).*

[106] **M. E. J. Newman**, *"Scientific collaboration networks: Network construction and fundamental results"*, Phys. Rev. E, *64(1), 016131 (2001).*

[107] **M. E. J. Newman**, *"Scientific collaboration networks: Shortest paths, weighted networks, and centrality"*, Phys. Rev. E., *64(1), 016132 (2001).*

[108] **L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley**, *"Classes of small-world networks"*, Proc. Natl. Acad. Sci. *(2000).*

[109] **J. Leskovec and E. Horvitz**, *"Planetary-scale views on an instant-messaging network", in "WWW '08: Proceeding of the 17th international conference on World Wide Web", 915, ACM, New York, NY, USA (2008).*

[110] **F. Karinthy**, Chain-links, *Everything is different, Budabest (1929).*

[111] **I. S. Pool and M. Kochen**, *"Contacts and influence"*, Social Networks, *1(1), 5 (1978).*

[112] **S. Milgram**, *"The small world problem"*, Psychology Today, *2(1), 60 (1967).*

[113] **Six Degrees Magazine**, *http://sixdegreesmag.com.*

[114] **SixDegrees**, *http://www.6d.fi.*

[115] **W. Stanzler, E. Laneuville, and D. Petrarca**, *"Six degrees", (2006).*

[116] **F. Schepisi**, *"Six degrees of separation", (1993).*

[117] **K. Bunyan**, *Six degrees facebook application.*

[118] **University of Virginia**, *"The oracle of bacon at virginia", http://oracleofbacon.org.*

[119] **M. S. Granovetter**, *"The strength of weak ties"*, Am. J. Sociology, *78(6), 1360 (1973).*

[120] **J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Menezes, K. Kaski, A. L. Barabási, and J. Kertész**, *"Analysis of a large-scale weighted network of one-to-one human communication"*, New J. Phys., *9(6), 179 (2007).*

[121] **J. Davidsen, H. Ebel, and S. Bornholdt**, *"Emergence of a small world from local interactions: Modeling acquaintance networks"*, Phys. Rev. Lett., *88(12), 128701 (2002).*

[122] **E. M. Jin, M. Girvan, and M. E. J. Newman**, *"Structure of growing social networks"*, Phys. Rev. E, *64(4), 046132 (2001).*

[123] **M. Marsili, F. Vega-Redondo, and F. Slanina**, *"The rise and fall of a networked society: A formal model"*, Proc. Natl. Acad. Sci., *101(6), 1439 (2004).*

[124] **J. M. Kumpula, J. P. Onnela, J. Saramäki, K. Kaski, and J. Kertész**, *"Emergence of communities in weighted networks"*, Phys. Rev. Lett., *99, 228701 (2007).*

[125] **R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. V. Dooren**, *"Geographical dispersal of mobile communication networks"*, Physica A, *387(21), 5317 (2008).*

[126] **M. C. González, P. G. Lind, and H. J. Herrmann**, *"System of mobile agents to model social networks"*, Phys.Rev.Lett., *96(8), 88702 (2006).*

[127] **M. C. González, P. G. Lind, and H. J. Herrmann**, *"Model of mobile agents for sexual interactions networks"*, Eur. Phys. J. B, *49(3), 371 (2006).*

[128] **A. Grönlund and P. Holme**, *"Networking the seceder model: Group formation in social and economic systems"*, Phys. Rev. E, *70(3), 036108 (2004).*

[129] **G. Robins, P. Pattison, Y. Kalish, and D. Lusher**, *"An introduction to exponential random graph (p*) models for social networks"*, Social Networks, *29(2), 173 (2007).*

[130] **M. Boguna, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas**, *"Models of social networks based on social distance attachment"*, Phys. Rev. E, *70(5), 056122 (2004).*

[131] **L. H. Wong, P. Pattison, and G. Robins**, *"A spatial model for social networks"*, Physica A, *360(1), 99 (2006).*

[132] **M. McPherson, L. Smith-Lovin, and J. M. Cook**, *"Birds of a feather: Homophily in social networks"*, Annu. Rev. Sociol., *27(1), 415 (2001).*

[133] **A. Vázquez**, *"Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations"*, Phys. Rev. E, *67(5), 056104 (2003).*

[134] **M. Hammer**, *"Predictability of social connections over time"*, Social Networks, *2, 165 (1980).*

[135] **R. Toivonen, L. Kovanen, M. Kivelä, J. P. Onnela, J. Saramäki, and K. Kaski**, *"A comparative study of stochastic algorithmic models for social networks"*, arXiv:0805.0512 *(2008).*

[136] **S. H. Lee, P.-J. Kim, and H. Jeong**, *"Statistical properties of sampled networks"*, Phys. Rev. E, *73(1), 016102 (2006).*

[137] **H. W. Hethcote**, *"The mathematics of infectious diseases"*, SIAM Review, *42(4), 599 (2000).*

[138] **E. M. Rogers**, Diffusion of Innovations, *Simon and Schuster (2003).*

[139] **A. Rapoport and A. M. Chammah**, Prisoner's Dilemma, *University of Michigan Press (1965).*

[140] **N. T. J. Bailey**, The Mathematical Theory of Infectious Diseases and its Applications, *London (1975).*

[141] **M. Kuperman and G. Abramson**, *"Small world effect in an epidemiological model"*, Phys. Rev. Lett., *86(13), 2909 (2001).*

[142] **D. J. Watts**, *"A simple model of global cascades on random networks"*, Proc. Natl. Acad. Sci., *99(9), 5766 (2002).*

[143] **Y. Moreno, M. Nekovee, and A. F. Pacheco**, *"Dynamics of rumor spreading in complex networks"*, Phys. Rev. E., *69(6), 066130 (2004).*

[144] **R. Pastor-Satorras and A. Vespignani**, *"Immunization of complex networks"*, Phys. Rev. E., *65(3), 036104 (2002).*

[145] **S. Lozano, A. Arenas, and A. Sánchez**, *"Mesoscopic structure conditions the emergence of cooperation on social networks"*, PLoS ONE, *3(4) (2008).*

[146] **X. Castelló, V. M. Eguíluz, and M. S. Miguel**, *"Ordering dynamics with two non-excluding options: bilingualism in language competition"*, New J. Phys., *8(12), 308 (2006).*

[147] **V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre**, *"Fast unfolding of communities in large networks"*, J. Stat. Mech., *P10008 (2008).*