Department of Biomedical Engineering and Computational Science

# Computational analysis of large and time-dependent social networks

Lauri Kovanen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
**DOCTORAL DISSERTATIONS** 81/2013

# Computational analysis of large and time-dependent social networks

**Lauri Kovanen**

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall F239a of the school on 16 May 2013 at 12.

**Aalto University**
**School of Science**
**Department of Biomedical Engineering and Computational Science**

**Supervising professor**
Jari Saramäki

**Thesis advisor**
Jari Saramäki

**Preliminary examiners**
Esteban Moro
Universidad Carlos III de Madrid
Spain

János Török
Budapest University of Technology and Economics
Hungary

**Opponents**
Renaud Lambiotte
University of Namur
Belgium

Aalto University publication series
**DOCTORAL DISSERTATIONS** 81/2013



ISBN 978-952-60-5165-9 (printed)
ISBN 978-952-60-5164-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5164-2

Unigrafia Oy
Helsinki 2013

Finland

441 697
Printed matter

Aalto University

# Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

**Author**
Lauri Kovanen

**Name of the doctoral dissertation**
Computational analysis of large and time-dependent social networks

**Publisher** School of Science

**Unit** Department of Biomedical Engineering and Computational Science

**Series** Aalto University publication series DOCTORAL DISSERTATIONS 81/2013

**Field of research**

**Manuscript submitted** 12 February 2013 **Date of the defence** 16 May 2013

**Permission to publish granted (date)** 27 March 2013 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

**Abstract**

Complex systems consist of a large number of elements that interact in a non-trivial way; for example the human brain, society, Internet, and biological organisms can all be modelled as complex systems. Complex systems can be naturally represented as networks, mathematical objects that consist of nodes and edges connecting these nodes, and the study of large networks based on empirical data has become known as complex networks. Since the first articles on complex networks appeared in the end of the 1990's, various technological, biological, and social networks have been analyzed. In recent years introductory text books on the subject have also been published.

The study of social networks of course has a longer history. Small social networks have been studied for decades in sociology, social psychology and anthropology, and the influence that social networks have on both performance and well being of individuals has been well documented. The availability of electronic communication records—mobile phone calls, emails, online social networking sites and even multiplayer computer games—have changed the scale and detail at which social networks can be analyzed. The largest data set studied so far includes over 700 million individuals, and the mobile phone call records studied in this Thesis contain information of over 6 million people. The combination of powerful computers and large data sets have enabled the emergence of computational social science.

Several aspects of large social networks are studied in this Thesis. Models of social networks are commonly used as a way to gain insight about the structure of these networks. The first article studies a number of models suggested for social networks and discusses their advantages and shortcomings. The community structure of various networks has also been a subject of great interest. It is widely accepted that nearly all networks have modular structure, evidenced by local densifications of connectivity. However, identifying communities in empirical data has turned out to be difficult both theoretically and in practice. We apply three state-of-art community detections methods to a large social network and evaluate the quality of the identified communities.

One important aspect of human interactions is omitted when analyzing networks: time. Temporal networks have become a common framework for studying data sets where the relations between nodes vary with time, and this framework can be readily applied to study mobile phone calls. The last part of this Thesis introduces the concept of temporal motifs—recurring patterns of events in temporal networks—that can be used to analyze the meso-scale structure of temporal networks.

**Keywords** complex systems, complex networks, social networks, temporal networks

**ISBN (printed)** 978-952-60-5165-9 **ISBN (pdf)** 978-952-60-5164-2

**ISSN-L** 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

**Location of publisher** Espoo **Location of printing** Helsinki **Year** 2013

**Pages** 180 **urn** http://urn.fi/URN:ISBN:978-952-60-5164-2

A! Aalto-yliopisto

**Tiivistelmä**

**Aalto-yliopisto, PL 11000, 00076 Aalto** www.aalto.fi

**Tekijä**
Lauri Kovanen

**Väitöskirjan nimi**
Suurten ja aikariippuvien sosiaalisten verkostojen laskennallinen analyysi

**Julkaisija** Perustieteiden korkeakoulu

**Yksikkö** Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

**Sarja** Aalto University publication series DOCTORAL DISSERTATIONS 81/2013

**Tutkimusala**

**Käsikirjoituksen pvm** 12.02.2013 **Väitöspäivä** 16.05.2013

**Julkaisuluvan myöntämispäivä** 27.03.2013 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

**Tiivistelmä**

Kompleksiset järjestelmät koostuvat suuresta määrästä alkioita, joiden keskinäiset vuorovaikutukset ovat monimutkaisia; esimerkiksi ihmisaivoja, yhteiskuntaa, internettiä ja biologisia organismeja voidaan kaikkia mallintaa kompleksisina järjestelminä. Matemaattisesti näitä järjestelmiä on luontevaa käsitellä verkostoina, jotka koostuvat solmuista ja niitä yhdistävistä kaarista. Suuria empiiriseen dataan perustuvia verkostoja tutkiva tieteenala tunnetaankin nykyään nimellä "kompleksiset verkostot". Ensimmäiset alaan liittyvät tutkimukset julkaistiin 90-luvun lopulla, ja sen jälkeen lähestymistapaa on käytetty erilaisten teknologisten, biologisten sekä sosiaalisten verkostojen analysointiin. Myös kompleksisia verkostoja käsitteleviä oppikirjoja on julkaistu.

Sosiaalisten verkostojen historia on kuitenkin tätä pidempi, sillä niitä on tutkittu jo vuosikymmeniä sosiologiassa, sosiaalipsykologiassa sekä antropologiassa. Sosiaalisen verkoston vaikutus yksilön toimintakykyyn ja hyvinvointiin on nykyään hyvin tunnettu. Elektroninen tiedonvälitys—matkapuhelut, sähköposti, internetin sosiaalisen median palvelut ja jopa monen pelaajan tietokonepelit—on kuitenkin muuttanut merkittävästi tutkittavissa olevien sosiaalisten verkostojen kokoa ja laatua. Suurin tähän asti analysoitu sosiaalinen verkosto sisältää tietoja yli 700 miljoonasta ihmisestä, ja tässä väitöskirjassa analysoitu matkapuhelindata käsittää yli 6 miljoonan ihmistä. Tehokkaat tietokoneet yhdessä valtavien tietokantojen kanssa ovat luomassa uutta, laskennallista sosiaalitiedettä.

Tässä väitöskirjassa tutkitaan useita suurten sosiaalisten verkostojen ominaisuuksia. Matemaattisia malleja käytetään usein apuna kun halutaan selittää sosiaalisten verkostojen rakennetta. Väitöskirjan ensimmäinen artikkeli tutkii erilaisten kirjallisuudessa esitettyjen mallien ominaisuuksia. Toinen paljon tutkittu aihe on verkostojen modulaarinen rakenne. On yleisesti tunnettua, että lähes kaikki empiiriset verkostot ovat modulaarisia: ne sisältävät paikallisia kaaritihentymiä. Moduulien tunnistaminen on kuitenkin osoittautunut huomattavan vaikeaksi ongelmaksi sekä teoreettisesti että käytännössä. Tunnistusmenetelmien toimintaa voidaan tutkia arvioimalla niiden empiirisestä verkostosta löytämiä moduuleja.

Aika on keskeinen elementti kaikessa vuorovaikutuksessa, mutta jää huomioimatta jos sosiaalista vuorovaikutusta tutkitaan vain verkostojen avulla. Temporaalisten verkostojen avulla voidaan käsitellä verkostomaista dataa, jossa vuorovaikutukset riippuvat ajasta, kuten matkapuheluiden muodostamaa rakennetta. Väitöskirjan viimeinen osa esittelee menetelmän, jonka avulla voidaan analysoida toistuvia rakenteita temporaalisissa verkostoissa.

**Avainsanat** kompleksiset systeemit, verkostoteoria, sosiaaliset verkostot, aikariippuvat verkostot

**ISBN (painettu)** 978-952-60-5165-9 **ISBN (pdf)** 978-952-60-5164-2

**ISSN-L** 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

**Julkaisupaikka** Espoo **Painopaikka** Helsinki **Vuosi** 2013

**Sivumäärä** 180 **urn** http://urn.fi/URN:ISBN:978-952-60-5164-2

# Preface

I began working on social networks in the summer of 2007 when I joined the Laboratory of Computational Engineering, the predecessor of BECS, as a summer student. Those first steps eventually lead to a Master's thesis, and as my knowledge of complex systems increased, so did my interest in studying them further. The doctoral studies I started four years ago are now coming to an end.

I want to thank my supervisor, professor Jari Saramäki for all his support and advice during the journey. I have had an exceptional freedom of carrying out research from the birth of new ideas to the final steps of publishing. There can hardly be a better way to learn what science is. Doing a doctoral thesis is not always easy, but that you told me before I started.

I also wish to express my gratitude to professor Kimmo Kaski for creating such an excellent research environment. BECS has the most favorable circumstances for scientific activities, and I have greatly enjoyed my time here. The bureaucracy has been kept to the minimum, and funding has not been an issue. I know that this is not something that simply happens. I am also grateful for all the opportunities to attend conferences and visit other universities, and all the discussions covering everything between life, science and politics.

Even though this Thesis has only one author, many people have contributed to the research that lies behind. I greatly appreciate the comments and instructions I have received from professor János Kertész, the informal discussions with professor Jukka-Pekka Onnela, the happiness and diligence of Riitta Toivonen, and the accurate and timely results from Gergely Tibély.

I am also grateful to my fellow doctoral students Mikko Kivelä and Gerardo Iñiguez for your company during these years, and all the practical and non-practical discussions we have had. And a great thank you to

everyone in the research group during these years. You have made it enjoyable to come to work every day: Talayeh Aledavood, Arnab Chatterjee, Richard Darst, Pietro Della Briotta Parolo, prof. Santo Fortunato, Tapio Heimo, Darko Hric, Jörkki Hyvönen, Hang-Hyun Jo, Markus Karppinen, Márton Karsai, Rainer Kujala, Jussi Kumpula, Marija Mitrović, Vasyl Palchykov, Raj Kumar Pan, Juan Perotti, and Taha Yasseri. And all the other people in BECS who have made it so much better to work here: Enrico Glerean, Julio Hernández Pavón, Riku Linna, Hanna Mäki, Tiina Näsi, Margareta Segerståhl, Jarno Vanhatalo, Panu Vesanen, Mikko Viinikainen, and everyone else I have talked with during all these years. I am supposed to call you colleagues, but you have become my friends.

I also wish to thank the administration for their kind help with the bureaucracy: Katri Kaunismaa, Katja Korpinummi, Eeva Lampinen, Laura Pyysalo, and Marita Stenman, and also the people who kept the computational resources running: Jukka Merinen, Mikko Hakala, Timo Aarnio, Jarkko Salmi, and Jari Siven. This work would not have been possible without your effort.

The people at the CABDyN Complexity Centre in Oxford were generous with their hospitality, and I much enjoyed my stay: Felix Reed-Tsochas, Elizabeth Leicht, Eduardo López, Jianguo Liu, and Griffith Rees. I also acknowledge the support of the Doctoral program Brain&Mind. Neuroscience is not my main topic, but it surely was an interesting digression. And for widening my scope outside techonology, I am grateful to everyone I have had the pleasure to meet in Aallonhuiput, Aalto Doctoral Student Association.

Finally, I wish to thank my dear Johanna for her support during these years. I did not always come home on time.

Espoo, April 19, 2013,

Lauri Kovanen

# Contents

**Preface** 1

**Contents** 3

**List of Publications** 5

**Author's Contribution** 7

**1. Introduction** 9

1.1 Complex systems . . . . . 9

1.2 Scope of this Thesis . . . . . 11

**2. Networks** 13

2.1 Graph theory . . . . . 14

2.2 Random graphs . . . . . 17

2.3 Complex networks . . . . . 21

**3. Social networks** 23

3.1 Social interactions . . . . . 24

3.2 Common properties of social networks . . . . . 25

3.2.1 Fat-tailed degree distribution . . . . . 26

3.2.2 Assortativity . . . . . 28

3.2.3 Clustering . . . . . 28

3.2.4 Contexts and communities . . . . . 28

3.2.5 Reciprocity . . . . . 29

3.3 Social network analysis . . . . . 30

3.3.1 Networks, performance and well-being . . . . . 30

3.3.2 Social networks and information technology . . . . . 31

3.4 Universal models . . . . . 34

**4. Communities in social networks** 37

4.1 Partitions . . . . . 39
4.1.1 Problems with modularity . . . . . 39
4.1.2 The Louvain method . . . . . 40
4.1.3 Infomap . . . . . 41
4.2 Overlapping communities . . . . . 42
4.2.1 Detecting overlapping communities . . . . . 44
4.3 Community detection in practice . . . . . 45

**5. Temporal networks** **47**
5.1 Motifs in static networks . . . . . 50
5.2 Temporal motifs . . . . . 50
5.2.1 Temporal subgraphs . . . . . 50
5.2.2 Isomorphism of temporal subgraphs . . . . . 52
5.2.3 Implementation . . . . . 53
5.3 Analyzing motif counts . . . . . 54
5.3.1 References . . . . . 54
5.3.2 Null models . . . . . 55
5.3.3 Reductive null models . . . . . 56

**6. Results** **59**
6.1 Models of social networks . . . . . 59
6.2 Structure of social networks . . . . . 60
6.3 Temporal motifs . . . . . 61

**7. Discussion** **63**

**Bibliography** **65**

**Publications** **77**

# 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** Riitta Toivonen, Lauri Kovanen, Mikko Kivelä, Jukka-Pekka Onnela, Jari Saramäki, Kimmo Kaski. A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks*, Volume 31, issue 4, pages 240-254, October 2009.

**II** Lauri Kovanen, Jari Saramäki, Kimmo Kaski. Reciprocity of mobile phone calls. *Dynamics of Socio-Economic Systems*, Volume 2, issue 2, pages 138-151, March 2011.

**III** Gergely Tibély, Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, Jari Saramäki. Communities and beyond: Mesoscopic analysis of a large social network with complementary methods. *Physical Review E*, Volume 83, issue 5, 056125, May 2011.

**IV** Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, Jari Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, Volume 2011, P11005, November 2011.

**V** Lauri Kovanen, Kimmo Kaski, János Kertész, Jari Saramäki. Temporal motifs reveal homophily, gender-related patterns and group talk in mobile communication networks. *pre-print*, arXiv:1302.2563, February 2013.

# Author's Contribution

**Publication I: "A comparative study of social network models: Network evolution models and nodal attribute models"**

Implemented the network models and the fitting of their parameters.

**Publication II: "Reciprocity of mobile phone calls"**

Design and implementation of the research. Primary writer.

**Publication III: "Communities and beyond: Mesoscopic analysis of a large social network with complementary methods"**

Contributed to designing the study and calculating results. Major role in writing the article.

**Publication IV: "Temporal motifs in time-dependent networks"**

Original concept of research, design and implementation of algorithms. Primary writer.

**Publication V: "Temporal motifs reveal homophily, gender-related patterns and group talk in mobile communication networks"**

Original concept of research, design and implementation of analysis. Major role in writing the article.

# 1. Introduction

## 1.1 Complex systems

Current interest in complex systems was sparked by two research articles published in the turn of the millennium. In 1998 Duncan Watts and Steven Strogatz observed that many systems—biological, social, and technological—had features that had so far been observed only in either random or regular systems, and therefore appeared to lie somewhere between these two extremes [1]. A year later Albert-László Barabási and Réka Albert showed that the distribution of node degrees in networks—the number of connections each element has—is very similar in a diverse set of empirical systems and provided a simple model to explain their observation [2]. These two articles, together with a sudden availability of data from large empirical systems, contributed to the birth of a new field that is today known as *complex networks*.

The idea of studying complex systems is not without predecessors. In fact, the classification of problems by their inherent complexity, put forward by Warren Weaver in 1948, is still relevant today [3]. According to Weaver, until the 20th century science was mostly concerned with problems of *simplicity*. Such problems contain relatively few variables and can typically be solved exactly, like calculating the paths of colliding billiard balls. In the 20th century statistical physics began addressing problems of *disorganized complexity* by calculating statistical properties of large systems with high accuracy. In some cases the same approach can also be applied to social systems: even though it is nearly impossible to predict when any single person will die, the average death rate can be predicted with high accuracy. In both cases we can make accurate claims about the system as a whole even though we have little knowledge of its con-

stituents.

The third kind of problems described by Weaver, and the ones primarily addressed by this Thesis, are the problems of *organized complexity*. These problems also typically involve a large number of entities, but now the interactions are *organized*; we are "dealing simultaneously with a sizable number of factors which are interrelated into an organic whole." Many important problems fall into this category, for example the surprising emergence of economic crises as a result of mostly uncoordinated but organized behavior of individuals. Biological systems also give rise to several problems of organized complexity. Determining the cause of cancer is notoriously difficult precisely because it involves a large number of non-trivial interactions.

Just like economic crises, other social phenomena also take us by surprise even though we ourselves make up the social system. A classic example is the *small-world phenomenon*, also known as "six degrees of separation", famously illustrated by a series of experiments carried out by Jeffrey Travers and Stanley Milgram in the late 1960's [4]. In one experiment 296 individuals in the US were asked to deliver a letter to a target person in Massachusetts by passing it on to a personal acquaintance they thought to be closer to the target. For the 64 letters that eventually reached the target, the average number of intermediaries was only $5.2$. Even though this experiment is far from proving the exact number of intermediaries, the fact that we live in a small world has since been verified over and over again [5].

It is hard to draw exact lines between the three different types of problems identified by Weaver. There is no theoretical difference between problems of simplicity and those of disorganized complexity: a system consisting of $10^{23}$ gas molecules follows the same laws of physics as a system of only two molecules, but it is impossible to study the larger system by tracing the paths taken by all molecules. In similar vein, the difference between disorganized and organized complexity is practical rather than theoretical. When the interactions are numerous and organized, the system cannot be treated with methods developed for problems of simplicity or disorganized complexity. It was also organized complexity that Herbert Simon studied in his seminal article on complexity [6], defining complex system as

> . . . one made up of a large number of parts that interact in a nonsimple way. In

> such systems, the whole is more than the sum of the parts, not in an ultimate, metaphysical sense, but in the important pragmatic sense that, given the properties of the parts and the laws of their interactions, it is not a trivial matter to infer the properties of the whole.

It might seem that by defining complex systems as those with nonsimple interactions we at the same time define them to be outside the scope of what can be studied. In addition, many researchers studying complex systems have background in statistical physics, and are effectively using methodology originally developed for disorganized complexity to study problems of organized complexity. How can we possibly answer problems of organized complexity? The fact is, we cannot—not in the same way as science answers problems of simplicity and disorganized complexity. Because of the intrinsic, pervasive complexity, we might never be able to model and predict complex systems with accuracy comparable to that required to send a man to the Moon—a task that is mostly a combination of problems of simplicity and disorganized complexity.

Even if exact predictions seem unlikely at the moment, there is still a lot to be gained. In fact, there exists already a complex systems measure that is being used daily by hundreds of millions of people. In the complex system terminology this measure falls into the category of eigenvector centrality measures and is best known as *PageRank* [7]. Most of its users, however, are probably more familiar with the name of the company founded by its inventors Larry Page and Sergei Brin, and the search engine carrying the same name: Google.

## 1.2 Scope of this Thesis

This Thesis studies the statistical properties of social networks by analyzing electronic communication records such as email messages, online social networks, and mobile phone calls.

Structural properties of small networks are analyzed by studying the performance of models that have been suggested to reproduce the structure of social networks. Community detection in social networks is studied by analyzing the communities detected in a mobile phone communication network by three state-of-the-art community detection methods.

The last section introduces the concept of temporal motifs that can be used to study recurring patterns of events in time-dependent networks.

In addition to introducing the necessary concepts and algorithms, I also discuss how the resulting motif counts can be analyzed.

# 2. Networks

Because complex systems by definition consist of interactions between a large number of elements, it is natural to represent them as *networks*. A network, or *graph* as they are also known, is a mathematical object that consists of elements and their pairwise relations.

The origin of graphs is commonly traced back to the great 18th century mathematician Leonhard Euler who, the story goes, was interested in the network of bridges in Königsberg and in the process gave birth to the field of mathematics we today know as *graph theory*.[1] For the modern audience the best known example of a network is no doubt found online.[2] With over 10 % of the world population in Facebook [9], social networking sites have taken a significant position in our lives. With unprecedented ease they also allow us to visualize the social structure we are part of, as illustrated by Figure 2.1 that shows the Facebook contacts of the author. Indeed, the term "graph" refers to this graphical nature of networks that is helpful in explaining the basic concepts of graph theory in the next Section. Unfortunately, large empirical networks are typically less easy to visualize, and this difficulty is one reason why more advanced methods are needed to analyze them.

[1] According to the story, a popular past-time for the citizens of Königsberg, present-day Kaliningrad, was to stroll about the town and try to find a route that would allow them to cross each bridge in the city exactly once. Euler recognized that all relevant aspects of the problem—the number of bridges between different land masses—can be captured in a graph, and proved that no such route can exist.

[2] Internet—the physical network of computers, servers and routers on which the online world exists—is obviously a network in itself, although most users are unaware of its structure.

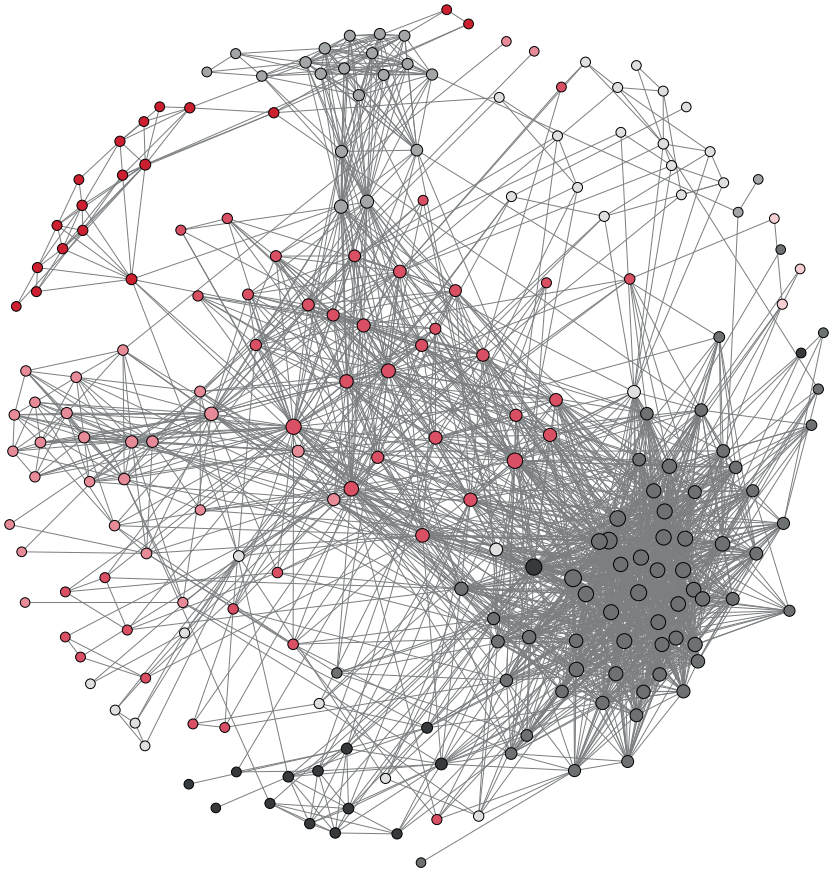


**Figure 2.1.** The *egonet* formed by my own Facebook contacts as recorded on January 10 2012. Each node—denoted by circles—is a person I have marked as a friend in Facebook, and there is an edge between two people if they have marked each other as friends. Node locations have been calculated so that densely connected groups of nodes are close to each other. The colors correspond to modules identified by the Louvain method [8] that tries to identify densely connected groups of nodes. Both the colors and the positioning of the nodes reveal the modular structure of the egonet, and looking at the people in those modules reveals that they correspond to different aspects of my life: family and relatives, studying at university, military service, student exchange abroad, and various clubs and associations I have been involved in; one module also corresponds to my colleagues while writing this Thesis. Only the largest connected component is shown; it includes 216 out of 224 nodes. The data was extracted with *netvizz* (`https://lab.digitalmethods.net/~brieder/facebook/netvizz/`) and plotted with *Gephi* (`https://gephi.org`).

## 2.1 Graph theory

This section briefly introduces the basic concepts of graph theory required in the Thesis. Formally, a graph $G = (V, L)$ consists of a set of $N$ *nodes*, denoted by $V = \{v_1, \ldots, v_N\}$, and a set of $M$ *edges*, $L = \{e_1, \ldots, e_M\}$, where each edge is a node pair: $e_k = (v_i, v_j)$. Two nodes with an edge

between them are said to be *adjacent*, *connected*, or *neighbors*. The term *dyad* means a node pair, either connected or not; this term is less used in mathematical discussions but occurs frequently in sociology.

If the network is *undirected* the order of the two nodes in $e_k$ is irrelevant. In *directed* networks the order does matter, and the two edges $(v_i, v_j)$ and $(v_j, v_i)$ are different entities. When graphs are drawn the directed edges are commonly denoted by arrows. Figure 2.2a shows an undirected graph with $V = \{a, b, c, d\}$ and $L = \{(a, b), (a, c), (c, b), (b, d)\}$. The directed graph in Figure 2.2b has the same set of nodes but edges $L = \{(c, a), (a, b), (b, a), (b, c), (d, b)\}$.

A network is said to be *simple* if it has neither self-edges like $(v_i, v_i)$ nor multiple edges between the same pair of nodes (in the same direction in the case of directed networks). The graphs in Figures 2.2a and 2.2b are both simple. An *empty graph* has no edges, and a *full graph* has an edge between every node pair.

One common extension of this basic framework is to associate each edge with an *edge weight* $w_{ij}$, resulting in a *weighted network* $G = (V, L, W)$ where $W$ defines the weight of each edge. In empirical networks edge weights are often used to denote the strength of the relation and therefore restricted to be strictly positive; in this case $w_{ij} = 0$ is equivalent to $(v_i, v_j) \notin L$. Other uses do exist, and for example in Section 3 we discuss networks that have both positive and negative weights.

Because networks are used in various fields of science even the most central concepts have more than one name. Multiple names for the same concept are routinely used even within a single research article. As already mentioned, *graph* and *network* are synonyms, although *graph* tends to be more commonly used in theoretical discussions and *network* when referring to empirical data. The term *web* has also been used to refer to networks, but occurs less often nowadays possibly because most people associate "web" with the World Wide Web. *Vertex* is a synonym for node, and both *link* and *tie* are synonyms for edge.

The following definitions will also be used in this Thesis:

**Node degree** The degree of a node is the number of neighbors it has. In directed networks the number of incoming and outgoing links can differ and the nodes have both an *out-degree* $k_{\text{out}}$ and an *in-degree* $k_{\text{in}}$. For example in Figure 2.2b node $a$ has $k_{\text{out}} = 1$ and $k_{\text{in}} = 2$.

**Walk, path and cycle** A sequence of nodes where two consecutive nodes

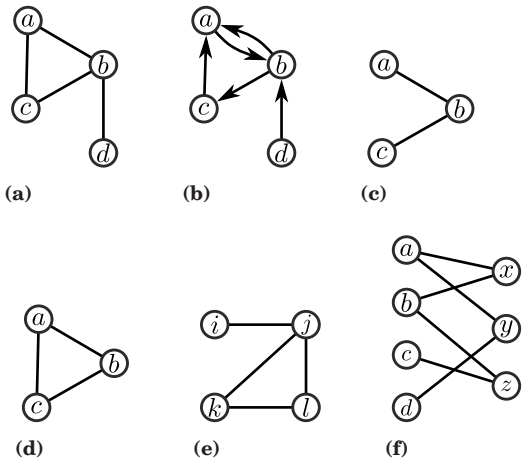



**Figure 2.2.** **(a)** An undirected graph with $N = 4$ and $M = 4$. **(b)** A directed graph with $N = 4$ and $M = 5$. **(c)** A subgraph of the graph in (a). Because the edge $(a, c)$ is missing this is not an induced subgraph. **(d)** An induced subgraph of the graph in (a). **(e)** This graph is isomorphic to the graph shown in (a). **(f)** A bipartite graph with $V_1 = \{a, b, c, d\}$ and $V_2 = \{x, y, z\}$. All edges have one end in $V_1$ and another in $V_2$.

are adjacent is called a *walk*, and the length of a walk is the number of edges traversed. In directed graphs the walk can typically progress only in the direction of the edges. A *path* is a walk where no node is repeated, and a *cycle* is a path that begins and ends at the same node.[3] For example, in the graph shown in Figure 2.2b the path $d \to b \to c \to a$ and the cycle $a \to b \to c \to a$ both have length 3. The shortest path between two nodes is also called a *geodesic path*.

**Subgraph** A graph $G' = (V', L')$ is a subgraph of $G = (V, L)$ if $V' \subseteq V$ and $L' \subseteq L$. We will mostly discuss *induced* subgraphs where for all $v_i, v_j \in V'$, if $(v_i, v_j) \in L$ then $(v_i, v_j) \in L'$; in other words, edges of the original graph are included in $G'$ whenever possible. Figure 2.2c shows a subgraph of the graph in Figure 2.2a that is however not induced; the subgraph in Figure 2.2d is induced.

**Clique** A subgraph that is a full graph is called a clique, and a clique with $k$ nodes is called a $k$-clique. For example the graph in Figure 2.2d is a 3-clique.

**Connectivity** An undirected graph is connected if there exists a path

[3]When a walk begins and ends at the same node it is called a *tour*. In reverence to Euler's early contribution to graph theory, a tour that traverses each edge of the graph exactly once is called an *Eulerian tour*.

between any two nodes; for directed graphs the issue of connectivity is more involved, as there may exist a path from $v_i$ to $v_j$ but not from $v_j$ to $v_i$ [10]. A (connected) component is a maximal sets of mutually connected nodes.

**Graph isomorphism** Two graphs $G_1$ and $G_2$ are isomorphic, denoted by $G_1 \cong G_2$, if they only differ by node labels. More formally, $G_1 \cong G_2$ if and only if there is a bijection $\pi : V_1 \to V_2$ such that $\pi(G_1) = G_2$. Here $\pi(G) = (\pi(V), \pi(L))$ where $\pi(V) = \{pi(v) \,|\, v \in V\}$ and $\pi(L) = \{(\pi(v_i), \pi(v_j)) \,|\, (v_i, v_j) \in L\}$. For example the two graphs shown in Figures 2.2a and 2.2e are isomorphic.

**Bipartite graph** A graph is bipartite if its nodes can be divided into two disjoint sets, $V_1$ and $V_2$, so that there are no edges between nodes in the same set. Figure 2.2f shows an example of such a graph. Many empirical systems can be represented as bipartite graphs. For example the authorship of scientific articles is captured by a network where $V_1$ are researchers, $V_2$ are articles, and edges connect articles to their authors.

## 2.2 Random graphs

Complex networks research is most closely related to the subfield of *random graphs*, initiated by Hungarian mathematicians Paul Erdős and Alfréd Rényi who in 1959 published an article discussing the properties of graphs obtained by placing edges at random between nodes [11]. It was this same Erdős–Rényi random graph that appeared in the first two papers on complex networks 40 years later, and in both cases because it fails to explain common properties of empirical networks: the high number of triangles in [1] and the shape of the degree distribution in [2]. Nevertheless, the Erdős–Rényi random graph has been of immense value to complex networks research. Its apparent simplicity also makes it a good starting point for discussing random graphs.

The term "random graph" is somewhat misleading; there is nothing random in any single graph. Random graph refers to an ensemble of graphs together with their occurrence probabilities. For example, the Erdős–Rényi random graph (ER graph) $G_{N,M}$ is the ensemble of all graphs with $N$ nodes and $M$ edges such that each graph is selected with an equal probability. An example with $N = 3$ and $M = 2$ is shown in Figure 2.3a. This

same random graph can also be described as a growth process: starting from an empty graph with $N$ nodes, $M$ times add an edge between two randomly selected nodes that are not yet connected. This growth process can be directly used to draw samples from $G_{N,M}$.

A slightly different variation of the ER graph is also commonly used. To draw a sample from the random graph $G_{N,p}$, start with an empty graph with $N$ nodes, go through each dyad and add an edge independently with probability $p$. This growth process corresponds to an ensemble of all graphs with $N$ nodes where a graph with $M$ edges occurs with probability $p^M(1-p)^{\binom{N}{2}-M}$; Figure 2.3b shows the ensemble when $N = 3$ and $p = 0.2$. Theoretical calculations with $G_{N,p}$ are often easier because edges are independent of each other; for example, the probability that two neighbors of a node are connected is exactly $p$. In the limit of large $N$, the two random graphs $G_{N,M}$ and $G_{N,p}$ become essentially equivalent because the number of edges in the latter is concentrated around the mean value $\langle M \rangle = p\binom{N}{2}$.

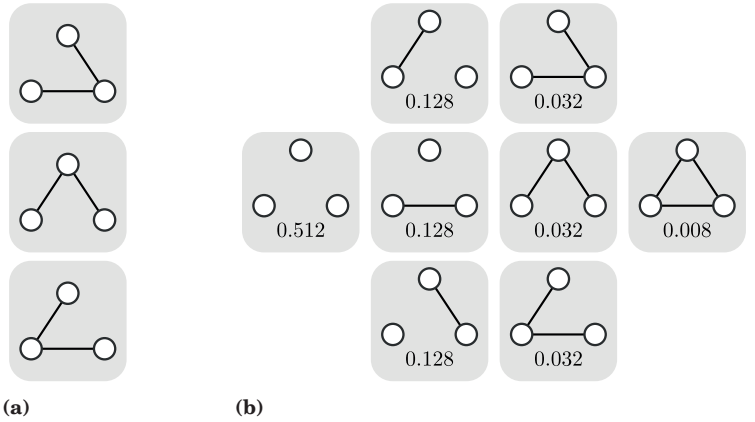
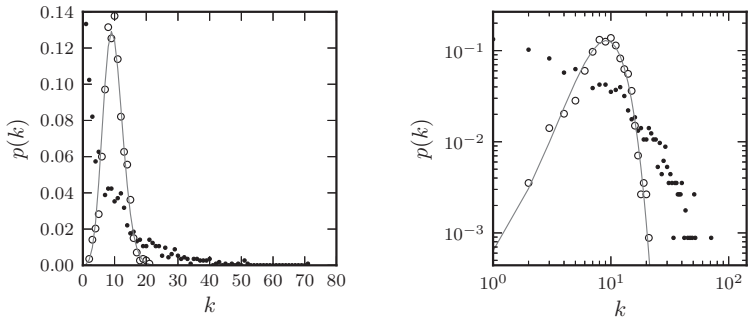



**Figure 2.3.** (**a**) The ensemble of graphs that makes up the Erdős–Rényi random graph $G_{N,M}$ with $N = 3$ and $M = 2$. In the $G_{N,M}$ model each graph with $N$ nodes and $M$ edges is equally probable. (**b**) The Erdős–Rényi random graph $G_{N,p}$ with $N = 3$ and $p = 0.2$. Now the ensemble consists of all graphs with $N$ nodes, and the probability of a graph with $M$ edges is $p^M(1-p)^{\binom{N}{2}-M}$, shown below each graph.

Although the Erdős–Rényi random graph is simple to describe, it has surprising properties. These properties are usually analyzed in the limit $N \to \infty$ so that the mean degree $\bar{k} = (N-1)p$ remains constant. One example of such a property is that when $\bar{k} > 1$, the mean size of the largest connected component has size proportional to $N$ (*giant connected component*, GCC), but when $\bar{k} < 1$ almost no graph has a GCC [10]. This kind of threshold behavior is typical for the ER graph.




**(a)** **(b)**

**Figure 2.4.** Black dots show the degree distribution of an email network with $N = 1133$ nodes and mean degree $\langle k \rangle \approx 9.62$ in **(a)** linear and **(b)** log-log coordinates. The degree distribution of an Erdős–Rényi random graph with the same mean degree is shown with circles, and the gray line denotes the asymptotic Poisson distribution. The email data is originally from [12], and the data shown here is identical to that used in Publication I.

$G_{N,p}$ also goes by the name of *Poisson random graph* because its degree distribution becomes a Poisson distribution when the mean degree $\bar{k} = (N-1)p$ is constant and $N \to \infty$.[4] The Poisson degree distribution means that most nodes have a degree close to the mean degree, while in most empirical networks the degrees instead have a very broad distribution, as shown in Figure 2.4. This discrepancy is one of the reasons why the ER random graph is not a suitable model for empirical graphs. A straightforward way to construct a model that takes into account the empirical degree distribution is to create a random graph conditional on the degrees of the nodes. This random graph is called the *configuration model* [13].

Unlike for the ER model, there is no simple algorithm for generating samples from the configuration model, but two different approaches are commonly used. The first algorithm starts from a network that has the required degree distribution and then switches the end nodes of randomly selected edge pairs; a sufficient number of switches must be made to ensure the results are statistically representative [14]. The alternative is to start with an empty graph where node $i$ has $k_i$ *stubs*, half-edges, with $k_i$ sampled from the degree distribution. Two random stubs are then connected to form an edge until no stubs are left. This algorithm is generally

[4] To see this, start from the exact binomial degree distribution $p(k) = \binom{N-1}{k} p^k (1-p)^{N-k-1}$. In the above limit $\binom{N-1}{k} p^k \approx \frac{(N-1)^k}{k!} p^k = \frac{\bar{k}^k}{k!}$ and $(1 - \frac{\bar{k}}{N-1})^{N-k-1} \xrightarrow{N \to \infty} e^{-\bar{k}}$, and combining these then gives $p(k) \to \frac{\bar{k}^k e^{-\bar{k}}}{k!}$.

faster, but the produced network is not guaranteed to be simple. However, when the network is large enough the possible self-edges and multi-edges can be removed with only minor harm to accuracy.

Because the degree distribution is often considered to be such an important property of empirical networks, the configuration model is often used as a *null model*, a reference system to compare empirical networks with; Discussing the configuration model, Newman, Strogatz & Watts suggest that "it is perhaps best to regard our random graph as a null model—a baseline from which our expectations about network structure should be measured" [13].[5] Indeed, the configuration model is commonly used to define communities in combination with modularity (Section 4.1) and to define statistically significant subgraphs in motif analysis (Section 5.1). Many properties of the configuration model can also be calculated analytically, such as the size of the largest component [13]. The

The configuration model is by definition unable to explain the origin of the degree distribution itself. It is a *phenomenological model* [16]: it makes no hypotheses about the mechanism underlying the data. Because understanding the reasons behind the common properties of various complex networks has been a central goal in the field, it has been common to create mechanistic models that could offer an explanations for those properties. One such example is the *preferential attachment model* that Barabási and Albert [2] proposed as an explanation for the broad degree distributions observed in many empirical networks. In this model a random network is constructed by starting from a small seed network. New nodes are added into the network, each connected to $m$ existing nodes so that these nodes are selected with probability proportional to their degree.

This model was, however, only the beginning of the story. As will be discussed in the following Sections, many other models produce a similar degree distribution, and it is not even obvious whether the empirical distributions really do have the shape produced by these models.

[5] The idea of comparing empirical data against some suitably randomized version of it is much older than this. Even though random graphs were introduced in 1959 [11], randomized networks were used in social network analysis already in the 1930's [15].

## 2.3 Complex networks

Complex networks research has been multidisciplinary since its inception. Because any system that consists of elements and their relations can be presented, at least approximately, as a network, it has been relatively easy to analyze data from different fields. The focus of this Thesis, social networks, will be discussed at length in the next section, but to give an idea of the generality of the approach this Section introduces a variety of empirical data studied so far in the complex networks literature.

The human brain is a network of interconnected neurons and a perfect example of a complex network in nature. However, it is not possible to study this network directly because of its extreme size—approximately 100 billion neurons each receiving from 1 to 100000 inputs from other neurons [17]—and the difficulties in obtaining data about the actual connections. One common way to study the brain network is via the *functional connectivity network* where nodes correspond to brain areas measured by functional magnetic resonance imaging (fMRI) and weighted edges denote the correlation of the activity in these areas. Networks like this have been used to study for example Alzheimer's disease [18] and the temporal properties of resting-state brain activity [19].

Other biological networks have also been studied. In food webs the nodes denote species and there is a directed edge from one species to another if the latter one eats the former [20, 21]. In *protein-protein interaction (PPI) networks* the nodes correspond to proteins and there is an edge between two nodes if those proteins have been observed to appear together in some protein complex [10]. One of the first identified properties of PPI networks was their *disassortativity*: the degree of a node and its neighbors' average degree are negatively correlated, which means that high-degree nodes tend to be connected to low-degree nodes [22]. Disassortativity in fact seems to be true for many biological and technological networks, but not for social networks [23]. *Network motifs* have also been studied extensively on biological networks as discussed in Section 5.1.

In comparison to biological networks, data on man-made networks is typically more readily available. Consider for example the World Wide Web whose structure can be extracted using a small program, a *web spider*, that follows links on web sites [24]. Web spiders are also commonly used to build data bases for WWW search engines [7]. With even more ease, *air traffic networks* can be constructed from flight schedules and

are useful for studying how global epidemics might spread via air traffic [25, 26, 27, 28].

*Citation networks* were among the first complex networks to be studied; scientists are understandably interested in the structure of science. In citation networks the nodes correspond to scientific articles and there is a directed edge from node $i$ to $j$ if article $i$ cites article $j$. It has long been known that the number of citations an article has received follows a broad distribution [29]. A more recent study found that the distribution of received citations follows a *shifted power law* $p(k_{\text{in}}) = (k_0 + k_{\text{in}})^{-\gamma}$, where $\gamma$ is between $5.6$ and $3.1$ [30]. Citation data can obviously also be used to study the structure of science, such as the division of science into separate fields [31] and the way these fields change in time [32].

While this section is by no means a complete listing of fields where complex networks can be used, it should give the reader a good idea of the generality of the approach. This generality, however, can also be a curse. Even though most methods can be readily applied to any data set, it is crucial to understand the context in order to be able to interpret the results. Blindly applying a method simply because it can be done has the potential to produce false results faster than anyone can counter.

# 3. Social networks

The credit for inventing social network analysis is commonly given to Jacob Moreno, who in the early 1930's used network analysis to explain an epidemic of runaways in a girl school in New York [34]. At the time collecting and analyzing network data was laborous, and studies were necessarily limited to relatively small networks. Figure 3.1 shows one very famous social network data set from another study, a network formed by 34 members of a karate club in a US university. In order to collect the data for this network Wayne Zachary observed the club for a duration of three years, from 1970 to 1972 [33]. What made Zachary's study so famous was in fact more or less a coincidence: during the period of observation the hired karate instructor and the club president fell into disagreement, and the club effectively split into two factions, one supporting the instructor and the other the president. These events allowed Zachary to study "how and why fission takes place in small bounded groups."

Since Zachary published his karate club study, computers have revolutionized the way social network data can be both collected and studied. A new field of *computational social science* has begun to emerge [35]. The largest social network studied so far includes over 700 million people [9]; a network of this size can only be handled with computers, and even then only with the most efficient algorithms. Curiously, in this era of computational science the data collected by Zachary is probably used more than ever. The division of the club into two factions, together with the small size of the network, has made it a convenient test bed for community detection algorithms.

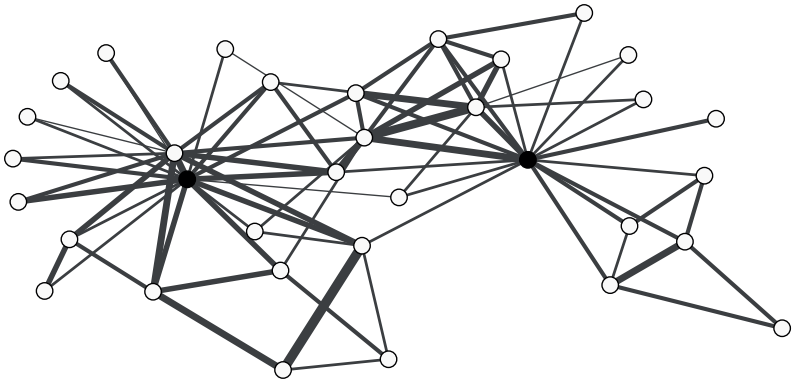


**Figure 3.1.** The social network formed by 34 members of the karate club studied by Zachary in the early 1970's [33]. The two black nodes denote the club president and the hired instructor; the network can be seen divided into two factions around these nodes. Edge widths denote the number of common contexts two people have. Eight different contexts were listed by Zachary, and the highest number of common contexts observed was seven. The data can be found in the original article, but it has seven non-symmetric edges (with $w_{ij} \neq w_{ji}$) even thought the relations are by definition symmetric. In these cases weight $\max\{w_{ij}, w_{ji}\}$ was used.

## 3.1 Social interactions

The common factor in all social networks is that nodes correspond to individuals and edges to relations between them. There are innumerable ways to define these relations, and while all networks can be analyzed with the same methods and share a number of common properties, there are important differences. Borgatti et al. [34] divided the relations into four categories, and this division is useful to get an idea of the range of possible relations:

**Similarities** of location, membership or attribute. For example in the weighted version of Zachary's karate club network the edge weights denote the number of common contexts, such as frequenting the same bar or attending the same karate competition [33]. Studying directors' co-membership in the boards of companies is useful for understanding the real structure of decision making in the industry [36]. Furthermore, it has been shown that recurring co-locations can be used to predict a social tie [37].

**Social relations** like kinship or other role, affection (likes, hates, ...), or cognitive relation (knows of, ...). Connections in online social networks such as Facebook typically denote just acquaintance, knowing

of the other person [38, 9].

**Interactions** such as mobile phone calls [39, 40], twitter messages [41, 42], face-to-face contacts [43] and emails [44, 45] have all been used to construct networks. Each form of communication yields a different network, but there is evidence that for example mobile phone calls and face-to-face interactions are experienced to be similar [46]. Other interactions have also been studied. For example the structure of sexual contact networks is important for modeling the spreading of sexually transmitted diseases [47, 48].

**Flows** of information, beliefs and resources can be used to define social networks directly, but often understanding and modeling such flows is the goal rather than the beginning of a study. The idea that social influence can flow along edges is as old as social networks analysis itself [34]. The spreading of infectious diseases has also received much attention [25, 49], and similar methods are now being used to study the flow of information [50, 51, 52, 53].

This classification is certainly not the only thinkable one, but it helps to clarify the different aspects of defining social networks. In many cases it is not possible to identify only a single class for the relation. Consider for example *core discussion networks* that are commonly used to study emotionally close friendships and are generated by asking people to name those with whom they have "discussed important personal matters during the last six months" [54]. While the question strictly speaking measures interaction, the nominations obviously reflect affection and emotional closeness. Similarly when people coordinate meetings by calling each other with mobile phones, mobile phone communication, face-to-face interactions, and similarity of location are all correlated [55].

## 3.2 Common properties of social networks

The structure of a social network naturally depends on how the relation is defined. However, many properties of social networks are common across a wide range of different relations, such as the existence of short paths between individuals discovered by Travers & Milgram [4]. More recent studies have however shown that we should be speaking of four rather than six degrees of separation [5]. Short geodesic paths are common also in other complex systems [1].

### 3.2.1 Fat-tailed degree distribution

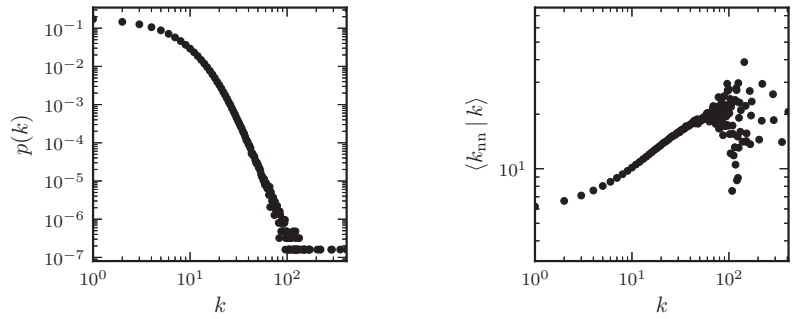



**Figure 3.2.** **(a)** Degree distribution for a mobile phone communication network with $6.24$ million nodes and mean degree $\langle k \rangle \approx 5.4$. The network is constructed from mobile phone call records of a single European mobile phone operator, and there is an edge between two nodes if there has been at least one call or SMS in both directions during a period of 6 months. **(b)** The average neighbors' degree $\langle k_{\mathrm{nn}} \mid k \rangle$ as function of node degree in the same mobile phone network. The assortativity coefficient that measures the correlation of the degrees of neighboring nodes is $0.285$.

Nearly all social networks have a broad, *fat-tailed* degree distribution, such as those shown in Figures 2.4 and 3.2a; this is in fact true even for sexual contact networks [47]. Because degree distributions often appear as straight lines in logarithmic coordinates they have been hypothesized to be *power laws* with $p(k) \propto k^{-\gamma}$. Networks with power law degree distributions have become known as *scale-free networks*, a term coined by Barabási and Albert [2].

Whether degree distributions of social networks and many other complex networks are truly power laws is still far from obvious. While statistical tests do exist [56], they are too rarely put to good use, and when they are used the evidence is often scarce [57].[1] And even though the tail of the distribution shown in Figure 3.2a does look like a power law, this is not the case for other large social networks [9, 58]. What makes the task even more difficult is that most empirical networks are samples of some larger networks, and the degree distribution of the sample and the actual network do not necessarily coincide [59]. Given these difficulties, the term "scale-free" is in practice used for nearly any network with a

[1] In fact, some published attempts to fit power laws to empirical data are bad enough to be amusing. You can even order a T-shirt with a bad power law fit from The Power Law Shop: `http://www.cafepress.com/thepowerlawshop`.

fat-tailed degree distribution.

Whatever the exact form of the distribution, the practical implication of the fat-tailed degree distribution is that there is no typical number of acquaintances: the majority of the nodes have a small degree, but there are always nodes with degree significantly higher than the mean. For example in May 2011 half of Facebook user had less than 100 contacts, the average number of contacts was about 200, and approximately 1 % of the users had over 1000 contacts[2] [9].

While the number of acquaintances varies from one individual to the next, this number is difficult to nail down also because it depends greatly on the definition of the relation. In the smaller end of the scale, core discussion networks measure the number of emotionally closest friends and typically place the average number of friends between 2 and 3 [54, 60]. On the other hand, estimates for the average number of acquaintances people know by name range between 1500 and 2000 [61]. The Facebook network is obviously somewhere between these two extremes.

A very different aspect of this question was addressed by Robin Dunbar when he suggested that our cognitive abilities place a limit for the number of social contacts we may sustain [62]. Dunbar also proposed a numerical value for this limit by studying the correlation between the *neocortex ratio*[3] and group size for non-human primates. By extrapolating this relation to humans with neocortex ratio of 4.1, the limiting group size turns out to be approximately 150 individuals, a value that has become known as *Dunbar's number*.

Dunbar's hypothesis is enticing, and it is plausible that cognitive constraints in some way limit the number of social contacts humans and other primates can sustain. However, presenting a single value for this limitation is misleading, as the 95% confidence interval for Dunbar's number ranges from $22.73$ to $446.2$.[4] As is evident from the discussion above, it is not even trivial to define what we should count when counting the number of social contacts, and defining social groups is no easier as we will see in Section 4. It seems that there is no simple answer to the simple question

[2]The maximum number of contacts was limited to 5000 at the time of the study.
[3]The neocortex makes up the majority of the cerebral cortex in the human brain. It is characterized by six layers of neurons, and is considered to be the most recent part of the cortex in evolutionary terms. Most higher cognitive functions are located in the neocortex [17]. Neocortex ratio is defined as the ratio of neocortex volume to the volume of the rest of the brain.
[4]The confidence interval originally given by Dunbar is smaller but incorrect [63].

about the average number of friends: it depends too greatly on both the individual and the definition of a friend.

### 3.2.2 Assortativity

One feature that sets social networks apart from many other networks is assortative mixing by node degree [64, 23]: high degree nodes are connected to other high degree nodes, or in other words, popular people have popular friends. Assortativity is usually quantified either by the Pearson correlation coefficient for the degrees of neighboring nodes, or by plotting the average neighbors' degree $\langle k_{\text{nn}} \rangle$ conditional on the node degree [39, 9], as shown in Figure 3.2b for the mobile phone data.

It is worth noting that even when there are no correlations between the degrees of neighboring nodes—as in the configuration model, for example—we have $\langle k_{\text{nn}} \rangle > \langle k \rangle$: on average your friends have more friends than you do. This is also very much true in social networks. But in uncorrelated networks $\langle k_{\text{nn}} \,|\, k \rangle$ is independent of $k$, while in social networks it grows with $k$ as shown in Figure 3.2b.

### 3.2.3 Clustering

Social networks are also known to have a very high number of triangles.[5] *Transitivity*, as this feature is also called, is often quantified by the *clustering coefficient*. Two slightly different definitions are commonly used: the *local* clustering coefficient $C_i$ is defined as the probability that two randomly selected neighbors of node $i$ are connected, while the *global* clustering coefficient $C$ is the probability that two randomly selected adjacent edges are in a triangle [65]. Depending on the networks and the definition used, values of the clustering coefficient typically ranges between $0.05$ and $0.5$; for the network used in Figure 3.2, $\langle C_i \rangle = 0.26$ and $C = 0.14$.

### 3.2.4 Contexts and communities

One explanation for the prevalence of triangles is the fact that most social relations have at least one *context*, such as family, work, or school, and those who share a common context are likely to know each other [66]. Contexts results in local densifications in the network and have been suggested to explain both clustering and assortativity [36]. And because people are generally knowledgeable of the contexts of their friends, contexts

[5] This is obviously not true for sexual contact networks.

might also explain how people were able to pass a message to an unknown target person in Milgram's famous experiment [67]. The existence of different contexts is also related to *homophily*, a widely studied feature of social networks that people tend to interact with others who are similar to them with respect to race, religion, education, or other socioeconomic property [68].

Community structure is central for understanding the structure of social networks, and the problem of identifying communities in empirical data has received significant attention in the literature: the review article on community detection by Santo Fortunato spans 100 pages and lists 457 references [69]. The challenges involved in defining and detecting communities will be discussed at length in Section 4.

### 3.2.5 Reciprocity

Many definitions of a social relationship are inherently reciprocal: if person $A$ is a friend of $B$, then $B$ is necessarily a friend of $A$. This is, however, not always the case, especially if we also limit the number of relations a single person may have. Consider for example networks constructed using *name generators*, that is, by asking subjects to list a small number of specific acquaintances such as five closest friends, or five collegues they prefer to go for advice. Because the nominations are not necessarily reciprocated, it makes sense to study the extent to which the resulting network is reciprocal.

The first studies on reciprocity date back to 1930's [15], and a large number of different measures for reciprocity have been proposed. One example is simply the fraction of reciprocated edges out of all edges [44]. More complicated measures can be seen to arise from attempts to put this simple number into context. For example, a large fraction of reciprocal edges is more surprising if the network is sparse, since in the limiting case of a full network all edges are necessarily reciprocal [70]. Similarly, when using name generators with a fixed number of nominations it makes sense to consider how the fixed out-degree affects reciprocity [15]; the effect of degree correlations can also be accounted for [71]. In mobile phone communication reciprocity has been shown to be a good predictor of the persistence of edges [72].

Data sets that contain some information on the strength of relationships—the number of emails between people, for example—allow studying reciprocity in more detail, and also at the level of individual edges. This kind

of *weighted reciprocity* has been studied to a lesser extent, but several measures have recently been proposed [73, 74, 75]; one measure was also defined in Publication II. Social networks are generally more reciprocal than other kinds of networks [75], but as was also shown in Publication II, there is still a large number of edges that deviate significantly from perfect reciprocity [73].

## 3.3 Social network analysis

Even though social network analysis was introduced already in the 1930's [15, 34], during much of the 20th century social networks were by and large missing from mainstream sociology. When social networks eventually started to gain currency in the 1950's they were often brushed aside as a "special method". Given the everyday observation about the importance of social relations, this omission now seems astonishing. As eloquently described by Mark Granovetter in "The Myth of Social Network Analysis as a Special Method in the Social Sciences" [76], this was to a large extent an unfortunate historical coincidence.

### 3.3.1 Networks, performance and well-being

The importance of social networks has since been confirmed in numerous studies It is has been well established that the position of individuals in the network—in addition to their personal properties and societal norms—often affects both their performance and well-being. The space here is too limited for even a shallow review of all relevant research—the Social Networks journal has been published since 1979—but the following examples should give some idea about the measurable effects of social networks.

The term *social capital* refers to the idea that some people have an advantage because of their position in the social network. For example, those who occupy "central" positions in the network might be able to affect the flow of information and therefore wield more power than their less "central" peers [34]. Because the idea of centrality is so vague, several different centrality measures have been proposed [10]. For example *betweenness centrality* measures the number of geodesic paths passing through an individual. Centrality of nodes is also important in other networks. PageRank used by Google is also a centrality measure: it gives

higher centrality to web pages that receive links from other high centrality pages [7].

Social capital may also arise from *structural holes*, a term coined by Ronald S. Burt to emphasize the importance of missing ties. Studying the performance of nearly 700 managers of a single US electronics company, Burt showed that those with more structural holes around them have higher salary, receive better evaluations from their supervisors, are more likely to be promoted, and most intriguingly, have better ideas [77]. The explanation proposed by Burt is that these people are members of multiple groups and thus exposed to a more diverse set of influences, a crucial ingredient for good ideas.

Several studies also connect social networks to the well-being of individuals. Evidence about the shrinking average size of the core discussion network—from three to two close friends between 1985 and 2004 in the United States—prompted new studies to find the cause of the change, as it could have wide-ranging social implications [60]. In addition to the size of this network, the number of contexts these most important contacts belong to might also play a role. A single strong community might give a better safety net, but at the cost of an increased pressure to conform to the common values of the community [54]. One particularly striking example is the fact that the structure of the friendship network of adolescents has been linked to suicidality [78].

### 3.3.2 Social networks and information technology

As illustrated by Zachary's karate club study, the means of collecting social network data were rather limited before the advent of electronic communication: questionnaires, interviews, and direct observations were the most commonly used methods. Since these methods are labor-intensive, the maximum size of social networks was typically limited to roughly 100 individuals. While this is enough for studying small and relatively isolated communities such as the karate club above, it was not possible to study empirically the structure and dynamics of social systems at the scale of the entire society.[6] Questionnaires are also limited to what people can recall—try listing all people you have talked with during the past

[6]Researchers of course came up with ingenious methods to circumvent this; even if it is not possible to measure the entire network, we can still measure some properties of it. Notable examples are Milgram's message passing study to estimate path lengths between people [4], and studies using family names in phone books to estimate the total number of acquaintances [61].

week—even though there are surely cases when peoples' perception of reality is more important than reality itself [34]. This can be ameliorated by asking the subjects to keep a diary on their social life, but most people find it too laborious to keep this up more than a week and may also alter their behavior knowing that they are being observed.

While large social networks derived from electronic communication have only been studied for roughly a decade, many interesting empirical results have already been obtained—results that would not have been possible with other methods.

One of the first big results obtained by analyzing large-scale empirical social networks was the verification of the so-called *Granovetter's hypothesis* [79]. The hypothesis, proposed by Mark Granovetter in 1973 in an article titled "The strength of weak ties", suggests that there should be a positive correlation between the strength of a relationship and the number of shared acquaintances. This correlation leads to the unintuitive result the title refers to, a network structure where weak ties are more important than their weight implies because they connect individuals who have only few common acquaintances. Strong ties, on the other hand, connect individual who have many common friends and therefore these ties are not as relevant to the large-scale connectivity of the network. In 2007 Onnela et al. [40] showed that this correlation can be observed in a mobile phone call network where edge weights corresponds to the total duration of calls between individuals.

Few years later Granovetter's hypothesis was also confirmed in another large-scale social network, this time using data from a massive multiplayer online game called *Pardus* [80]. Online games are a very promising source of social data since they allow recording and studying all interactions between people. Another study based on the same data verified the *social balance theory* that predicts that in a signed social network some triangles are more common than others (see Figure 3.3) [81]. The same study found also that the balance was obtained mostly by the addition of new relations, not by altering existing ones.

Another study showed that network structure, this time the number of distinct contexts, affects the probability of accepting an invitation to join Facebook [82]: if two or more of your friends are already using Facebook, it is more likely that you join if those two friends do not know each other. The way network structure affects the spreading of diseases is obviously an important problem. Based on data from a Brazilian web site used for

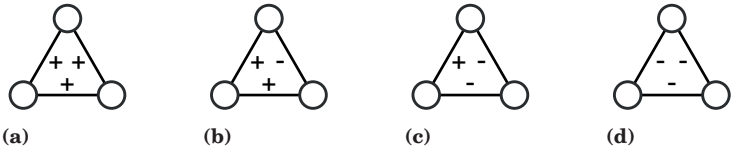



**Figure 3.3.** The four possible triangles when there can be both positive and negative relations. According to the social balance theory, triangles (a) and (c) are stable. Triangle (b) is unstable because "the friend of a friend is and enemy"; it can be balanced by turning one positive link negative or the negative link positive. Triangle (d) is unstable only in the strong formulation of the social balance theory. It was found to be underrepresented in the online game data of [81], but not as strongly as triangle (b). The reason this triangle is expected to be unstable in the online game is that any two players in it could become friends after realizing that they have a common enemy.

ranking prostitutes, Rocha et al. concluded that prostitution alone cannot explain the prevalence of most sexually transmitted diseases [48].

Online social networks also allow carrying out controlled experiments, sometimes at unprecedented scale. By using the web platform that allows direct control of the structure of the social network, Damon Centola showed that health behavior is adopted more readily in a clustered network than in a network with only a small number of triangles [83]. The clustered network allows the messages from acquaintances to synchronize temporally, and multiple concurrent messages are more effective in eliciting a reaction. Another study by Centola showed that health behavior spreads more efficiently in homophilous networks [84].

The recent experiment carried out by Bond et al. [85] is a good illustration of the research potential of online social networks. Bond et al. devised an experiment where 61.2 million US Facebook users were shown an advertisement to vote in the congressional elections in November 2010. For most users this ad included the profile pictures of six randomly selected Facebook contacts who had already reported to have voted. 0.6 million users saw the same ad without the profile pictures, and yet another 0.6 million users in the control group were shown no ad at all. The massive size of the experiment allowed detecting a statistically significant increase of 0.39 % in voting when users where shown the profile pictures; there was no difference between showing the ad without profile pictures and the control group.

### 3.4 Universal models

Researchers with a background in physics brought in a very different mindset to social network research. The contribution of the 1999 article by Barabási and Albert [2] was not so much the model they proposed to explain the power law degree distribution,[7] but the idea that networks with very different origins—an actor collaboration network, world wide web, and a power grid—might have been generated by the same, universal mechanism. The fact that this common feature was a power law distribution did little to curb the excitement, as power laws were known to be signatures of self-organized systems: "[A power law degree distribution] indicates that large networks self-organize into a scale-free state, a feature unpredicted by all existing random network models" [2].

This idea of universality is still very much present in complex networks literature. Therefore, instead of looking at social networks exclusively, many studies include data from a wide variety of different systems and hypothesize the existence of a common reason to explain their similarity. Thus an article titled "Universal features of correlated bursty behavior" suggests a common explanation for burstiness—the tendency of events to occur in bursts—that is observed not only in human communication, but also in the occurrence of earthquakes and the firing patterns of neurons [87]. In similar vein, "A universal model for mobility and migration patterns" suggests a single model to explain the commuter traffic between US counties, long-term migration patterns, number of phone calls between municipalities, and freight traffic in the US [88]. New methods are also often applied to a variety of different networks to see whether those systems can be divided into "universality classes". For example motif analysis has been used to identify "superfamilies" of networks [89].

When a common underlying mechanism is proposed, it typically takes the form of a model built on a small set of assumptions. This approach has been immensely successful in physics, but with complex systems matters are, as always, more complicated. The fit between the model and empirical data is nearly always only qualitative. While this is hardly sur-

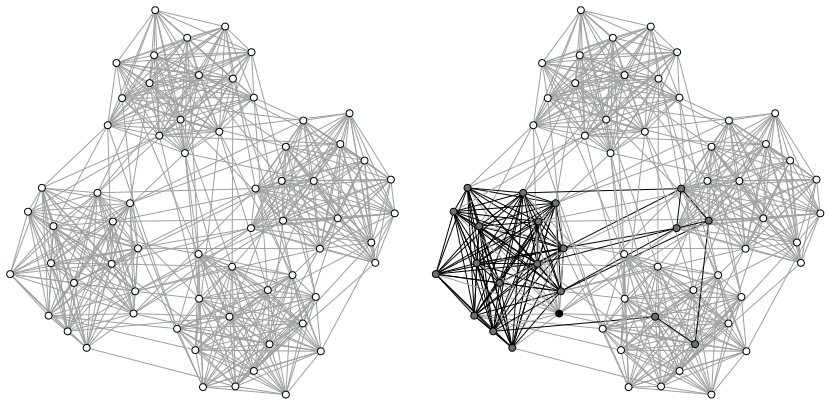
[7] It soon turned out that the idea of "preferential attachment" had been already rediscovered several times, as documented by Evelyn Fox Keller [86]. In 1965 Derek de Solla Price used the model to explain the distribution of citations of scientific articles, calling it the "cumulative advantage" effect. De Solla Price, however, appears to have been unaware that the same model was studied by Herbert Simon in 1955, who in turn attributes it to G.U. Yule in 1925.

prising with complex systems, it means that multiple models with very different assumptions and mechanisms may fit the data equally well—or equally badly—making it nearly impossible to decide which model is correct. Indeed, many models that produce power law degree distributions have been proposed since the preferential attachment model in 1999 [86, 90]. The inverse logic used by the modeling approach—that the model is correct *because* it fits the data—is broken by the imperfect fit and the existence of multiple, equally plausible models. The degree distribution *might* be due to self-organization, or it might not; the models will never tell.

In fact, in the case of power law degree distributions it is not clear whether any model is needed at all. As a generalization of the central limit theorem, power laws are obtained by summing variables with fat-tailed distributions [86], and a simple multiplicative random process is enough to produce log-normal distributions that are difficult to tell apart from power laws [90]. If simplicity is the criterion for the best explanation, both of these should be preferred over any model.

These problems are in no way limited to power law distributions, but are encountered every time any model is fitted to data [86]. Models of complex systems, however, are particularly vulnerable because the fits are nearly always only qualitative, and typically there are also other aspects of the data that are *not* explained by the model. The theoretical and empirical justificiations for the mechanism being proposed are also often weaker than in natural sciences. All this, of course, does not mean that the models are useless. Even false models are useful in many ways [91], for example to illustrate possible mechanisms and as necessary stepping stones to better models. These issues should however be kept in mind when model are proposed to explain empirical observations.

# 4. Communities in social networks

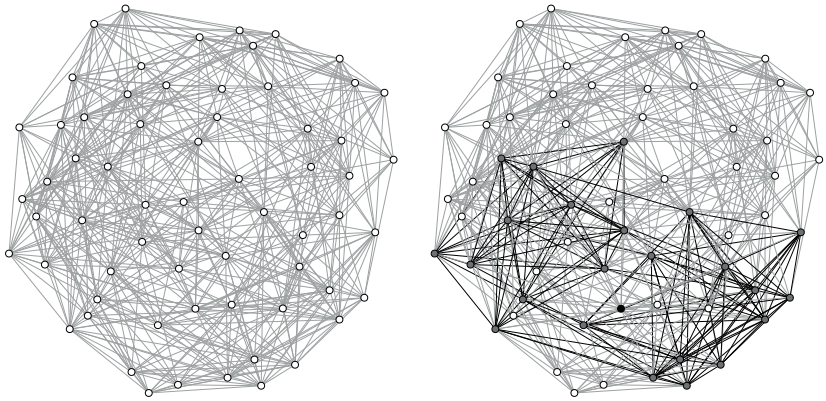


**(a)** **(b)**

**Figure 4.1.** (**a**) An artificial network with non-overlapping communities. The network has 64 nodes and 4 communities with 16 nodes each. Two nodes in the same community are always connected, and nodes in different communities are connected with probability $p = 0.05$. The resulting network has mean degree $\langle k \rangle = 17.55$. Even though the community assignment is not explicitly marked, the communities are clearly visible because of the force directed layout algorithm used to calculate the position of the nodes. (**b**) The same network, but the connections between the neighbors (gray nodes) of a single random node (black node) have been highlighted. Most neighbors are obviously in the same community as the node itself.

Most of the network properties discussed so far consider either the micro-scale—properties of individual nodes and their immediate neighbors—or the macro-scale, such as the degree distribution and other statistics of the entire network. Nearly all empirical networks also have non-trivial structure between these two scales, and one of the most studied *meso-scale* structures is the division of nodes into *communities*. Again, given the popularity of the idea in different fields the terms *module* and *cluster* are commonly used as synonyms for community.

Even though there appears to be a consensus that most empirical net-



**(a)** **(b)**

**Figure 4.2.** (**a**) An artificial network with overlapping communities. The network has 64 nodes and 12 communities; each node has been assigned to two random communities. If two nodes share at least one community they are connected with probability $p = 0.9$, otherwise they are never connected. The resulting network has mean degree $\langle k \rangle = 18.04$. (**b**) The same network, but the connections between the neighbors (gray nodes) of a single random node (black node) have been highlighted. Even though it is impossible to discern any communities in the network as a whole, the modular structure is obvious around each node.

works do have communities, there is little agreement on their exact definition. What most researchers do agree on is that communities should be connected subgraphs that are in some sense "dense".[1] Typically communities are defined implicitly by the community detection algorithm used to identify them.

Community structure is qualitatively very different depending on the number of communities per node. When the network has the kind of modular structure as shown in Figure 4.1, it is reasonable to assign each node into exactly one community. The resulting community structure is called a *partition*. Most community detection algorithms introduced so far are partition-based methods [69], and these methods have also been widely applied to social networks, for example to Facebook data [93, 94, 95] and mobile phone call networks [8, 96]. As a simple proof-of-concept, most new partition-based methods are also applied to the unweighted version of Zachary's karate club to see if they can detect the two factions described by Zachary.

When a node may simultaneously belong to multiple communities, the

[1]Non-connected groups of nodes have also been studied. For example block models can be used to identify groups of nodes that are similar with respect to their connectivity to other nodes, but are not necessarily connected themselves [92].

community structure is called a *cover* and the communities are said to overlap. This is for example the case with Facebook contact networks, as is evident from the egonet shown in Figure 2.1. One problem with overlapping communities is that they are nearly impossible to visualize at the network level, as shown in Figure 4.2. Although covers are necessarily more complicated than partitions, studying them is important because many empirical data sets appear to have overlapping communities [97].

## 4.1 Partitions

In a good partition the dense parts of the network should be contained inside communities. Probably the most widely used measure for the goodness of a partition is *modularity*, first introduced by Mark Newman in 2003 to study assortative mixing in networks [23] and then applied to community detection by Newman and Girvan in 2004 [98]. Modularity is motivated by the idea that a good partition should have more edges inside communities than expected at random. Given a partition $\mathbb{P}$, its quality is therefore measured by the modularity

$$Q(\mathbb{P}) = \sum_{c \in \mathbb{P}} \frac{l_c}{M} - \frac{E[l_c]}{M} \tag{4.1}$$

where $l_c$ is the number of edges inside community $c$ and $E[l_c]$ is the number of edges in the community expected at random. This expectation is usually defined by the configuration model, where the probability of a link between two nodes with degrees $k_i$ and $k_j$ is approximately[2] $\frac{k_i k_j}{2M}$, and thus the expected number of edges inside community $c$ is $E[l_c] = \frac{1}{2}\sum_{i \in c}\sum_{j \in c}\frac{k_i k_j}{2M} = \frac{k_c^2}{4M}$, where $k_c = \sum_{i \in c} k_i$. The modularity then becomes

$$Q(\mathbb{P}) = \sum_{c \in \mathbb{P}} \frac{l_c}{M} - \left(\frac{k_c}{2M}\right)^2 \tag{4.2}$$

### 4.1.1 Problems with modularity

At first sight modularity seems like a very reasonable measure of partition quality. It however has some counter-intuitive properties, the *resolu-*

[2] One way to draw samples from the configuration model is to start with an empty network where each node has $k_i$ *stubs* (half-links) and then connect two randomly chosen stubs $M$ times. The probability of first selecting any stub of node $i$ is $\frac{k_i}{2M}$, and the probability of then selecting any stub of node $j$ is $\frac{k_j}{2M}$. Since we can select the same nodes in either order, the total probability of adding the link $(i, j)$ during a single step is $2\frac{k_i k_j}{4M^2}$, and the probability of adding a link at any of the $M$ steps is approximately $\frac{k_i k_j}{2M}$.

*tion limit* being probably the best known. Fortunato and Barthélemy [99] showed that if a network has two modules with $m$ edges in each, and these modules are connected by a single link to each other and the rest of the network, then in the partition corresponding to the maximum modularity these two modules are assigned in the same community if $m < \sqrt{M/2}$. In general, the existence of a resolution limit means that modularity maximization cannot identify communities smaller than some threshold, no matter how good those communities otherwise appear. The basic problem is that even though modularity can be expressed as a sum of the contribution of individual communities, the terms in this sum are connected via $M$ and therefore even distant and seemingly unrelated edges affect the goodness of a community. The resolution limit can be circumvented to some extent by maximizing modularity again in each identified community [99], or by extending the definition of modularity so that it is possible to adjust the resolution and detect communities of different size [100, 101]. What makes this approach less effective is that empirical networks can have communities with highly varying sizes [102].

Modularity is also notoriously difficult to optimize. Modularity optimization has been proven to be an NP-complete problem [103]; in addition, there can be multiple local optima that correspond to very different partitions even though all have a modularity close to the global optimum [104]. It has also been shown that trees can have a high modularity; thus high modularity alone does not guarantee a modular structure [105].

### 4.1.2 The Louvain method

Because of the NP-completeness, modularity can be optimized only with heuristic methods for all but the smallest networks. To give an idea of one such heuristic, consider the *Louvain method* [8]. The Louvain method builds a hierarchy of partitions, each level a local optimum of modularity. The algorithm is initialized by creating $N$ communities, each consisting of a single node. The communities are then processed in a round-robin manner, joining each with the neighboring community that results in the largest increase of modularity. When no improvement is possible, one level of the hierarchical community structure is obtained. The communities are then turned into "supernodes" and the optimization is repeated for the new network. The algorithm finishes when modularity can no longer be improved.

One reason why the Louvain method in particular has become so pop-

ular is its good performance in the comparison of community detection methods by Lancichinetti & Fortunato [106]. This comparison also highlights another problematic feature of community detection methods: the heuristic used for optimization affects the result at least as much as the choice of the target function. The comparison includes four different modularity optimization algorithms and shows that they all give different results; in addition, the most direct attempt to optimize modularity, using simulated annealing, is outperformed by other methods. The Louvain method performs well not because it optimizes modularity, but because of the heuristic used in the optimization. The Louvain method has also been reported to yield partitions where most communities are smaller than the resolution limit [93], which essentially proves that the detected partition does not correspond to the optimal modularity.

### 4.1.3 Infomap

Many other partition-based methods have also been proposed in addition to those trying to optimize modularity. Another method that performed well in the comparison mentioned above is *Infomap* [31]. Infomap is based on the following idea. Suppose a random walker is released into the network so that at every time step it moves to a random neighboring node. If the network is modular, the random walker should remain trapped inside single communities for relatively long consecutive time periods and less often move from one community to another. Infomap identifies the best partition by finding the optimal two-level coding scheme for describing the path taken by this random walker. This coding scheme has one upper level code book for describing the steps between communities, and a single code book for each community to describe the walk inside that community. The optimal code book is a balance between having too many communities (long codes are used frequently when moving between communities) and too few (long codes are needed to describe the walks inside communities). The coding scheme that gives the shortest description corresponds to the optimal partition.

Infomap is a good example of a recurring pattern in community detection algorithms: it starts from an intuitive and reasonable idea on what good communities should be like and then introduces an algorithm to identify such communities. This approach would be most commendable if it wasn't for the disturbing observation that different intuitive ideas nearly always lead to different algorithms and consequently result in dif-

ferent partitions. This happens, of course, because there is no single "best" way to define communities. This offers little consolation for someone who would like to detect communities in an empirical network, as little research exists on the circumstances under which a given method detects the communities correctly.

### 4.2 Overlapping communities

There is a significant problem in using partition-based methods to identify social communities: the assumption that everyone has only one community is blatantly wrong. Some might even consider this to be self-evident, especially after seeing pictures of Facebook networks such as that in Figure 2.1. Indeed, the idea that people have multiple social contexts is not new; already in 1981 Scott L. Feld [66] introduced the term *focus* to denote the various contexts people have in their life and around which their social relations are organized; family, colleagues and hobby clubs are examples of different foci.

However, the idea of multiple contexts has been very often ignored. When this happens, the assumption of a single community is nearly always implicit, such as using a partition-based method to identify social communities. Curiously enough, this omission is not limited only to complex networks but appears also in social psychology [107]. The same implicit assumption was also made by Granovetter in his model to explain the strength of weak ties [79]. The model is based on the assumption that all strong triangles are closed—that is, any two close friends always know each other. However, if people have multiple contexts there is no reason to expect this to be even approximately true: two friends from different contexts generally do not know each other. In fact, it has been directly shown that among the emotionally closest relations 40 % of triangles are not closed [54]. Another direct confirmation of multiple contexts is based on Facebook data: all 20 subjects interviewed by Lampinen et al. [107] readily agreed with the idea that their Facebook contacts are divided into multiple contexts.

In complex networks literature the idea of multiple contexts has become known as *overlapping communities*, in contrast to thinking of communities as disjoint sets of nodes. Ahn et al. [97] coined the term *pervasive overlap* to describe the situation where nearly all nodes have multiple communities and showed that pervasive overlap is not limited only to so-

cial networks. A similar conclusion was reached by Reid et al. [108] who showed that in many networks partitions break maximal cliques that one would intuitively expect to always lie inside communities, and therefore no partition is likely to correspond to the "real" community structure. Publication III shows that partitions of social networks are unrealistic also in other ways.[3]

Multiple contexts have been shown to have a measurable effect on social contagion. Ugander et al. [82] showed that the probability of accepting an invitation to join Facebook depends only on the number of different contexts the person already has in Facebook, not on the total number of contacts.[4] In a similar vein, Reid & Hurley [52] studied complex contagion, a spreading process where multiple neighbors must be infected for the spreading to occur, and showed that this kind of spreading is faster in a network with overlapping community structure.

One particularly useful way to think of overlapping communities is to represent the social structure as a bipartite network where one set of nodes corresponds to people and the other to contexts. Newman and Park [65, 36] studied the theoretical properties of such networks when the contexts were independent and two people know each other with some probability $p$ if they share at least one context (this model was used to create Figure 4.2). The next step to a more realistic model would be to consider the contexts to be correlated. As shown in [54], people have a tendency to unify the context of closest friends, most often by inviting them home. As a result, the friends from different contexts become acquainted, edges have multiple contexts, and most people end up having one significant alter with whom they share both multiple contexts and many mutual friends: their spouse. Omitting the possibility of multiple contexts per edge could lead to qualitatively misleading ideas of social networks. Indeed, if the contexts were distinct, the Facebook network shown in Figure 2.1 would consist of multiple connected components. The fact that nearly all nodes

[3]The methods used both in [108] and Publication III are among those with the best performance in the recent comparison based on artificial benchmark networks [106]. The fact that these methods perform well on artificial benchmarks but badly on empirical networks illustrates one major shortcoming of using artificial networks to carry out the comparison: it is extremely difficult to verify that the artificial networks have the properties that are relevant for detecting and defining communities. Constructing a good benchmark networks is essentially as difficult as creating a realistic network model.

[4]The number of different contexts was quantified by the number of connected components in the network made up of Facebook users who had added the email address of the invited person into their address book in Facebook.

belong to the same component shows that every context includes someone with acquaintances in another context.

### 4.2.1 Detecting overlapping communities

Community detection methods that identify overlapping communities are still much more rare than partition-based methods. Introduced by Palla et al. in 2005, *clique percolation* was among the first methods proposed [109]. Clique percolation defines communities as subgraphs consisting of percolating $k$-cliques, as illustrated in Figure 4.3. Unlike the Louvain method and Infomap, clique percolation is entirely deterministic, and does not necessarily assign a community for each node: nodes not included in any $k$-clique are left outside the community structure. The most notable difference, however, is that the communities are defined explicitly, not just as the output of a given algorithm. This has allowed others to develop faster algorithms that identify exactly the same communities [110, 111], but most importantly, it means that anyone wishing to identify communities in an empirical network can judge whether the idea of community according to clique percolation corresponds to their own idea of a good community.

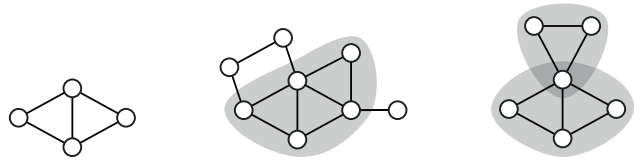



**Figure 4.3.** (In all Figures $k = 3$.) (**a**) Two $k$-cliques are adjacent if they have $k - 1$ nodes in common. In this case the two triangles are adjacent because they share the two nodes in the middle. (**b**) A $k$-clique community (gray background) is a maximal subgraph consisting of $k$-cliques such that there is a sequence of adjacent $k$-cliques between any two; a $k$-clique *percolates* through the community. The three nodes outside this community do not belong to any community because they are not included in any triangle. (**c**) The communities overlap because the central node belongs to both of them.

Despite these benefits, clique percolation often fails in practice. The only parameter is the clique size $k$, which is often too coarse: 3-cliques may percolate through the whole network, resulting in one very large community, but only a small fraction of nodes are included in any 4-clique. The requirement of a complete clique is inflexible; a 5-clique has 10 edges, but loosing even one of them makes it invisible for clique percolation. For these reasons clique percolation performs badly when the network is too

sparse [112]—which is also the case for the mobile phone data used in Publication III—or when the network contains too few cliques, as the random benchmark networks used in the recent comparison of community detection algorithms [106].

Other methods for detecting overlapping communities have also been suggested, although they are not as numerous as partition-based methods. For example, to show that pervasive overlap is common in empirical networks Ahn et al. [97] introduced a simple hierarchical partitioning of edges that consecutively joins edge pairs with the highest similarity, with similarity defined as

$$S(e_{ik}, e_{jk}) = \frac{n_+(i) \cap n_+(j)}{n_+(i) \cup n_+(j)} \tag{4.3}$$

where $n_+(i)$ is the neighborhood of node $i$ including node $i$ itself. Another method, proposed by Lee et al. [113], uses maximal cliques as seeds for communities, adds nodes to these seeds according to a local fitness measure and finally prunes communities that are too similar to each other.

## 4.3 Community detection in practice

It is worth pointing out that in some cases social communities can be successfully identified with partition-based methods. The unweighted version of Zachary's karate club network is commonly used as a simple test bed for partition-based methods, and most methods successfully recover the division described by Zachary. This network, however, consists of only *one* context: the karate club. Ironically, the overlapping nature of social communities was well known to Zachary: in the weighted version of the graph the edge weights correspond to the number of common contexts two people have. In addition to single context networks, partitions can also be useful approximations for identifying contexts in egonets, as illustrated in Figure 2.1.

It is also possible that partitions provide a good approximation for the social structure at larger scales. Humankind is divided by political borders into countries, states, and cities, and all of these divisions are partitions. Languages divide us even more. Belgium has two large monolingual communities that obviously cannot have much communication between them, which might explain why modularity optimization works so well with the Belgian mobile phone data [8]. Whether the boundary between pervasive overlap and partition is at the level of languages, states,

or cities is an open question.

Most publications on community structure either introduce new methods or discuss the performance, properties and limitations of various existing methods. In this frenzy to publish new algorithms it is easy to lose sight of the main goal of community detection: to reliably detect communities in empirical networks. Currently it is very easy to identify communities with a number of different methods, but nearly impossible to determine which answer is the correct one. Surprisingly little research exists to help select the correct method, and often the method of choice is just the most popular method one has heard of [69].

# 5. Temporal networks

Although electronic communication records, such as mobile phone calls, have been used in research for several years now, most studies have not made use of the calls directly but have instead considered networks where nodes correspond to people and edge weights denote, for example, the number of calls, emails, or tweets during some time period [39, 40, 58, 55, 114, 115]. While this approach has certainly revealed many interesting things about the structure of social networks, it leaves out one crucial dimension of communication: time. Temporal properties of human communication are now known to be highly non-trivial and far from a simple Poisson process [116, 117, 118, 119], and for example the structure of email networks has been shown to vary greatly with time [45]. One common feature is *burstiness*, the tendency of communication events to be clustered in time [120, 121, 87], leading to fat-tailed distributions of inter-events times.

This kind of temporal inhomogeneity has been shown to be important for *spreading dynamics*. Consider for example the simple susceptible-infected (SI) model that is often used to study the spreading of diseases, fads, and information. In the SI model all nodes are initially susceptible, except for a small number of infected nodes. The infection then spreads from the infected nodes to their susceptible neighbors whenever the nodes interact. Now, if the interactions occur in bursts, the infection is typically transmitted by the first event in the burs. Because the remaining events in the burst are redundant, spreading with bursty interactions is slower when compared to a spreading process with the same number of non-bursty interactions [122, 123, 51, 124, 119]— unless, of course, repeated contacts are necessary for infection [125], something that has been shown to be true at least for enlisting in a health web site [83].[1]

[1]Although there are several studies about spreading via communication events,

*Temporal networks* is a formalism for studying the properties of empirical networks where relations between nodes vary in time. Formally a temporal network $G_T = (V, E)$ consists of a set of nodes $V$—exactly as normal, static networks—and a set of events $E$. An event $e_i = (v_{i,0}, v_{i,1}, t_i, d_i) \in E$ takes place between the two nodes $v_{i,0}$ and $v_{i,1}$, begins at time $t_i$ and has duration $d_i$. Extensions of this basic formulation have also been suggested. For example, suppose the events denote the existence of connections between nodes: when there is an event between two nodes, a messages may be passed between them. Now we can introduce a *latency* $\lambda_i$ for each event $e_i$ such that when a message is sent from $v_{i,0}$ to $v_{i,1}$ at time $t$, the message is received only at time $t + \lambda_i$ (assuming $t_i \leq t \geq t_i + d_i - \lambda_i$) [126].

The *aggregate network* $G_{T_0,T_1}(G_T) = (V, L)$ is a static network with the same set of nodes as the temporal network and an edge between nodes $v_i$ and $v_j$ if there is at least one event between them during the time interval $[T_0, T_1]$; edge weights are often used to record the total number or duration of events during the time interval. When the aggregation period is the full time interval for which the data is available, the resulting analysis no longer has any temporal dimension. It is this kind of aggregate networks that have been often used to study for example mobile phone communication [39, 40] and Twitter [42].

Alternatively we can create a sequence of networks by aggregating over consecutive time intervals; such networks have been used to study for example face-to-face interactions [127], email communication [45, 127], and even the transportation of bovines [128]. The advantage of this approach is that all measures defined for static networks, such as node degree, clustering coefficient and assortativity, can be readily calculated for each aggregate network. For example the evolution of communities in time has been studied by relating clique percolation communities identified in consecutive time intervals [129]. Selecting the length of the time interval is however a balancing act. If the interval is too long, relevant dynamics are lost because they take place inside the aggregated networks. If the interval is too short, the aggregate networks become too sparse and the relevant quantities cannot be calculated. Both of these problems can be circumvented if we instead study the temporal networks directly, without the aid of the aggregate network. This approach, of course, requires new

it seems that in reality intentional information spreading almost never occurs beyond a few steps [123, 118].

methods and measures developed specifically for temporal networks.

Many definitions from graph theory have been generalized to temporal networks. However, because temporal networks are more expressive than static networks there is often no unique generalization; the problems encountered are in many ways similar to those that arise when generalizing quantities from unweighted to weighted networks [130]. Consider for example the concept of path in static networks. First of all, in temporal networks we are typically interested only in paths that respect the arrow of time, called *time-respecting paths* or *journeys* [131]. The concept of shortest path can now be generalized in at least three different ways, as shown in Figure 5.1.

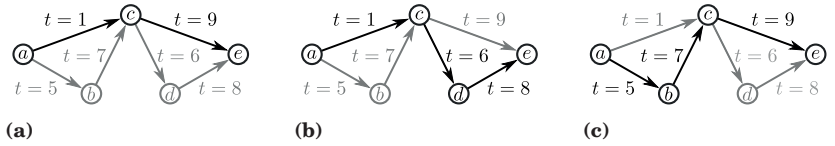



**Figure 5.1.** Example of the different definitions for the most cost efficient journey as defined in [131]. Suppose we start at time $t = 0$ and want to find the optimal path from node $a$ to $e$. **(a)** The *shortest journey* uses the smallest number of events, in this case only two. **(b)** The *foremost journey* arrives the fastest, here at time instance $t = 8$. **(c)** The *fastest journey* takes the least time from start to finish.

For a thorough discussion on temporal networks, see the review article by Holme and Saramäki [132]. The rest of this Chapter discusses *temporal motifs*. Just like paths, temporal motifs can be understood as a generalization of a concept originally introduced for static networks. The concepts presented here and originally in Publication IV are most similar to the "communication motifs" defined by Zhao et al. [133]. The term "temporal motif" and the related "dynamic motif" occur more often in the literature but have no established meaning. Thus "temporal motif" has been used as a synonym for a time-respecting path [128] or to denote the pattern of consecutive edits in Wikipedia [134]. In similar fashion, "dynamic motif" may refer to studying static motifs in consecutive aggregate networks [45], the stability of different subgraphs in a dynamic process [135], or the simultaneous activation of edges during a dynamic process [136].

## 5.1 Motifs in static networks

The concept of *network motif* was introduced in Milo et al. [20] and Shen-Orr et al. [137] in 2002, and motifs have since become widely used especially in the analysis of biological networks [138, 139]. As it was originally defined, network motifs are classes of subgraphs more common in the data than expected. The expected occurrence is typically defined by the configuration model, and the *class* of a subgraphs is defined by graph isomorphism. For example, let $C(m)$ denote the number of subgraphs that are isomorphic to a 3-cycle in a given network. If $C(m)$ is so high that it is unlikely to occur in the configuration model, then 3-cycle is a *motif*.

They are several open problems in the definition and use of motifs [140]. In particular, using a random graph to define the statistical significance of motifs is problematic. Because this same problem is faced also when analyzing temporal motifs, it is to be discussed in more detail in Section 5.3.2. However, because of this problem we adopt the usage of [141] and use the term "motif" more generally to refer to an equivalence class of subgraphs, independent of their statistical significance in comparison to some reference.

## 5.2 Temporal motifs

Motifs are equivalence classes of isomorphic subgraphs. Therefore, to generalize this concept to temporal networks we need to generalize both the concept of "subgraph" and "isomorphism".

The treatment here, as well as that in Publication IV, is limited to temporal networks where nodes have at most one event at a time, as is the case with mobile phone calls. This is easily extended to the case where events have no duration, even if they do sometimes occur simultaneously. The extension to the general case where events have durations and nodes may have any number of simultaneous events is more involved, for reasons discussed below. All concepts and methods are defined for directed events; only small changes are needed to deal with undirected events.

### 5.2.1 Temporal subgraphs

There are several alternatives for defining the subgraphs being counted in the static motif analysis, but most studies consider connected, induced

subgraphs [138]. Connectivity in static networks is a purely topological concept. In temporal networks time also has a role to play, exactly as when generalizing paths.

Given a time window $\Delta t$, we define two events to be $\Delta t$-*adjacent* if they have at least one node in common and the time difference between them—measured from the end of the first event to the beginning of the second—is no longer than $\Delta t$. This definition of adjacency directly leads to a definition of connectivity: two events are $\Delta t$-*connected* if there is a sequence of $\Delta t$-adjacent events between them. Note that unlike in journeys, the events in this sequence need not be ordered in time. These concepts are illustrated in Figure 5.2.

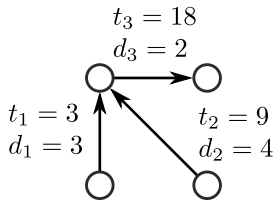



**Figure 5.2.** A small temporal network with three events. With $\Delta t = 10$ the first two events are $\Delta t$-adjacent ($t_2 - (t_1 + d_1) = 3 < \Delta t$), as well as the latter two ($t_3 - (t_2 + d_2) = 5 < \Delta t$), and therefore all three events are $\Delta t$-connected.

Any set of $\Delta t$-connected events is called a *temporal subgraph*.[2] There is however one problem with this definition, analogous to why static motifs are based on *induced* subgraphs instead of all subgraphs: in some cases the number of temporal subgraphs is not linear with respect to the number of events. Consider for example an $n$-star—a single person making $n$ consecutive calls—where all events take place within $\Delta t$. Because any two events are $\Delta t$-connected, there are $\binom{n}{k}$ temporal subgraphs with $k$ events. Not only does this distort motif counts, but counting the subgraphs becomes infeasible even for reasonable values of $n$ and $k$.

To avoid these problems we only consider *valid temporal subgraphs* in which for any two events $e_i$ and $e_j$ that share a node we always include all other events that share the same node and occur between $e_i$ and $e_j$ in time. This can be seen analogous to induced subgraphs that always include any edge between two nodes in the subgraph; in a temporal network the word *between* also refers to time, not only topology. This constraint effectively limits the number of subgraphs in the above example: there are $n-k+1 =$

[2]Note that a temporal subgraph need not include all mutually $\Delta t$-connected events.

$O(n)$ valid temporal subgraphs with $k$ events, instead of $O(n^k)$.[3]

The exact definition of temporal subgraph should ultimately depend on the purpose of the study. Whatever the definition, it should satisfy two criteria: there must be an efficient algorithm for counting subgraphs in the data, and it should allow an intuitive interpretation for the resulting counts. The definition given here, although certainly not the only one possible, is sensible for most data where events do not overlap in time. It is much more difficult to come up with a suitable definition of temporal subgraph in the more general case when events may overlap in time. The primary reason for this difficulty is the large number of ways in which events can overlap both in time and topology.

### 5.2.2 Isomorphism of temporal subgraphs

To obtain temporal motifs the subgraphs defined above must be divided into equivalence classes. Because temporal subgraphs are not really graphs—there is additional temporal information—we cannot use graph isomorphism directly. Again, there is no unique way to define this equivalence. When should two temporal subgraphs be considered similar?

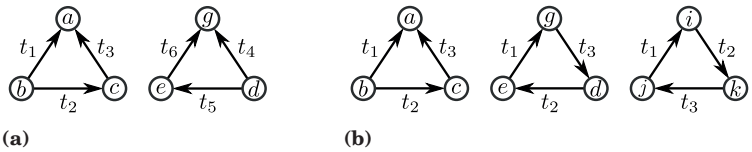



**Figure 5.3.** Here we assume $t_i < t_j$ if $i < j$; event durations have been omitted for simplicity. (**a**) Two temporal subgraphs that correspond to the same motif: the subgraphs have identical topology and the events take place in the same order. (**b**) Three temporal subgraphs that each correspond to a different motif. The first two have different topology, and in the latter two the events occur in different order.

As with static motifs we start by discarding node labels. This is however not enough: two temporal subgraphs would be equivalent only if they have the same underlying graph *and* identical event times and durations. This requirement is obviously too specific under most circumstances. In-

[3]The complexity of temporal networks is often considered to be a disadvantage, but higher complexity also means higher flexibility: the solution presented here has no counterpart in static networks. An $n$-star in a static network contains $\binom{n}{k}$ induced subgraphs with $k$ edges if we allow *arbitrary overlap of nodes and edges* between the subgraphs. Two ways to limit this number are given in [138]. If we allow *no overlap* there is only one subgraph—the central node cannot be reused—and if we allow *overlap of nodes but not edges* we obtain $\lfloor n/k \rfloor$ subgraphs; in this case a 3-star and a 5-star would both contain only one subgraph with 3 edges.

stead, we consider two subgraphs to be equivalent if the *temporal order* of their events is identical. This definition is illustrated in Figure 5.3.

### 5.2.3 Implementation

So far we have not discussed how to actually obtain the motif counts; because the purpose is to study large data sets, inefficient implementations will not do. Counting the occurrence of motifs consists of two steps: enumerating all temporal subgraphs in the data, and identifying the corresponding motif for each subgraph. The first step—detecting all temporal subgraphs—is easy to solve efficiently: one algorithm for doing this is given in Publication IV, another in [142].

To identify motifs, temporal subgraphs are first mapped into directed colored graphs as shown in Figure 5.4. A colored graph is a graph where each node has an additional property, commonly called *color*. This mapping retains all information about topology and temporal order of events, but because the result is a normal graph, existing tools developed for solving graph isomorphism can be used to divide the temporal subgraphs into equivalence classes. In practice this is done by identifying for each graph its *canonical form* that is by definition identical for two graphs if and only if they are isomorphic. Calculating the canonical form is a non-trivial task, but efficient tools have been developed for solving it.[4]

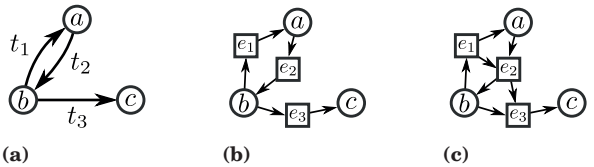



**Figure 5.4.** In order to identify temporal motifs the subgraphs are mapped into colored directed graphs for which the canonical form can be solved with existing tools. (**a**) The original temporal subgraph with three events. (**b**) An additional vertex is created for each event, with a distinct color (here denoted by a square shape) to distinguish the new vertices from those corresponding to original nodes. (**c**) Additional links are added between the event vertices to denote their temporal order: from the first event to the second, from the second to the third. This graph is now a normal directed, colored graph that contains all information about the topology and temporal order of events in the original temporal subgraph.

The algorithm we have used for identifying canonical forms, *bliss* by

[4]Canonical forms can be directly used to solve graph isomorphism, and no polynomial-time algorithm is known for solving graph isomorphism. In fact, determining whether two graphs are isomorphic is one of the few problems not proven to be in either P or NP-complete.

Junttila and Kaski [143], is an improved version of the algorithm proposed by McKay already in the late 1970's [144]. Not only are these algorithms efficient, but they are founded on using colored graphs and can therefore be directly applied to identify the canonical form of the graph shown in Figure 5.4c. Furthermore, because colored graphs are in any case being used, there is little extra cost in using node colors already in the original temporal network to denote different node types. And because events are also mapped into vertices before calculating the canonical form, different event types may be used as well.

## 5.3 Analyzing motif counts

The implementation described above enables counting temporal motifs in temporal networks with up to $10^9$ events. At this point we face the same problem that occurs when analyzing static network motifs: the motif counts alone are not very informative. In order to say whether a given motif count is high or low we need something to compare with, and very often no obvious reference is available.

### 5.3.1 References

Another data set generated by the same—or very similar—process would be an excellent reference. This approach has been very successful in the study of brain networks. Functional magnetic resonance imaging (fMRI) can be used to measure the pattern of functional connectivity between different brain areas, and by repeating the same measurement with multiple individuals it is possible to obtain an averaged functional connectivity network of the human brain. These networks can then be used to study differences in functional brain connectivity for example between age groups [145] or when performing different tasks [18].

While this approach is good for studying the brain, it is not feasible when the system under study is a social network formed by millions of individuals. We cannot obtain multiple samples of Belgium, and comparing two different countries, say France and Belgium, is more like comparing the brain of a rooster to that of a lion: while the comparison will surely reveal differences, most differences are explained by well-known reasons—population size, demographics, economy, and history—and unless we can factor out the effect of these reasons the analysis is unlikely

to be useful. And even if we were able to account for these differences, we would still be only comparing the equivalent of two brains, unable to say whether any differences are real or due to randomness. Yet, while it does not seem possible to transfer methodology from brain studies to social systems, the other direction is very much possible: any new method developed for studying social networks should also make sense when applied to brain networks.

### 5.3.2 Null models

As discussed above, the configuration model has been often used as a reference in the study of static motifs. The general idea is to construct a "null model," a suitably randomized versions of the empirical data that can be used as a target of comparison. Given the empirical motif count $C(m)$, and the mean $\mu_m$ and standard deviation $\sigma_m$ of the counts in the null model, one usually then calculates the $z$-score

$$z(m) = \frac{C(m) - \mu_m}{\sigma_m} \quad . \tag{5.1}$$

If the empirical data does not differ from the null model, the $z$-scores are expected to have zero mean and unit variance. Statistically significant deviation from this expectation suggests that whatever was retained in the null model—the degree distribution in the case of configuration model—is not enough to explain the occurrence of motifs.

Null models continue to be used even though the problems of this approach have been well documented [140].[5] The main problem is that using the configuration model as a reference is arbitrary. Artzy-Randrup et al. illustrated this with a simple toy network consisting of a $30 \times 30$ grid of nodes where nearby nodes were connected with higher probability. When compared against the configuration model, the same motifs were found to be overrepresented in this toy network as in the neural network studied by Milo et al. [20], even though the toy network is random. Comparing against a random graph does not reveal non-randomness; the only thing it reveals is whether the data was generated by that particular random graph.

[5]The comment about the problems of null models by Artzy-Randrup et al. [140] did not come from a void. Similar ideas were used in ecology already in the 1970, and the validity of the null model approach appears to have been fiercely debated [146].

### 5.3.3 Reductive null models

To better understand the logic of null models it is helpful to write out the corresponding null hypothesis. Because the only property retained in the configuration model is the degree sequence of nodes, the null hypothesis being tested by this comparison is

$H_0$ : Motif counts depend only on the degree sequence.

If $H_0$ is true, we expect the $z$-scores to be distributed with zero mean and unit variance; when this is not the case, we can conclude that the null hypothesis is likely to be false—that is, we can conclude that the data was not generated by the configuration model. Note that it makes little sense to study individual motifs separately. A single large $z$-score is sufficient evidence to conclude that $H_0$ is false, and if $H_0$ is false, it cannot be the explanation for motifs with small $z$-scores either.

In general, when studying the prevalence of measure $Q$ by comparing against a null model that retains properties $P$ of the empirical system the null hypothesis always has form

$H_0$ : $Q$ depends only on $P$.

The only information we can obtain is whether $H_0$ is true or false. In most cases the answer is already known—all empirical networks differ from simple random graphs—and therefore the information gained is negligible. Suppose, however, that we could instead study a null hypothesis of the form

$H_0$ : $Q$ does not depend on $P$.

Even if $H_0$ is now found to be false—which is still the most likely outcome for the same reasons as above—the test still continues to be useful: $E[Q]$, when calculated using the null model, is the expected value of $Q$ when the effect of $P$ has been factored out. When the set $P$ is explicitly known and contains only a small number of elements, the value $Q/E[Q]$ has an intuitive interpretation: it tells how much larger $Q$ is *because of* $P$.

The major difficulty in this approach is that instead of constructing a null model that includes only $P$, we need a null model that excludes only $P$. This can be tricky, and for many choices of $P$ even impossible. However, in Publication V we show how this can be done to test the null hypothesis

$H_0$**:** Motif counts do not depend on node types.

The constructed null model isolates the effect of node types in the occurrence of motifs, and the value $C(m)/\mu_m$ therefore reveals how much more or less common a motif is because of the node types in it. If the node type denotes for example gender, we can study the influence that gender has in time-dependent mobile phone communication patterns.

# 6. Results

## 6.1 Models of social networks

Publication I studies eight models that have been put forth as models of social networks. The goal of the study was to see whether these models could match the qualitative features common in nearly all social networks, such as a broad degree distribution, high assortativity, clustering that decreases by degree, large cliques, and the existence of a dense core. To be able to compare the models their parameters were fitted to low-level properties of two empirical data sets: the friendship network of Finnish users of the music recommendation size *last.fm* ($N = 8330$, $\langle k \rangle = 4.2$), and the network of email communication in a single university ($N = 1133$ and $\langle k \rangle = 9.6$).

None of the models is able to match all of the features of social networks. The comparison does however reveal common features among the models. Seven of the eight models studied can be divided into two categories: two *nodal attribute models* work by assigning attributes to nodes and then connect nodes according to these attribute values, whereas five *network evolution models* are based on an algorithm that defines how the network grows or changes in time. The models in each category turned out to be qualitatively similar. The nodal attribute models are able to reproduce assortativity and community structure of the empirical networks, while the network evolution models could match the degree distribution and clustering spectra. Thus it seems that neither nodal attributes or network evolution is enough to model the empirical data.

The last model, *the exponential random graph model* (ERGM), was the only phenomenological model included in the study; it does not make any assumptions about why the data is the way it is but only tries to match

the properties of the data. The main result concerning ERGM is the instability of its parameter space that made it very difficult to match its features to the empirical networks.

Publication I shows that one should be careful in using models to draw conclusions about real social networks. False models can, however, still be useful in many different ways [91]. The models analysed here show that many properties of social networks can be generated with relatively simple mechanisms, and it is plausible that several of the proposed mechanisms play some role in the formation of social networks.

### 6.2 Structure of social networks

Publication II studies the reciprocity of mobile phone communication. The main finding is that non-reciprocal relations are very common: in approximately 40% of edges one partner initiates over 70% of calls. Very similar results were obtained in [73] for relative instead of absolute number of calls. Furthermore, it was shown that lack of reciprocity is not only due to differences in activity, but rather a property of the dyad itself. High-weight edges are slightly more reciprocal; no connection between reciprocity and edge overlap was observed when the known correlation between edge weights and overlap was taken into account.

Lack of reciprocity in communication does not directly imply that the underlying relationship itself lacks reciprocity. It would be interesting to study how the measured reciprocity of communication—or the lack of it—differs from the reciprocity perceived by the people involved, and how the reciprocity of communication correlates with other ways to measure the reciprocity of relationships.

Publication III compares three highly regarded community detection methods: Infomap [31], the Louvain method [8], and clique percolation [109]. Instead of using artificial benchmark networks to carry out the comparison as done in [106] we used the three methods to identify communities in a mobile phone call network and then studied whether the detected communities correspond to intuitive conceptions of what social communities should be like, and whether the methods identify the same communities.

In general the identified communities corresponded poorly to ideas of social communities. Some communities are extremely large with well over 1000 nodes, while others are tree-like. The three methods also identify

different communities, although the Louvain communities tend to be included inside the Infomap communities. The two stochastic methods, Louvain and Infomap, output different communities on different runs.

All of these results confirm that the truthfulness of communities detected by any method should always be taken with a grain of salt. However, even if the identified communities are not "correct", analyzing their properties is useful because it reveals something about the mesoscale structure of the network: the prevalence of tree-like communities suggests that the network itself has tree-like subgraphs.

It is still not easy to summarize the structure of a large-scale social network, and describing the community structure has proven particularly elusive. At the moment there is little need for more community detection methods, even if there are still relatively few methods for identifying overlapping communities. Instead, two research lines could significantly advance the field. First, there is a pressing need for a large-scale data set with explicitly known communities to use as a reference; as shown in Publication III, methods that work well with artificial benchmarks are not guaranteed to work with empirical data sets. Second, theoretical work is needed to define communities more explicitly. On one hand we need to know what kinds of communities different data sets may contain; on the other, what kinds of communities are detected by each methods. Such theoretical work would allow selecting the most suitable community detection method for each data set.

### 6.3 Temporal motifs

Publication IV introduces the concept of temporal motifs and also the algorithms needed to identify them efficiently. Publication V then shows how temporal motifs can be used to analyze data where nodes have multiple types. This includes constructing a reference system that only excludes the effect of node and edge types. We show that this references system correctly gives a null result when there are no temporal differences between node types. All differences identified using this null model are purely temporal, in the sense that they cannot be observed from the aggregate network.

Using mobile phone call records and available meta data about the customers (age, gender and subscription type) we show that there are systematic differences between different user types. We find temporal ho-

mophily, the tendency of similar people to appear in common patterns more often than expected based on the aggregated network alone. Some of the identified patterns are easy to explain—for example returned calls are most common when the first caller is a prepaid customer and the second postpaid—but other, equally persistent patterns have no equally obvious explanation: for example motifs where one person calls two others are more common when the two recipients have similar age. We also identify temporal differences between dense and sparse neighborhoods. Both repeated calls and returned calls are more common on edges where the two nodes have few common neighbors, whereas temporal motifs that involve three nodes are more common on edges where the nodes have multiple common neighbors. As also this result is independent of the weighted, aggregate network, it can be seen as an extension of Granvetter's hypothesis to temporal networks.

The work presented in Publication IV and Publication V contains only the first steps in using temporal motifs to analyze time-dependent networks. The methods presented in these articles can be readily applied to any temporal network where nodes are involved in at most one event at a time. The generalization to multiple simultaneous events, although not simple, would significantly extend the applicability of temporal motifs. Furthermore, by using a null model that takes into account the structure of the aggregate network we can be sure that our results are independent of those obtained for static networks. Only time will tell whether these ideas will prove useful in the analysis of temporal networks.

# 7. Discussion

Massive data sets on social systems can be a treasure stash of new information. But because these data sets are so different from the small social networks studied earlier, new methods—and even new kinds of people—are needed to analyze them. All algorithms used to process massive data sets must obviously be efficient. Efficiency, however, must be achieved without sacrificing accuracy: false results that are calculated quickly and correct results impossible to calculate are both equally useless.

With such unprecedented data sets even statistical analysis requires more attention. Displaying friends' photos in a Facebook message was shown to increases voting by 0.39 %; a small difference, but statistically significant when the sample size is 61 million. Had the same study been carried out with pen and paper, arduously collecting up to 1000 replies, the result would have been exactly the opposite: no statistically significant difference between the two groups. With large data sets the focus of the discussion should be shifted from statistical significance to *effect size*—actual real world significance—but the latter term is still rarely used, in part surely because it is hard to define precisely. These challenges are in many ways similar to those encountered in genetical studies where the availability of massive genetic data sets has allowed "hunting for biological surprises," often resulting in false positives [147].

Large social data sets have yet another feature that sets them aside from older data sets: they have not been collected exclusively for research purposes. Mobile phone call records exist primarily because the operators need to bill their customers, and Twitter messages are sent to be read, not to allow their senders to be scrutinized. Consequently, even though these data sets are unprecedented both in scale and accuracy, they do not include a large number of demographic variables that are commonly used in sociology. Large data sets are *discovered* rather than *collected*, and this

might even require us to change the way we come up with research questions: instead of thinking about what data must be collected to answer a given question, we need to think up questions that can be answered with the data that is available.

Yet the most significant problem with the various data sets is not what they do or do not contain, but who has access to them. Because of privacy concerns very few, if any, large scale social networks are publicly available to the research community; the mobile phone data used in this Thesis came with a strict non-disclosure agreement required by the mobile phone operator. Facebook and other social networking sites have also financial reasons not to give out their data: the product they sell is the information they have about the users. Lack of access to data may significantly slow down research on large social networks.

When Warren Weaver introduced his classification of problems by their complexity in 1948, he also made several uplifting predictions about the future of complex systems research [3]. Weaver was certainly correct in predicting that computers would come to play a significant role; the research carried out for this Thesis would not have been possible without computers, nor would the data exist without computer-aided communication. Surprisingly, one of Weaver's predictions that is yet to materialize has nothing to do with technology. Weaver predicted that *mixed teams*, multi-disciplinary research groups, would contribute greatly to our understanding of complex systems. Although inter-disciplinary research certainly exists, ignorance about the progress made in other fields continues to be a problem; after all, preferential attachment was discovered independently at least three times during a period of 75 years.

While all of these challenges must be addressed, they are hardly too great to overcome—especially given the high benefits of understanding complex systems. It seems certain that social science is slowly but permanently being transformed into a computational science where hypotheses about the structure and dynamics of entire societies can finally be studied empirically.

# Bibliography


[1] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[2] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999.

[3] Weaver Warren. Science and Complexity. *American Scientist*, 36(536), 1948.

[4] Jeffrey Travers and Stanley Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, December 1969.

[5] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four Degrees of Separation, January 2012.

[6] Herbert A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482, December 1962.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[8] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+, July 2008.

[9] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The Anatomy of the Facebook Social Graph. *pre-print*, November 2011. arXiv:1111.4503.

[10] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, first edition, 2010.

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

[12] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68(6):065103+, December 2003.

[13] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118+, July 2001.

[14] Yael A. Randrup and Lewi Stone. Generating uniformly distributed random networks. *Physical Review E*, 72:056708+, November 2005.

[15] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, chapter 13. Structural analysis in the social sciences. Cambridge University Press, first edition, November 1994.

[16] Roman Frigg and Stephan Hartmann. Models in Science. http://plato.stanford.edu/archives/fall2012/entries/models-science/, 2012.

[17] Dale Purves, George J. Augustine, David Fitzpatrick, William C. Hall, Anthony-Samuel LaMantia, and Leonard E. White. *Neuroscience*. Sinauer Associates, Sunderland, MA, USA, fifth edition, 2012.

[18] Randy L. Buckner, Jorge Sepulcre, Tanveer Talukdar, Fenna M. Krienen, Hesheng Liu, Trey Hedden, Jessica R. Andrews-Hanna, Reisa A. Sperling, and Keith A. Johnson. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(6):1860–1873, February 2009.

[19] Stephen M. Smith, Karla L. Miller, Steen Moeller, Junqian Xu, Edward J. Auerbach, Mark W. Woolrich, Christian F. Beckmann, Mark Jenkinson, Jesper Andersson, Matthew F. Glasser, David C. Van Essen, David A. Feinberg, Essa S. Yacoub, and Kamil Ugurbil. Temporally-independent functional modes of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 109(8):3131–3136, February 2012.

[20] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, October 2002.

[21] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.

[22] Sergei Maslov and Kim Sneppen. Specificity and Stability in Topology of Protein Networks. *Science*, 296(5569):910–913, May 2002.

[23] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126+, February 2003.

[24] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, September 1999.

[25] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, December 2009.

[26] Raj K. Pan and Jari Saramäki. Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, 84(1):016105+, July 2011.

[27] Christian M. Schneider, Tamara Mihaljev, Shlomo Havlin, and Hans J. Herrmann. Suppressing epidemics with a limited amount of immunization units. *Physical Review E*, 84(6):061911+, December 2011.

[28] Alcides V. Esquivel and Martin Rosvall. Compression of Flow Can Reveal Overlapping-Module Organization in Networks. *Physical Review X*, 1(021025), December 2011.

[29] Derek de Solla Price. A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[30] Young-Ho Eom and Santo Fortunato. Characterizing and Modeling Citation Dynamics. *PLoS ONE*, 6(9):e24926+, September 2011.

[31] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, January 2008.

[32] Martin Rosvall and Carl T. Bergstrom. Mapping Change in Large Networks. *PLoS ONE*, 5(1):e8694+, January 2010.

[33] Wayne W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

[34] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network Analysis in the Social Sciences. *Science*, 323(5916):892–895, February 2009.

[35] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational Social Science. *Science*, 323(5915):721–723, February 2009.

[36] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122+, September 2003.

[37] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, December 2010.

[38] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4):330–342, October 2008.

[39] Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M. Argollo de Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179+, June 2007.

[40] Jukka-Pekka 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. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.

[41] Peter S. Dodds, Kameron D. Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE*, 6(12):e26752+, December 2011.

[42] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María P. Pérez, Gonzalo Ruiz, Francisco Sanz, Fermín Serrano, Cristina Viñas, Alfonso Tarancón, and Yamir Moreno. Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study. *PLoS ONE*, 6(8):e23883+, August 2011.

[43] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *PLoS ONE*, 6(8):e23176+, August 2011.

[44] M. E. J. Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101+, September 2002.

[45] Dan Braha and Yaneer Bar-Yam. Time-Dependent Complex Networks:Dynamic Centrality, Dynamic Motifs, and Cycles of SocialInteraction. In Thilo Gross and Hiroki Sayama, editors, *Adaptive Networks: Theory, Models and Applications*, Springer Studies on Complexity, pages 39–50. Springer Berlin / Heidelberg, 2009.

[46] Ruth Rettie. Mobile Phone Communication: Extending Goffman to Mediated Interaction. *Sociology*, 43(3):421–438, June 2009.

[47] Fredrik Liljeros, Christofer R. Edling, Luis A. Amaral, H. Eugene Stanley, and Yvonne Aberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, June 2001.

[48] Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts. *PLoS Computational Biology*, 7(3):e1001109+, March 2011.

[49] Simon Cauchemez, Achuyt Bhattarai, Tiffany L. Marchbanks, Ryan P. Fagan, Stephen Ostroff, Neil M. Ferguson, David Swerdlow, and the Pennsylvania H1N1 working group. Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108(7):2825–2830, February 2011.

[50] Dashun Wang, Zhen Wen, Hanghang Tong, Ching-Yung Lin, Chaoming Song, and Albert-László Barabási. Information Spreading in Context. In *20th International World Wide Web Conference*, 2011.

[51] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102+, April 2011.

[52] Fergal Reid and Neil Hurley. Diffusion in Networks With Overlapping Community Structure. *pre-print*, September 2011. arXiv:1105.5849.

[53] José L. Iribarren and Esteban Moro. Affinity Paths and Information Diffusion in Social Networks. *Social Networks*, 33(2):134–142, May 2011.

[54] Gerald Mollenhorst, Beate Völker, and Henk Flap. Shared contexts and triadic closure in core discussion networks. *Social Networks*, 33(4):292–302, October 2011.

[55] Francesco Calabrese, Zbigniew Smoreda, Vincent D. Blondel, and Carlo Ratti. Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data. *PLoS ONE*, 6(7):e20814+, July 2011.

[56] Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, February 2009.

[57] Michael P. H. Stumpf and Mason A. Porter. Critical Truths About Power Laws. *Science*, 335(6069):665–666, February 2012.

[58] Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskovec. Mobile call graphs: beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 596–604, New York, NY, USA, 2008. ACM.

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

[60] Matthew E. Brashears. Small networks and high isolation? A reexamination of American discussion networks. *Social Networks*, 33(4):331–341, October 2011.

[61] Peter D. Killworth, Eugene C. Johnsen, Bernard, Gene Ann Shelley, and Christopher McCarty. Estimating the size of personal networks. *Social Networks*, 12(4):289–312, December 1990.

[62] R. I. M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(04):681–694, November 1993.

[63] Charles H. Janson. Primate group size, brains and communication: A New World perspective. *Behavioral and Brain Sciences*, 16(4):711–712, November 1993.

[64] M. E. J. Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701+, October 2002.

[65] M. E. J. Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):026121+, August 2003.

[66] Scott L. Feld. The Focused Organization of Social Ties. *American Journal of Sociology*, 86(5):1015–1035, 1981.

[67] Duncan J. Watts, Peter S. Dodds, and M. E. J. Newman. Identity and Search in Social Networks. *Science*, 296(5571):1302–1305, May 2002.

[68] Miller McPherson, Lynn S. Lovin, and James M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[69] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, February 2010.

[70] Diego Garlaschelli and Maria I. Loffredo. Patterns of Link Reciprocity in Directed Networks. *Physical Review Letters*, 93(26), 2004.

[71] Gorka Zamora-López, Vinko Zlatić, Changsong Zhou, Hrvoje Štefančić, and Jürgen Kurths. Reciprocity of networks with degree correlations and arbitrary degree sequences. *Physical Review E*, 77(1):016106+, January 2008.

[72] Cesar A. Hidalgo and C. Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, May 2008.

[73] Cheng Wang, Anthony Strathman, Omar Lizardo, David Hachen, Zoltan Toroczkai, and Nitesh V. Chawla. Weighted reciprocity in human communication networks. *pre-print*, September 2011. arXiv:1108.2822.

[74] Leman Akoglu, Pedro O. S. Vaz de Melo, and Christos Faloutsos. Quantifying reciprocity in large weighted communication networks. In *Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, PAKDD'12, pages 85–96, Berlin, Heidelberg, 2012. Springer-Verlag.

[75] Tiziano Squartini, Francesco Picciolo, Franco Ruzzenenti, and Diego Garlaschelli. Reciprocity of weighted networks, August 2012. arXiv:1208.4208.

[76] Mark Granovetter. The Myth of Social Network Analysis as a Special Method in the Social Sciences. *Connections*, 13(2):13–16, 1990.

[77] Ronald S. Burt. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–399, September 2004.

[78] Peter S. Bearman and James Moody. Suicide and friendships among American adolescents. *American journal of public health*, 94(1):89–95, January 2004.

[79] Mark S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[80] Michael Szell and Stefan Thurner. Measuring social dynamics in a massive multiplayer online game. *Social Networks*, 32(4):313–329, October 2010.

[81] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, August 2010.

[82] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, April 2012.

[83] Damon Centola. The Spread of Behavior in an Online Social Network Experiment. *Science*, 329(5996):1194–1197, September 2010.

[84] Damon Centola. An Experimental Study of Homophily in the Adoption of Health Behavior. *Science*, 334(6060):1269–1272, December 2011.

[85] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, September 2012.

[86] Evelyn Fox Keller. Revisiting "scale-free" networks. *Bioessays*, 27(10):1060–1068, October 2005.

[87] Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2:397+, May 2012.

[88] Filippo Simini, Marta C. González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, February 2012.

[89] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of Evolved and Designed Networks. *Science*, 303(5663):1538–1542, March 2004.

[90] Guido Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology (Oxford Finance)*. Oxford University Press, USA, June 2007.

[91] W. C. Wimsatt. False Models as Means to Truer Theories. In M. Nitecki and A. Hoffman, editors, *Neutral Models in Biology*, chapter 6, pages 23–55. Oxford University Press, December 1987.

[92] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107+, January 2011.

[93] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Alessandro Provetti. The role of strong and weak ties in Facebook: a community structure perspective. *pre-print*, March 2012. arXiv:1203.0535.

[94] Emilio Ferrara. A Large-Scale Community Structure Analysis In Facebook, March 2012. arXiv:1106.2503.

[95] Amanda L. Traud, Eric D. Kelsic, Peter J. Mucha, and Mason A. Porter. Community Structure in Online Collegiate Social Networks. *pre-print*, October 2010. arXiv:0809.0690.

[96] Paul Expert, Tim S. Evans, Vincent D. Blondel, and Renaud Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, May 2011.

[97] Yong-Yeol Y. Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, August 2010.

[98] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+, August 2003.

[99] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, January 2007.

[100] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész. Limited resolution in complex network community detection with Potts model approach. *The European Physical Journal B*, 56(1):41–45, March 2007.

[101] Jussi M. Kumpula, Jari Saramäki, Kimmo Kaski, and János Kertész. Limited resolution and multiresolution methods in complex network community detection. *Fluctuation and Noise Letters*, 7(3):206–214, September 2007.

[102] Andrea Lancichinetti and Santo Fortunato. Limits of modularity maximization in community detection. *Physical Review E*, 84(6):066122+, December 2011.

[103] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On Finding Graph Clusterings with Maximum Modularity Graph-Theoretic Concepts in Computer Science. In Andreas Brandstädt, Dieter Kratsch, and Haiko Müller, editors, *Graph-Theoretic Concepts in Computer Science*, volume 4769 of *Lecture Notes in Computer Science*, chapter 12, pages 121–132. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2007.

[104] Benjamin H. Good, Yves A. de Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106+, April 2010.

[105] James P. Bagrow. Communities and bottlenecks: Trees and treelike networks have high modularity. *Physical Review E*, 85(6):066118+, June 2012.

[106] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117+, November 2009.

[107] Airi Lampinen, Sakari Tamminen, and Antti Oulasvirta. All My People Right Here, Right Now: management of group co-presence on a social networking site. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 281–290, New York, NY, USA, 2009. ACM.

[108] Fergal Reid, Aaron McDaid, and Neil Hurley. Partitioning Breaks Communities. *pre-print*, May 2011. arXiv:1105.5344.

[109] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[110] Jussi M. Kumpula, Mikko Kivela, Kimmo Kaski, and Jari Saramaki. A sequential algorithm for fast clique percolation. *Physical Review E*, 78(2):026109+, July 2008.

[111] Fergal Reid, Aaron McDaid, and Neil Hurley. Percolation Computation in Complex Networks, April 2012. arXiv:1205.0038.

[112] E. N. Sawardecker, M. Sales-Pardo, and L. A. Amaral. Detection of node group membership in networks with group overlap. *The European Physical Journal B*, 67(3):277–284, November 2008.

[113] Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley. Detecting highly overlapping community structure by greedy clique expansion. *pre-print*, June 2010. arXiv:1002.1827.

[114] Renaud Lambiotte, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, September 2008.

[115] J. Saramaki, E. A. Leicht, E. Lopez, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar. The persistence of social signatures in human communication. *pre-print*, April 2012. arXiv:1204.5602.

[116] João G. Oliveira and Albert-László Barabási. Human dynamics: Darwin and Einstein correspondence patterns. *Nature*, 437(7063):1251, October 2005.

[117] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, May 2005.

[118] Fernando Peruani and Lionel Tabourier. Directedness of Information Flow in Mobile Phone Communication Networks. *PLoS ONE*, 6(12):e28860+, December 2011.

[119] Mikko Kivelä, Raj K. Pan, Kimmo Kaski, János Kertész, Jari Saramäki, and Márton Karsai. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03005+, March 2012.

[120] K. Goh and A. Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002+, February 2008.

[121] Márton Karsai, Kimmo Kaski, and János Kertész. Correlated Dynamics in Egocentric Communication Networks. *PLoS ONE*, 7(7):e40612+, July 2012.

[122] Alexei Vazquez, Balázs Rácz, András Lukács, and Albert L. Barabási. Impact of Non-Poissonian Activity Patterns on Spreading Processes. *Physical Review Letters*, 98(15):158702+, April 2007.

[123] José L. Iribarren and Esteban Moro. Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Physical Review Letters*, 103(3):038702+, July 2009.

[124] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A. L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102+, February 2011.

[125] Taro Takaguchi, Naoki Masuda, and Petter Holme. Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics. *pre-print*, June 2012. arXiv:1206.2097.

[126] Nicola Santoro, Walter Quattrociocchi, Paola Flocchini, Arnaud Casteigts, and Frederic Amblard. Time-Varying Graphs and Social Network Analysis: Temporal Indicators and Metrics. *pre-print*, February 2011. arXiv:1102.0629.

[127] John Tang, Mirco Musolesi, Cecilia Mascolo, and Vito Latora. Temporal Distance Metrics for Social Network Analysis. In *Proceedings of the Second ACM SIGCOMM Workshop on Online Social Networks*, WOSN '09, pages 31–36, New York, NY, USA, August 2009. ACM.

[128] Paolo Bajardi, Alain Barrat, Fabrizio Natale, Lara Savini, and Vittoria Colizza. Dynamical Patterns of Cattle Trade Movements. *PLoS ONE*, 6(5):e19869+, May 2011.

[129] Gergely Palla, Albert-László L. Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, April 2007.

[130] Jari Saramäki, Mikko Kivelä, Jukka P. Onnela, Kimmo Kaski, and János Kertész. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105+, February 2007.

[131] Bui B. Xuan, A. Ferreira, and A. Jarry. Computing shortest, fastest, and foremost journeys in dynamic networks. *International Journal of Foundations of Computer Science*, 14(2):267–285, 2003.

[132] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, October 2012.

[133] Qiankun Zhao, Yuan Tian, Qi He, Nuria Oliver, Ruoming Jin, and Wang-Chien Lee. Communication Motifs: A Tool to Characterize Social Communications. In *CIKM*, pages 1645–1648, October 2010.

[134] David Jurgens and Tsai-Ching Lu. Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia. In *Sixth International AAAI Conference on Weblogs and Social Media*, June 2012.

[135] Robert J. Prill, Pablo A. Iglesias, and Andre Levchenko. Dynamic Properties of Network Motifs Contribute to Biological Network Organization. *PLoS Biol*, 3(11):e343+, October 2005.

[136] Lisa Schramm, Vander V. Martins, Yaochu Jin, and Bernhard Sendhoff. Analysis of Gene Regulatory Network Motifs in Evolutionary Development of Multicellular Organisms. In *12th International Conference on the Synthesis and Simulation of Living Systems (ALife XII)*, October 2010.

[137] Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68, April 2002.

[138] Giovanni Ciriello and Concettina Guerra. A review on models and algorithms for motif discovery in protein–protein interaction networks. *Briefings in Functional Genomics & Proteomics*, 7(2):147–156, March 2008.

[139] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, June 2007.

[140] Yael Artzy-Randrup, Sarel J. Fleishman, Nir Ben-Tal, and Lewi Stone. Comment on "Network Motifs: Simple Building Blocks of Complex Networks" and "Superfamilies of Evolved and Designed Networks". *Science*, 305(5687):1107, August 2004.

[141] Jukka-Pekka Onnela, Jari Saramäki, Janos Kertész, and Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103+, June 2005.

[142] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs. In Peter Holme and Jari Saramäki, editors, *Temporal Networks*, Springer Complexity Series. Springer, 2013. In press.

[143] Tommi Junttila and Petteri Kaski. Engineering an efficient canonical labeling tool for large and sparse graphs. In David Applegate, Gerth S. Brodal, Daniel Panario, and Robert Sedgewick, editors, *Proceedings of the Ninth Workshop on Algorithm Engineeringand Experiments and the Fourth Workshop on Analytic Algorithms and Combinatorics*, pages 135–149. SIAM, 2007.

[144] Brendan D. McKay. Practical Graph Isomorphism. *Congressus Numerantium*, 30:45–87, 1981.

[145] Stefanie Brassen, Matthias Gamer, Jan Peters, Sebastian Gluth, and Christian Büchel. Don't Look Back in Anger! Responsiveness to Missed Chances in Successful and Nonsuccessful Aging. *Science*, 336(6081):612–614, May 2012.

[146] Nicholas J. Gotelli. Research frontiers in null model analysis. *Global Ecology & Biogeography*, 10(4):337–343, 2001.

[147] Daniel MacArthur. Methods: Face up to false positives. *Nature*, 487(7408):427–428, July 2012.

Aalto-DD 81/2013

ISBN 978-952-60-5164-2
ISBN 978-952-60-5166-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

**Aalto University**
**School of Science**
**Department of Biomedical Engineering and Computational Science**

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

**DOCTORAL DISSERTATIONS**