Time-Series-Based Reconstruction and Analysis of Complex Networks -

Methods for Quantitative Comparison of Dynamic Processes

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

der Naturwissenschaftlichen Fakultät II
des Fachbereichs Physik
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von
Dipl. Physiker Mirko Kämpf
geb. am 16. April 1975 in Stollberg / Erzg.

Gutachter:
PD Dr. J. W. Kantelhardt
Prof. Dr. S. Trimper
Prof. Dr. H. Kantz

Datum der Verteidigung: 29. Mai 2017

Contents

1. Preface

The Wikipedia project is an excellent example of great success in several contexts. First to mention is the huge number of contributors and users worldwide which made Wikipedia one of most often used web sites on the world wide web[1]. This is even more noticeable as no commercial interest is behind Wikipedia and the Wikimedia Foundation. A self-organized global community of enthusiasts achieved a remarkable result within one decade. They created a public encyclopedia which represents the world's knowledge in more than 245 languages[2]. Before the public break-through of the free global online encyclopedia Wikipedia's completeness and accuracy seemed to be possible only for commercial publishers. Instead of relying on a strong editorial process based on a few experts, the contributions of a large public crowd are the fundamental base of Wikipedia. Second, the software behind Wikipedia was developed by a self-organized group of open source developers. This illustrates again (after the success of Linux as a free and open operating system for computers) how free open software can be seen as a catalyst in the process of forming open knowledge bases in a public space, such as the Internet. This kind of knowledge management should be considered to be a cultural achievement of recent history[3].

A lot of different people use Wikipedia in many different contexts. While the majority of Wikipedia users just reads available articles, a still large number of people contribute actively to this public knowledge base[4]. Many different research projects are focused on Wikipedia. Some commercial products use Wikipedia to enrich their own data and even the recent gamification trend was not ignored by Wikipedia. *The Wikipedia Adventure* is an interactive game based tutorial to teach people how to contribute to Wikipedia. This way, the whole editorial process is less random and better organized while cultural differences and diversity still exist.

This work follows a generic approach as it uses Wikipedia data as a proxy for social media applications (SMA). Most SMAs are online communication networks. They connect collaborating users and also allow a transparent communication among controversial users. This social network aspect is obviously a very dynamical part, and can be found in a variety of complex systems. A more static aspect is related to the inherent structure of content - usually a set of interlinked web resources. This aspect is called static, but in reality it is dynamically too. It evolves over time but on a different time scale. Both aspects can be modeled as individual networks. Obviously, the two different aspects can also be combined in one model which represents a complex system.

Wikipedia shows many properties of a complex system. Because the entire system should not always be truncated into slices but rather be treated as a holistic system, nowadays Networks of Networks (NoN) are used. NoNs became the representation of complex systems in more and more research projects. This work contributes to the field of NoNs as it provides data acquisition and preparation techniques to model functional aspects of complex systems based on data as layered and integrated networks, one network for each individual functional aspect, which all, if combined, form a NoN. Initially, this work started with individual sub projects focused on time series analysis. The final goal is the development of a formalism and a methodology for advanced data-driven social media analysis - including dynamics, structure, and evolution of structural properties as a function of time within several domain specific contexts. Recent relevance studies about news articles and media coverage [1] and market analysis [2] are example use cases which can be generalized to technical systems such as sensor data analysis for predictive maintenance in industry, for traffic control, as well as to risk analysis and fraud detection in financial services. This relates the work to the young field of *Econophysics* introduced by Mantegna and Stanley [3].

Since the advent of Econophysics, which supports a totally new approach in interdisciplinary research and requires a connection between social science, economics, and natural science, many studies have investigated properties of social networks and especially socio-technical systems. However, while many social systems are intuitively connected with each other, little research exists on inter-connections between coexisting dynamic aspects such as usage and growth. If a system grows, it is usually not in equilibrium. This means, we have non-stationary processes and can not expect to measure stationary time series.

A combination of different public available information sources from and about Wikipedia allows us to identify trends and to normalize measured data using a non-parametric approach. We can thus describe the time evolution of a complex system, such as an emerging market, by deriving characteristic properties from a variety of directly measurable variables. We have used the emerging Big Data market as a case study [4]. Wikipedia provides both, primary data and background information topics belonging to a particular topic of interest.

[1] According to `https://en.wikipedia.org/wiki/List_of_most_popular_websites` Wikipedia ranks as number six in the Alexa Traffic Rank in August 2015

[2] In October 2015, Wikipedia has 245 sub projects which contain 10 or more articles and which received 10 or more edits in last month (see: `https://stats.wikimedia.org/EN/Sitemap.htm`)

[3] According to Gesellschaft zur Förderung Freien Wissens e.V. Wikipedia deserves recognition and protection as UNESCO's first digital World Cultural Heritage Site.

[4] The column labeled *participants* in the first table on `https://stats.wikimedia.org/EN/Sitemap.htm` illustrates the ratio between active speakers of a language and the number of Wikipedia editors contributing to that language.

Furthermore, we can compare the representation of a topic - such as the Big Data market - within Wikipedia. Our new approach is based on data from arbitrary Wikipedia topics and includes page content, link structure, and usage patterns. The growth rate and editorial dynamics of selected pages provide useful information. The growth of Wikipedia is studied based on the edit history of Wikipedia pages, while the information retrieval process is studied with hourly click count data - both provided by the Wikimedia Foundation [5].

Especially the recent emergence of such large public data sets together with cloud based computation methods and highly scalable implementations of network analysis algorithms define another context for this work. Further motivation for creation of correlation and dependency networks comes from recently introduced network measures for multiplex networks.

Finally, the combination of available open data, new data preparation strategies, and new analysis methods can be seen as an important contribution to modern interdisciplinary data-driven research. Wikipedia was in the focus of this research project, and it also was a great source of inspiration. The open public data, which Wikipedia consists of, was produced by a community of people using open software. Now we can see clearly that there is again a need for open software, which allows efficient analysis of all this public information. Analysis of search traffic and stock market prices are established, but many more applications are possible, especially if usage data and content analysis are combined. The latest example for using open software systems in order to increase transparency in economy and politics is called: "The Panama Papers." One has to be really careful, since such methods, if combined with the right data sets, can also be abused in order to increase the political pressure on a group of people or individuals.

By establishing an open data culture and a feedback loop between data creation and analysis we should be able to generate more transparency. All this can support knowledge creation and sustainable usage patterns for Wikipedia in particular, and other Social Media Applications (SMA) in general after the big hype is over.

In general, research questions and hypotheses have to be defined carefully. Especially if data is already available before a question is asked or a hypothesis is defined, it is possible to align the research question too much with already existing data. This approach is dangerous and can lead to wrong results, although data exploration is an important part in the data science process [6]. According to [7] (p.16) "*A well-formulated hypothesis will be both quantifiable and testable, that is, involve measurable quantities or refer to items that may be assigned to mutually exclusive categories.*" This means, a good hypothesis is one which can be verified and validated by application of statistical tests to data.

Potentially disturbing effects of measurement procedures (often indirect measurements[5]) and data collection techniques on the underlying system have to be evaluated and minimized. This ensures, that the process which is in the focus of a study is not manipulated in an unwanted and unpredictable way.

The number of measurements and simulation runs - in general the number of conductible experiments - is limited. One has to select the right sampling methods and an appropriate definition of the study focus to eliminate the selection bias as much as possible. If bias cannot be avoided then it must be quantified.

During the rise of Big Data technology also data analysis methods had (and still have) to be developed further. Especially new study types, based on an increasing variety of larger and more diverse data sets allow integration of different but so far isolated scientific disciplines. It is necessary to identify the right scope, e.g., the appropriate resolution for time series and spatial data. The ranges for sliding window analysis, and the right number of neighbor nodes within networks have to be chosen with care, especially when data is collected from multiple individual systems. Combining data sets often means also combining different measurement techniques. That's why normalization, filtering and (re)sampling become fundamental elements in study preparation and experiment design. Finally, terminology has to be integrated or translated between different scientific disciplines in order to have a benefit from existing best practices. Finally, this leads to new requirements for data analysis software especially regarding management of metadata.

1.1. Addressed Problems

The following four specific problems have been addressed in this work within several sub-projects:

(P1) - *Uni-variate and multi-variate time series analysis*: Analysis of (a) traffic data (measured and simulated), (b) evacuation simulation results, and (c) social media usage data was done using established methods. Results have been published in [8, 9, 10]

(P2) - *Study design for interdisciplinary computational science*: Common studies on social online media are affected by a strong selection bias. In many cases not much is known about users. In many cases, demographic, cultural, geographic, economic, and even political issues influence online platforms. Studies on massive online data

[5] A quantitative analysis is not possible directly in many cases because the variables, one is interested in, are not accessible directly. This means, one can not measure interactions or relations between elements or subsystems directly. Indirect measurements have to be used instead. Such indirect measurements are pretty common and based on well known or assumed relations between the accessible variable and the target variable. Correlation analysis reveals even more details than analysis of individual measurement results, e.g., relations between variables which again can lead to an indirect measurement procedure. One has to differentiate aspects, which are simply not measurable but exist and such effects, which do not exist in isolated systems. In general, not all hidden variables also lead to emerging phenomena in complex systems.

can be improved and validated only if comparable reference data is available and by using normalization, which has to be adaptive and context sensitive. We consider Wikipedia as a potential source for context networks to identify trends and contextual bias.

(P3) - *Detrending of raw media usage data:* Development of new normalization methods and measures for media usage analysis allows integrated analysis of different communication channels and finally a comparison of those. In general, one can analyze individual communication channels as already done by many researchers, but the remaining challenge is to understand how the importance, acceptance, and reliability of different channels changes over time, while the system is far away from equilibrium. Results related to P2 and P3 have been published in [4].

(P4) - *Integration of multiple facets of complex systems*: Generalization and integration of existing network creation methods and analysis techniques enables us to develop a unified framework for large-scale simulations and data analysis as a tool for complex systems research on top of distributed IT infrastructures. First results have been published in [11].

1.2. Structure of This Work

This work consists of three main parts followed by an additional case study (which was published in a slightly shortened version [4]) and the reproduction of three publications [12, 13, 9] in an appendix.

Part I - *Introduction*: Chapter 2 provides an introduction to the scientific discipline *complex systems research* (CSR). A summary of relevant aspects from network theory is presented in chapter 3 followed by an introduction to social network research in chapter 4. There, it is shown why Wikipedia can be considered as a stub for user interest in particular topics, and that Wikipedia is a complex system. Finally, all used time series analysis methods - most importantly Detrended Fluctuation Analysis (DFA), Return Interval Statistics (RIS), Cross-Correlation (CC), and Event-Synchronization (ES) for pairs of time series - are introduced in chapter 5. The chapter finishes with a discussion of data generation techniques, which allow creation of correlated time series pairs and time series with long-term correlations.

Part II - *Novel analysis methods and data preparation methods are introduced in part II.*: Chapter 6 introduces the idea of neighborhood networks, which define the context for a given network node. Such neighborhood graphs allow a contextual normalization and contextual detrending of raw data. More details about study design and data preparation procedures are described in chapter 7. Especially in case of long-term studies it is important to know the stability of structural properties of the system one is interested in. In chapter 8 we discuss the problem of non-stationary real networks and how to overcome the limitations. As an example, we study and compare the life cycle of several Wikipedia sub-projects in different languages. We introduce a generic framework for network reconstruction from time series of multiple different types in chapter 9. Creation of functional networks is the central element of this work. Thus, the selection of the right measure for link strength calculation is essential. Chapter 10 shows difficulties during link strength interpretation. How spurious links and real links can be separated will be shown, and an auto-adaptive filter method is demonstrated. Chapter 11 introduces two new measures: the *representation index*, and the *time-resolved relevance index*. Both can be used to classify network resources and to distinguish local and global relevance. Dynamically changing network structures can be found in many functional networks. One needs a method to quantify the underlying dynamics. Based on the idea of a force directed layout, which was initially developed to create natural and aesthetic representations of networks, a novel concept, called *structure induced stress* is introduced in chapter 12.

Part III - *Analysis results from several sub-projects are combined in part III.*: Chapter 13 shows a characterization of the overall editorial and knowledge consumption processes in Wikipedia. Chapter 14 gives a conclusion and provides an outlook into the next steps of my research.

Part IV - *Additional results and published work is presented in part IV.*: A purely data-driven market study, based on Wikipedia pages in multiple languages is presented in Chapter 15. Here, we compare the two approaches based on Wikipedia data (completely open) and Google Trends data (free, but not open). The new methods developed in this work are finally applied to data from financial markets. Three previously published articles are presented in part IV. Initial results of our Wikipedia based study of co-evolving networks are reproduced in chapter 16. Chapter 17 shows our approach to study stationarity of an agent based evacuation simulation. A comparison of an agent based traffic simulation and real traffic data shows a difference, especially regarding inherent correlation properties. Chapter 18 closes this work with a reproduction of an article about cross-correlation behavior in traffic data. We studied traffic phases based on correlation properties of multiple metrics collected on a highway near Madrid.

Frankleben, December the 18th, 2016

Mirko Kämpf

Part I.

Introduction

2. Complex Systems

> Reductionism, as a paradigm, is expired, and complexity, as a field, is tired. Data-based mathematical models of complex systems are offering a fresh perspective, rapidly developing into a new discipline: network science.
>
> *(Albert-László Barabási, Nature Physics, **8**, 2012)*

Complex systems are focused by many interdisciplinary research projects. Initially this branch of research was called *Systems Theory*. Later, the field evolved into *Cybernetics*, which finally is called *Complex Systems Research* (CSR). CSR is much more than a theory and has applications in many different fields. According to Newman [14] *"A complex system is a system composed of many interacting parts, often called agents, which displays collective behavior that does not follow trivially from the behaviors of the individual parts."*

These individual parts exist on several scales. Microscopic elements, mesoscopic components, and also macroscopic sub-systems can be parts of a complex system, but they can also show complexity as an inherent property on their own. This is why abstraction is required to represent them in a meaningful way. Complex systems are very often represented as networks. A network consists of nodes and links. Nodes represent the objects, and links represent the relations between objects, which can be either just conceptual links or real interactions. Alternatively, the term *graph* is preferably used in the mathematical context. In principle it has the same meaning or expressiveness like the term *network*. The graph consists of *vertices* (nodes) and *edges* (links).

The following section introduces a philosophical principal called *mechanistic mindset* - which precedes the era of CSR - in order to connect both. Important high-level concepts relevant for CSR are listed to summarize typical properties together with examples from recent research projects.

According to Pietschmann [15] a dominant philosophical principle or paradigm is the mechanistic mindset (original: "mechanistisches Denken der Neuzeit") which is based on four components:

(1) - *"Everything which can be measured should be measured."* The philosopher, mathematician, physicist, and astronomer Galileo Galilei (1564-1642) influenced the evolution in science, especially in natural science during the first half of the 17-th century.

(2) - *"Everything can be decomposed in smaller sub components."* The theory and publications of René Descartes (1596-1650) build the base for this principle.

(3) - *"Either ... or."* Although Aristoteles, had been one of the most important and most influencing philosophers, who lived hundreds of years before Galilei and Descartes (384 B.C. to 322 B.C.), his thinking influenced our culture after the year 1200 because his work and wisdom was brought to Europe at this time together with mathematical and numerical concepts from the Middle East.

(4) - *"Cause and effect."* This fourth cornerstone of a mechanistical mindset is based on work of Isaac Newton (1643-1727).

Pietschmann says: *"one should differentiate without separation."* In this way he addresses the methods for future research. Reality should be represented and studied in a way that does not affect the outcome because of negative influences of the research method on the initial system. Complex systems analysis in general requires detailed investigation of individual elements in the presence of their real context (without separation).

One goal of this work is, to develop an analysis procedure which does not require a complete isolation of elements but allows embedding and contextualization of data. Data-driven contextualization is a kind of normalization of measured data in a temporary and spatially limited context. We apply relative measurement procedures which have been found to be reasonable methods and in agreement with these recommendations.

2.1. What are Complex Systems?

According to the online etymology dictionary the term *complex* means *composed of parts*, derived from the French *complexe*, which means "complicated, complex, intricate" (17-th century), from the Latin *complexus* "surrounding, encompassing". The meaning "not easily analyzed" was first recorded 1715. This shows that the concept of complex systems is not new, but the scientific field related to complex systems is pretty young.

Luis M. Rocha published an article called "*Complex systems modeling: using metaphors from nature in simulation and scientific models*" [16]. This document provides the following three definitions for complex systems:

(1) - *[Advances in Complex Systems Journal]*: A complex system is *"a system comprised of a (usually large) number of (usually strongly) interacting entities, processes, or agents where understanding it requires the development, or the use of, new scientific tools, nonlinear models, out-of equilibrium descriptions and computer simulations."*

(2) - *[Herbert Simon]*: A complex system is *"a system that can be analyzed and decomposed into many components having relatively many relations among them, so that the behavior of each component depends on the behavior of others."*

(3) - *[Jerome Singer]*: A complex system is *"a system that involves numerous interacting agents whose aggregate behaviors are to be understood. Such aggregate activity is nonlinear, hence it cannot simply be derived from summation of individual components behavior."*

In general, one can use the following properties to verify if a system can or should be called complex - since those properties are common and found in many complex systems they can be seen as a reference. They are: (a) presence of feedback loops, (b) spontaneous ordering and emergent organization (self-organization), (c) structural stability and robustness, (d) hierarchical organization, and (e) non-linear dynamics.

It is not required to find all of those properties in a system, some are obviously visible, others can be hidden and require special treatment to be identified. This is the purpose of complex systems research. Based on these principles it is possible to describe real world systems. Rocha, e.g., describes methods for modeling and simulation which allow much more feasible virtual experiments. These properties are good criteria for a classification of systems in general [16]. Next I explain the properties (a-e) in a more detailed way:

(a) Coexisting directed links lead to *feedback loops* which can be positive (amplifying) or negative (damping). The positive activity level of one element has the potential to influence others and most importantly the future of that particular element. This way also the history is related to the current state of an element.

(b) Interference, or resonance effects are, e.g., related to self-organized structure formation. This way, complex systems exhibit *emergent phenomena*. Emergent behavior is the result of nonlinear overlapping activity of all elements and can not be identified be looking into the disconnected part only.

(c) The multiple ways of coupling (especially the strong links and short distances) lead to *cascading failures* which may have catastrophic consequences on the overall system behavior.

(d) As a result of coupling multiple systems on different scales coexist. *Complex systems are nested* and exhibit a hierarchical structure. Nested complex systems can again be complex and so on. Multi-scale models have been proposed and analyzed to handle large-scale systems during the SOCIONICAL project. During model integration we found, that in many cases the individual low-level interactions can be replaced by simplified relations. This is a reasonable compromise between simplification and keeping the embedding within the original environment.

(e) *Nonlinear interactions* between many components lead to a complex system in which the concept of superposition can not be applied. One consequence is the emergence of specific phenomena, which can not be explained based on the known behavior of individual elements. One has to handle all interacting elements and their interactions, which is usually a very large number.

Furthermore, *complex systems may be open* and consist of a huge number of interacting elements. Open systems are often far away from equilibrium. One can find coexistence of fluctuations and stable patterns.

Due to feedback loops, damped coupling, and delays, *complex systems have a memory*. This means that history and current state of the system as well as history and current state of the environment are important and influence the future behavior of a complex system. Like in magnetic materials hysteresis can be found in many complex systems.

Coupling of complex systems with other complex systems - which both are represented as networks - leads to *networks of networks*. Depending on inherent structural properties of the subsystems the combination leads not necessarily to a superposition of both.

2.2. SOCIONICAL

SOCIONICAL provided the context of this work. It is an international, interdisciplinary research project - funded under European Seventh Framework Program (FP7) - with focus on information and communication technologies.

The project goal was, to develop new methods for complexity science using large- and multi-scale modeling approaches in order to implement simulation and prediction methods for large-scale socio-technical systems.

The following illustration shows the context of this work, especially the SOCIONICAL project, as a multi-layer network.

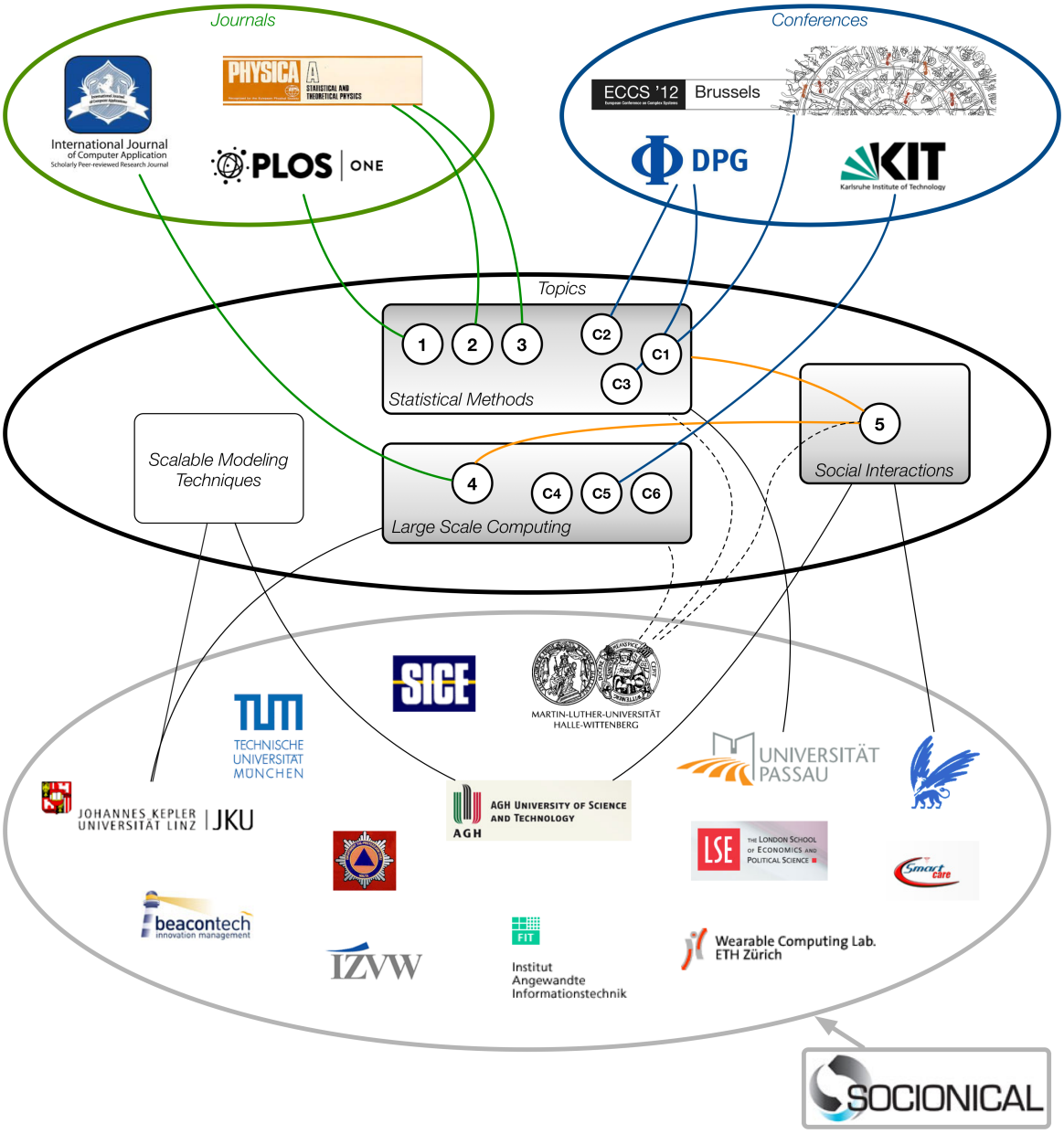

Figure 2.1.: **Real world scenarios lead often to hierarchical networks.** Reduction of complexity is done by focusing on individual groups (ovals). Links exist between elements (of the same type in the same group and of different types from different groups. Also groups can be related to elements and the other way around. Chopping the system into slices leads to information loss. *This example shows a (partial) context network of this work.* Such context networks are also simplifications. Finally, a comparison of results, derived from different approaches allows one to qualify the impact of simplifications during modeling. For illustration purposes I group (a) scientific articles (1,2,3,4,5) and (b) conference contributions (C1,C2,C3,C4,C5,C6) by topic (black border). The topics are related to research groups (black straight line) which collaborate within the SOCIONICAL project (gray border). Finally, the articles were published in different journals (green border) and I attended multiple conferences (blue border). The figure represents only a subset of possible relations to highlight the concept of contextual embedding, thus it is a partial context network.

According to the project website[1]: *SOCIONICAL focuses on the specific example of Ambient Intelligence (AmI) based smart environments. A key component of such environments is the ability to monitor user actions and to adjust its configuration and functionality accordingly. Thus, the system reacts to human behavior while at the same influencing it. This creates a feedback loop and leads to a tight entanglement between the human and*

[1] http://www.socionical.eu/

the technical system. At the same time there is dynamic, heterogeneous human-human, human-technology, and technology-technology communication leading to ad-hoc coupling between components and different feedback loops.

2.3. My Steps Towards Complex System Research

Complex systems research (CSR) builds on well established scientific approaches and combines those with pretty new techniques, including Big Data analysis for advanced analysis of huge data sets from real world, experiments, and numerical simulation. High performance computing (HPC) for large-scale simulations is also of great importance in CSR. My thesis is focused on three example scenarios: (a) interaction of a global user community with Wikipedia, (b) Social force model for simulation of evacuation scenarios, and (c) traffic data collected on a highway and gathered from numerical simulations. We also study global (system wide) properties and emergent phenomena (local properties) that arise in AmI based socio-technical systems. Figure 3.2 illustrates a global feedback loop formed by users interacting with Wikipedia. This work started as part of the SOCIONICAL. Several experiments with real persons were conducted, e.g., the movement of groups of people had been studied based on recorded trajectories. While people moved inside a building, or even in an open space during large public events - such as the Vienna marathon, the Lord Mayor's show in London, and a traditional festival in Malta - their mobile phones were used to record and collect motion profiles of individual persons. Furthermore, data from other real world scenarios, such as traffic flow data from M30 highway in Madrid and specific computational frameworks were combined. In many cases, the integration level we achieved in SOCIONICAL was a pairwise approach, e.g., fluctuation properties in traffic flow data were compared with results from a commercial traffic simulation tool (see [9]) in order to learn more about the consistency of the numerical model and reality. We found that fluctuation properties were not fully present in simulated data. This means, that if fluctuations are the cause of an emergent phenomenon in a complex system, one may miss this particular phenomenon in the simulation in which our data was generated.

Next, we developed and studied an agent based evacuation-scenario with interacting individuals on top of a complex geometry. The motion of agents was calculated, not measured as in the examples before. We implemented persons as cellular automata and combined this technique with another numerical simulation based on Helbing's *social force model* [17]. Return interval statistics reveals that such an evacuation process is non-stationary [13]. This result has an impact on a second type of simulation, on a different scale. While the per-room evacuation was investigated, we found appropriate properties to control a simulation covering many rooms connected to floors, which finally form a complex geometry. Initially, in the second evacuation simulation we placed new agents at a constant rate to the simulation grid, but according to the previous results, we knew that a time-dependent distribution represents reality better. In both cases the analysis results told us more about dynamic parameters per subsystem which allowed a multi-scale model integration at the end.

During the project I found several very inspiring books and articles related to complex systems research but also applicable in interdisciplinary studies. An overwhelming amount of scientific publications exists, also many books about complexity were published. Obviously, there is not one single book which can teach all relevant topics. From within my personal perspective I want to highlight the book "*Neuland des Denkens*" by Frederick Vester (published in 1998). It was the most inspiring book for me so far and woke my interest in networks, at a time before I could use the emerging Internet on my own. Second, the book "*Fraktale und Finanzen: Märkte zwischen Risiko, Rendite und Ruin*" published in 2005 by B. B. Mandelbrot and R. L. Hudson illustrated impressively the relation of mathematical theory, statistical concepts, applied science, economy, and physics. Covering fractals and time series properties in theory and in applications, this book initially guided me to the topic of time series analysis. Furthermore, we needed to learn more about the concepts of network theory. Those are well presented in the book "*The Structure of Complex Networks: Theory and Applications*" by E. Estrada (published in 2011).

Beside ideas and theories, data and tools are very important in CSR. In order to have a quick start into the development of new computational methods it is time saving to build on top of or to adopt existing open source software. In this work we used Gephi [18] for network analysis and visualization and the simulation toolbox NetLogo [19]. Both are written in Java and highly customizable and very helpful during development of simulation software. Self made analysis software was written in Java and it turned out that existing data analysis software like R [20] and the commercial product Matlab [21] were worth the time for learning their syntax. But, on the long run, a scalable data management platform - in this case Apache Hadoop combined with Apache HBase - was crucial in order to manage the growing complexity in our methodology. It was not just the growing volume of the data sets but also the variety of parameters and algorithms which had to be applied in a consistent way.

This chapter will be closed with important examples of complex systems and network science.

2.4. Examples

Physiological Networks: Bashan *et al.* [22] describe the human organism as an integrated network of complex physiological (sub)systems which all have individual regulatory mechanisms. They all interact continuously, and a

failure of one system can trigger a breakdown of the entire network - which leads potentially to death. The authors have developed a method for probing interactions among multiple physiological subsystems and represent this as a physiological network. They found physiological states, which are characterized by a specific network structure. This indicates a robust interplay between network topology and function.

International Stock and Currency Markets: Our global economy, and especially financial markets are created by interacting people, representing institutions or even whole countries. All those different participants in the market contribute and consume. In general there are overlapping interests. The interest of different actors can also be seen as competition. The relation of one participant to multiple subsystems can finally lead to complex structures and cyclic dependencies. All activity in such markets is based on information. Thus, all participants consume information, and produce such. Already the growing demand for information about an asset is interpreted as a precursor of increased demand of the particular asset. Consumption of information generates second order data, the information about information. For example, Google Trends data is based on the frequency of search terms, which represent interest in a particular topic. Wikipedia click count statistics also provide this kind of information - but in a slightly different shape. Both data sources are used in this work.

Information flow via news channels is established by multiple technologies. However, all have one thing in common - each transmission leads to a delay and this has an effect on the trading activity in financial markets - no matter if initiated by a person or fully automatically.

Kenett *et al.* [23] demonstrate how international stock markets depend on each other via dependency networks, which were derived from market price time series. It is clear that the interdependence of complex systems leads to even more complex systems. Individual simplification of each system would lead to simpler models, but one has to expect information loss. In order to overcome this limitation it is important to develop methods, which take the coupling between such diverse systems into account.

R. Smith [24] gives another example, related to economy. Visualization of the spread of credit crisis using the example of the US equity markets shows a cascade or epidemic flow like model along the stock correlations.

Climate Networks: This work was also influenced by climate research. We use results of recent climate studies, because the identification of relevant links and the interpretation of calculated relations between nodes in a meaningful way need special methods. Donges *et al.* [25, 26, 27] describe the procedure of network creation based on a fixed grid of points, placed on the surface of the globe on multiple layers in different heights. Palus *et al.* [28] use a fixed threshold for filtering significant links (see chapter 10), and Berezin *et al.* [29] demonstrates how to modify the link creation procedure in order to have more reliable network representations for stability analysis.

Citation and Collaboration Networks: Scientific collaboration networks illustrate the relations among scientists. If affiliation data can be incorporated, it is possible to describe relations between institutions. If data enrichment is continued, one can also identify relations between nations, in the context of scientific research. The measured information is the number of collaborations between pairs of persons, and the result describes the structure of a pretty complex system, which consists of many institutions, embedded into multiple social, cultural, economic, and political relations. The same approach is also applied to *flavor network*[2] in order to study the principles of food pairing and to *human disease networks*.

Further examples for application of correlation based network construction are earthquake studies [30], human mobility networks, e.g., based on available airline connections or daily commute traffic via train, bus, or car [31].

There are many more examples which emphasize the relevance of network science. Havlin *et al.* [32] go one step further by showing the demand for a revolution in network science which should lead to a significant increase of our understanding of social infrastructures and of interdependent natural systems. They list eight opportunities in the context of current challenges and possible applications. According to Havlin *et al.* one fundamental aspect is the combination of methods from complex network theory and the proposed theory of coupled and interdependent networks. The variety of different examples highlights the versatility of network analysis as a generic scientific method.

[2] A flavor network is constructed from a set of ingredients for cooking. Two ingredients are linked if they both appear in a recipe. By analyzing many typical recipes one can compare different regions based on typical food.

3. Network Theory

> Although cascading failures may appear random and unpredictable, they follow reproducible laws that can be quantified and even predicted using the tools of network science. First, to avoid damaging cascades, we must understand the structure of the network on which the cascade propagates. Second, we must be able to model the dynamical processes taking place on these networks, like the flow of electricity. Finally, we need to uncover how the interplay between the network structure and dynamics affects the robustness of the whole system.
>
> *(Albert-László Barabási, Network Science, 2016)*

Networks are formed by many individual - usually distinguishable - elements (called nodes) including their relationships with each other (called links). Networks allow calculation of topological properties which describe the overall system. Using simple *averaging procedures* like, e.g., the average value of temperature, or average weight the system structure and existing interactions between elements are not taken into account. Network measures on the other hand provide values for comparison of networks since the structure is reflected in the result. This differentiates network theory from statistical thermodynamics, where the individual element to element interactions are not handled individually, but rather by using a mean field approach, or based on potentials and effective potentials.

3.1. Overview

Properties of objects of different size - ranging from atoms, cells, animals, human beings with manifold social and technological interactions, to countries or even planets and other objects in space - can be measured and represented as random variables X and Y, and in case of time-dependent properties as $X(t)$ and $Y(t)$. A relationship between such a pair of object properties may exist and be obviously visible or even hidden. An obvious link is also called *implicit relation* or *structural link*. Usually it is measurable or detectable as a property of one object only, but it can also depend on combined information from both objects. Hidden relations can be calculated from pairs or tuples of properties. For this purpose, appropriate distance or similarity measures are required as explained in more detail in chapter 9.2. The objects are simply called *node* or *vertex* and relations are called *link* or *edge*. Links have their own identity (at least for computation and storing the data). Even if the link is not a real physical entity we use another variable $Z(t)$ to represent the time-dependent link property as a variable. In many cases we want to describe real world phenomena. Therefore we need multiple different types of interactions or relations between individual nodes or sub components of the systems, formed by groups of nodes. This can be achieved with multiple layers of edges or by multivariate random variables $\boldsymbol{Z}(t)$, where $\boldsymbol{Z}$ is a vector with multiple components z representing multiple links from different layers. Finally, one can differentiate structural and functional links and study how both depend on each other.

Furthermore, we differentiate two modes of network creation: **(1)** - *Reconstruction:* We start with data and extract information to describe node and link properties based on calculations. Creation of functional, dependency and correlation networks from time series are examples.

(2) - *Construction:* We define rules to describe a growth process or a re-wiring procedure which changes properties of an already existing graph in a specific well known way. Network generators, based on theoretical network models or growth models are examples.

3.2. Typical Network Properties

If the directions of links are relevant the graph is denoted *directed*. For example, dependency networks (DN) are directed. Also Bayesian networks, which allow logical reasoning based on probabilities, are directed. They represent a set of random variables together with their conditional dependencies. Directed acyclic graphs (DAG) are a special case of directed graphs without loops. Trees are special graphs without loops or closed triangles, but not necessarily directed graphs. Not only the orientation of links is used to classify graphs and networks. Some very important graph properties are node degree distribution, link density, diameter, and clustering coefficients. Fundamental network properties are defined as follows:

Size and density: The network size is given by the number of nodes z_n or N and the number of links z_l or L. k_in is the number of links which end on a node. The number of links starting at a node is k_out. The total number of links of a particular node is $k_\mathrm{directed} = k_\mathrm{in} + k_\mathrm{out}$ in a directed network and $k_\mathrm{nondirected} = (k_\mathrm{in} + k_\mathrm{out})/2$ in case of a non-directed network. The maximum number of possible links in a network is $z_{\mathrm{l}_\mathrm{max}} = z_\mathrm{n} \cdot (z_\mathrm{n} - 1)$ without self links or simply $z_{\mathrm{l}_\mathrm{max}} = z_\mathrm{n}^2$ if self loops are allowed.

The network density ρ is defined as ratio of existing links and all possible links $\rho = z_\mathrm{l}/z_{\mathrm{l}_\mathrm{max}}$ where $z_{\mathrm{l}_\mathrm{max}}$ depends on restrictions like symmetry and possibility of self loops.

Node degree k and degree distribution $P(k)$: The probability distribution function for node degree $P(k)$ gives the probability to find a node with degree k in the considered graph.

Clustering coefficients can be calculated for each node. The global clustering coefficient describes the full network. In order to calculate the global clustering coefficient one has to calculate the ratio of all linked node triples z_triple (sets of three nodes with two links and no self link) and all completely linked triangles z_triangle to get: $C_\mathrm{global} = \frac{3 \cdot z_\mathrm{triangle}}{z_\mathrm{triple}}$.

Watts and Strogatz [33] defined a local clustering coefficient by simply dividing the number of existing links in the neighborhood around a node i by the maximal possible number of links in this subgraph to get: $C_i = \frac{2n}{k_i(k_i-1)}$, with i as node index, for the undirected graph. The result has to be multiplied by 2 for directed graphs. Small-World networks usually have a high clustering coefficient, while random graphs show small clustering coefficients.

Community structure can be revealed in many ways. One approach is based on random walks on networks [34]. An information theoretical approach was related to the previously mentioned method of random walks [35]. Hierarchical block clustering [36], minimum cut approach [37], and statistical inferences are possible alternatives. I recommend using Wikipedia as a starting point to explore this broad field. A massive amount of literature is available, and common text books about networks cover this topic extensively.

Figure 3.1 illustrates structural differences of (a) a random graph, also known as Erdös-Renyi network, (b) a scale-free network (a special example is the Barabasi-Albert network), and (c) a Watts-Strogatz network (which is an example of a small-world network)[1] using a radial axis layout[2] to emphasize structural differences visually.

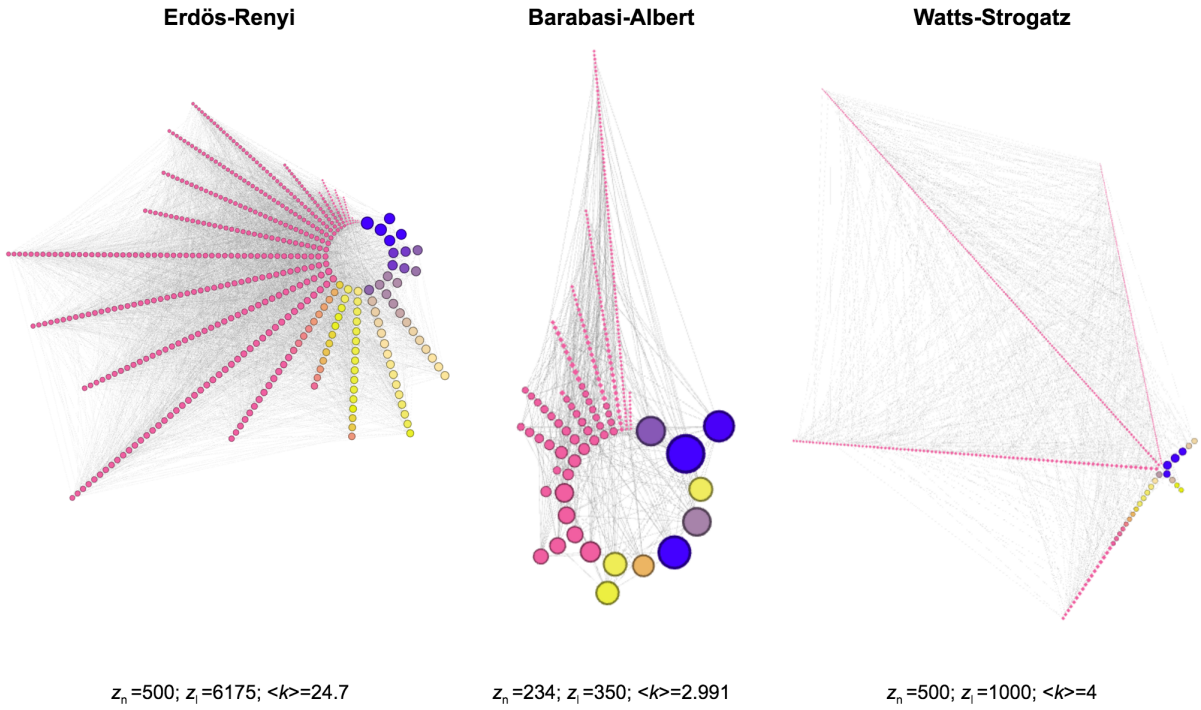

Figure 3.1.: **Comparison of network models**. Qualitative comparison of Erdös-Renyi, Barabasi-Albert, and Watts-Strogatz networks for small graphs (rendered with Gephi using a *radial axis layout* [18]). This layout emphasizes structural properties. Nodes are grouped by *degree* k which defines the position on the circle, and ordered by *betweenness centrality* in a clockwise way. Also, color-coding represents betweenness centrality [38] here.

Random Graphs: A random graph is one in which all possible links $L_\mathrm{max} = N(N-1)$ (or $L_\mathrm{max} = N^2$ in case of self loops) appear with a given probability p. Thus, the parameters p and N define one instance of a random graph belonging to an ensemble of such random graphs. The model $G_{\mathrm{N,p}}$ is called Gilbert model with

[1] Also other networks such as the Barabasi-Albert network can show the small-world property.

[2] Nodes are grouped using an attribute or a metric (degree, betweenness centrality) for radial axis layout. The groups are drawn on axes radiating outwards from a central circle.

link probability p (which equals graph density ρ in this case), N as the number of nodes, and a binomial degree distribution $P(k)$.

According to Dorogovstev [39] one can distinguish graphs in equilibrium and such, which are not in equilibrium. As long as the statistical weights of the system evolve over time and as long as new network nodes are added, the system is far from equilibrium, such as the growing Wikipedia network (see figure 8.2). Wikipedia, e.g., is characterized by non stationary user activity and continuous growth at variable growth rates. Even a network with a fixed number of nodes can be far from equilibrium, because link or node property modifications can lead to a change of inherent statistical weights.

The Erdös-Renyi random graph also labeled with $G_{\mathrm{L,N}}$ is a statistical ensemble whose members are all possible labeled graphs for a given number of nodes N and number of links L where all these members have equal statistical weight (see [39] p. 9f). All the members of this ensemble satisfy the same restrictions and appear with the same probability. The concepts *canonical ensemble* (with a fixed number of links) and *grand canonical ensemble* (in which the chemical potential is fixed) allow a comparison of graph models with methods from statistical mechanics. The Gilbert model and the ER-model converge for large $N \to \infty$. A classical random graph is a maximally random network with average degree $< k >= p(N-1)$ resulting from link probability p and size N. For small random graphs $P(k)$ is a Poisson distribution.

Detailed structural investigations are possible already with the degree distribution function $P(k)$. The degree distribution can easily be obtained by counting the number of edges for each node - as long as those edges are available in the data set - without any special graph analysis algorithm.

Diameter: The *diameter* d of a network is the length of the longest shortest path between two network nodes in all possible node pairs (n_i, n_j): $d = max(dist_{\mathrm{shortest}}(n_i, n_j))$ with $dist_{\mathrm{shortest}}(n_i, n_j)$ as graph distance, which represents the number of steps one has to go on the network in order to reach n_j starting in n_i. A calculation of the graph diameter, on the other hand, is pretty expensive for very large graphs. An estimation technique called *PseudoDiameter* is implemented by the Wolfram GraphUtilities [40]. Kang *et al.* [41] invented HADI, an algorithm for fast diameter estimation and data mining on large graphs using Hadoop. This allowed them to analyze the largest public web graph ever analyzed - the Yahoo web graph. This graph had 1.4 billion nodes and 6.6 billion edges, spanning hundreds of Gigabytes in 2008. They found, that the previous estimation of the diameter of the WWW by Albert *et al.* [42] was over-pessimistic. According to Albert *et al.* [42] it is assumed that the diameter of the web grows as log N. With HADI Kang *et al.* found a much smaller size of 15.64 as opposed to 19.2, which was predicted by Albert *et al.* previously.

Several categories or classes of networks can be distinguished according to their topology. Depending on the **characteristic topology** which means depending on **structural properties** one can distinguish random graphs, scale-free networks, and small world networks (see figure 3.1).

3.2.1. Scale-Free Networks and Small Worlds

To explain the *scale-free property* of a network I start with a definition of the scale of a random network. If $P(k)$ follows a Gaussian distribution defined by μ_k and σ_k, the degree of a randomly chosen node is k and typically close to N and all moments of $P(k)$ are well defined. For a Poisson distribution as found in typical random graphs $\sigma_k = \sqrt{< k >}$, and this is why the degrees are also bound to a limited range around $< k >$. In case of a Power law distribution $P(k) \sim k^{-\gamma}$ the second moment diverges if $N \to \infty$ and σ_k is not defined. This means that no typical scale or no typical size in terms of average node degree exists for this network. In many real world networks the exponent γ is between 2 and 3 (see table 4.1 in [43]). A consequence of this power law distribution is the presence of a few dominating hubs. These hubs are nodes with a very high degree. Hubs are heavily connected, but the majority of nodes in a scale-free network are not well connected.

If nodes are heavily connected within a given local neighborhood, we find a high local clustering coefficient. In this case, many nodes can be reached with just a small number of steps, and the network is robust, because of many redundant links. Many hubs means, there is not one central node but a decentralized network structure exists. What happens if such local hubs are connected to each other? This process leads to another important property of real world networks which is called *small world phenomenon.* Connections between hubs are global links, typically called long-distance links. They act as a kind of shortcut and allow pretty short paths between nodes, which are physically far away from each other. According to Stanley Milgram [44], the average shortest path between any two people on the globe is approximately 6, this so called small world phenomenon is also known as "*six degrees of separation*". The Watts and Strogatz model [33] (see figure 3.1) allows one to describe and construct this phenomenon in graphs with only local links. They start with a ring lattice and rewire randomly chosen nodes. As a consequence the shortest path length drops already after a few steps of replacing local links by long distance links. Recent research results reported by Backstrom *et al.* [45] show that in Facebook's social graph the average distance is less than 4, even shorter than found by Milgram. In general, a classical random graph has a small number of closed triangles, but it can show the small world phenomenon as well (see Dorogovstev [39] p.18).

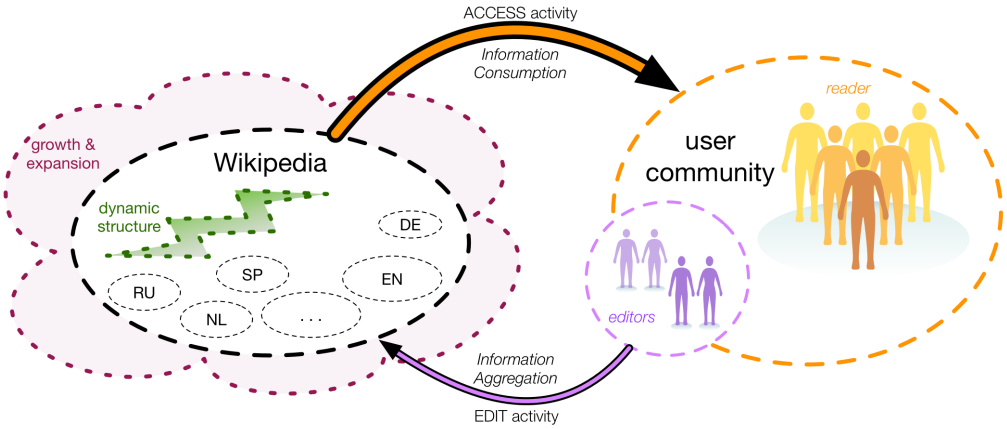

Figure 3.2.: **Networks as representation of complex systems**. The Wikipedia network consists of interlinked content - the Wikipages - and the social network of users. Users contribute to the system with content creation and content improvements. The majority of users only consume this free public content. Thus, user communities are asynchronously connected to the content they share. Such an interconnected system can be seen as a global knowledge base, or as a stub for public interest in multiple topics, provided in multiple languages. The Wikipedia system works in general like an associative memory.

3.2.2. Bipartite and k-partite Graphs

Traditionally, a graph consists of one type of nodes and edges, which represent node and link properties at a given point in time or within a time range during which the system is not changing. Multiple facets of a system are modeled in multiple layers by multiple sets of links. But sometimes, time-dependent properties and also multiple node and link types are required - or even combinations of those (see figure 3.5).

For example, the modification of Wikipedia pages or collaboration on research articles can be modeled as networks and lead to implicit social networks among the collaborating persons. Such networks consist of two node types, people and resources, and thus they are called bipartite (see figure 3.2 and figure 3.3).

Bercovitz [46] concludes, that the bipartite network structure contains useful information on the implicit social network, even if this cannot be seen directly, since social relations are not explicitly provided.

For example, simultaneous activity on multiple resources can be interpreted as an implicit social link between collaborating people. Similar to this, a correlation network expresses implicit relations between resources, based on similar activity patterns caused by coincidence of interactions. In this way a dynamic view of the system can be generated.

Network nodes can be partitioned into k disjoint sets so that no two nodes within the same set are linked to each other (although there are no isolated clusters) form a k-partite network (see fig. 3.2 for illustration). Links always connect two nodes of different types or from different sets. In the special case of a bipartite network only two different types of nodes are available. This kind of assignment of all nodes to disjoint sets is also called partitioning[3].

In order to compare networks of different types, for example a bipartite network and a graph with only one type of nodes, a technique called *bipartite projection* is used (see figure 3.3). This allows also aggregation of the resulting graphs since they are of one final node type now. Beside this, the amount of data is reduced, but if weights are ignored, information can be lost. It is not possible to do the inverse transformation of a bipartite projection.

So far all discussed network types represent only snapshots of a system at a given point in time. But time-dependent networks are very important in order to study dynamics of non stationary open systems.

[3]The term partition defines a group of nodes with common properties (at least one).
Multi-partite networks can be defined based on obvious or derived node properties. Clustering algorithms identify groups of common nodes, according to the structure of a network. Links between those clusters can be separated from intra-cluster links. It is important to consider, that the term is used in multiple ways. A partition is also a set of records in a database table - even if the records represent different node types - the partition strategy inside the database may differ from graph partitions.

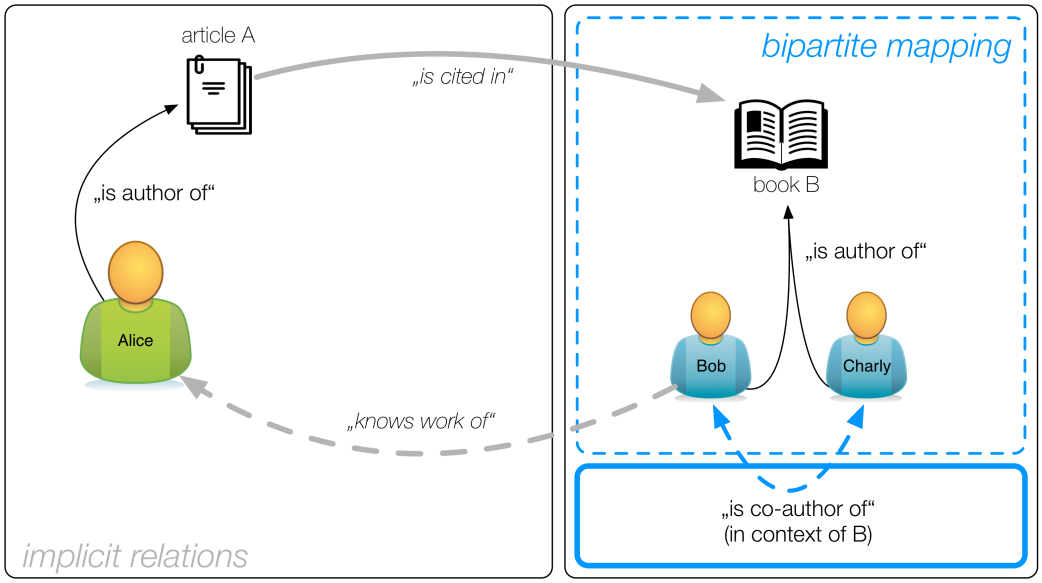

Figure 3.3.: **Complex networks can be formed by implicit and explicit relations.**. Authorship networks are bipartite networks. If an article A is cited in a book B, one assumes, that authors of B know the content of A. This assumption is expressed as implicit link (see dashed gray link in left box). Citation and co-authorship networks are unipartite networks. They contain only nodes of one type. Bipartite mappings allow a transformation of bipartite networks by replacing one node type and its associated links (book and authorship) by a new link type (*'is co-author of'*) shown as a dashed blue arrow in the right box.

3.3. Obvious and Hidden Links in Networks

Links can exist for a very long time, for ever, or just sporadically, depending on the use case. For example, chemical reactions and social interactions within society can be described by networks. The concepts are similar but the time scales are very different. Temporal links can be special types, such as dependency links which are the result of processes on top of the network, e.g., communication processes. The link only exists as long as communication lasts or as long entities are in an effective range of each other. Temporal links can even exist between nodes which are not connected by a static network and because of this, such temporal links are sometimes hidden links and not accessible directly.

Structural links between nodes represent obvious or well known relationships between them. Different link types express different types of relations. One has to distinguish between directed and non directed relations. A directed relation can be a property which is only visible in the context of one node. Starting from one node, one can ask: "What are the node's children?" This kind of relation implies another inverse relation such as: "What are the child's parents?" Directed links can be of a specific category and they can have a weight too. Link types can be manifold. The more types or the more possible values one has to distinguish the more complicated and less expressive the analysis procedures become. For practical reasons, one can replace two of such directed links - if they are the inverse of each other - by one non directed link as illustrated in figure 3.3 (implicit relations). In other cases a link only exists if the weight exceeds a defined threshold.

Links between web pages, or in our special case between Wikipedia articles, are easy to detect and to extract. But not all structural links are obviously visible. A family tree is an example of a structural network. Due to limitations of data availability and access limitations it is often hard to create a comprehensive family tree which spans several centuries. Friendship and follower networks are modern examples of structural networks. In many social networks, the structure is not actively maintained. This means, links are created over time and they can disappear as well. The network grows but as soon as some system properties change, the network is not updated immediately. Due to this fact, the network is not well aligned to reality. Such a network is not a valid representation of reality any more. E.g., friendship emerges, it stays for a certain time but friendship can also end. One can clearly see a *network life cycle* which spans a certain time range. Within this range, friendship networks are considered to be structural networks.

One has to distinguish quasi-static structural networks from functional networks, although they may change over time as a consequence of structural changes. Functional networks usually overlay structural networks or they can be caused by processes on such structural networks. Functional links are not defined directly like in the case of friendship or parent child relation. It is more of an indirect relation like co-occurrence, or co-location of elements. Both are consequences of activity. In case of co-occurrence of extreme events in two time series, one can conclude

that a relation between both elements which causes the extreme event might exist. One has to be careful, there is no easy way to prove that the relation is not a random correlation. On the over hand, co-location can cause or be interpreted as a structural link. A chemical reaction can be used to explain this. Only if the atoms or molecules are within a certain range in space at a given time and other appropriate conditions are met, a reaction can happen. Before the reaction, reactants are individual entities. During the reaction they are co-located and also part of a functional network. After the reaction a new molecule exists, which means that a structural link between two formerly disconnected entities has emerged. A functional network also can be reconstructed from social media data. If people work together over a certain period of time than their collaboration can be interpreted as a functional network link.

This can lead to friendship and other social relations. On the other hand, friendship and the implicated trust between friends can be the reason for deep successful collaboration. This way, network structures may lead to memory effects which can be identified as long-term correlations in time series data.

Correlation links are used to describe a symmetric relation between two nodes. Causation networks are directed and their dependency links are used to express the influence one node has on others.

A recent example of such a dependency measure is called *DebtRank*. It was introduced by Battiston *et al.* [47] with the goal to determine the systemically important nodes in a network. Analysis of systemic risk in financial systems is essential in order to understand critical situations or trends towards critical destructive eruptions like the global financial crisis in 2008. They show, that not only the size of a system component but even more the centrality in the network matters. They suggest that the debate on too-big-to-fail institutions should include the even more serious issue of too-central-to-fail. If this happens than a single measure is replaced by a structural measures derived from the entire system instead of just one node.

Similarity and distance measures are used to compute *functional links*. Functional links are often the consequence of hidden processes or collective activity of or on network links and nodes[4].

A *structural link* is typically an obvious link, but as explained before, the data might not easily be accessible.

Structural networks can be directed or non directed. Correlation networks are usually non directed networks as they are calculated from symmetric functions. In section 5.4.3 I describe event-synchronization which allows calculation of a directed network from activity time series in order to overcome the limitations of the cross-correlation based approach. More specific types of functional networks exist, e.g., climate networks reconstructed from mutual information (see Donges *et al.* [26]) or dependency networks reconstructed from time series triples also using the concept of mutual information (see Kenett *et al.* [23]).

3.3.1. Highly Dynamical Networks: Temporal Networks

All previously mentioned networks were modeled following a connectivity driven approach. Alternatively, networks can be created in an activity driven way. According to Perra *et al.* [48] *"network modeling plays a critical role in identifying statistical regularities and structural principles common to many systems. The structural patterns of the network are at the basis of the mechanisms ruling the network formation. Because connectivity driven models necessarily provide a time-constant or time-aggregated representation, they may fail to describe the instantaneous and fluctuating dynamics of many real world systems."* They address this challenge by defining the *activity potential*. An activity potential is a time invariant function characterizing the node's interactions. They construct an activity driven model capable of encoding the instantaneous time description of the network dynamics. Especially if network formation or transformation processes are of interest, this model has an advantage as it is able to explain structural features such as the presence of hubs, which simply originate from the heterogeneous activity of agents. Highly dynamical networks can even be described analytically. This allows to overcome the limitations of time aggregated data.

Petter Holme shows in his colloquium about temporal networks [49] many examples of temporal networks: (a) *human and animal proximity networks*, (b) *citation and collaboration networks*, (c) *economic and ecological networks*, and (d) *distributed computing*. For example, proximity networks can be obtained by mobile communication devices which record the position via GPS as a function of time. A practical application is the optimization of treatment in health care. A study by Liljeros *et al.* [50] investigated the temporal network of 295, 108 Swedish patients over two years. In order to optimize the system one has to reorganize processes in the hospital in such a way that leads to a more static network by reducing the temporal network. More examples are traffic and transportation networks, brain networks, and networks in complex materials. Casteigts *et al.* [51] describe dynamic networks as *time-varying graphs*.

The article *"Walking and searching on time-varying-networks"* (see Perra *et al.* [52]) is another example which illustrates the importance of new approaches in network science which go beyond static network layers and simple topology.

Already in 2003, Ch. Ivanov Plamen called a "dynamics network" *the network formed by many individual nonlinear systems which have their own output and at the same time are interacting with each other* [53]. In

[4]Collective activity on a network means interaction with it by one or many individuals during a given time interval.

order to quantify the interactions he suggested linear measures, such as cross-correlations. In case of nonlinear interactions he suggested usage of synchronization measures.

In this work we use the term *'functional networks'* and calculate metrics for pairwise node properties (from time series data). The metrics can be interpreted as connectivity link between pairs of nodes. In addition, we use time windows to handle multiple time scales. Based on sliding window techniques it is possible to reconstruct activity networks or time-varying networks this way. Similarity measures and concepts from information theory are used to define connections between network nodes even if such links do not exist physically or if they are not visible obviously.

3.4. Networks of Networks

Recently, a new concept, called networks of networks (NoN) was introduced (see [27],[54],[55],[56], and [57]).

In a simple network, nodes are all of the same type. Links between pairs of different nodes exist in bi-partite networks. Now we combine both concepts (see figure 3.5). Nodes from one type (one sub-network) can also be linked to nodes in other sub-networks directly. Directed links between networks can introduce dependencies - especially in case of feedback conditions where dependency links exist between both components in both directions. Such inter-dependencies can lead to cascades of failures as a result (see [55]). Figure 3.5 illustrates also indirect or hidden links, which are introduced by links to a third component. The green subsystem B acts as a connector layer and thus has no internal links. This causes the light gray colored area in the matrix. Real connectors can have an internal structure but this is omitted for simplicity.

Dependency links are different form correlation links. A correlation link describes that the two nodes have something in common but if the one node disappears, the other one is not affected. In case of a dependency link, the dependent node also disappears in case of a node failure. One has to be careful with this terminology, e.g., dependency links are also defined by calculation of mutual correlation (see [23]) but such links do not express any real dependency, but rather an assumed direction of influence between the nodes in cases where the term causation cannot be used in general.

Based on the idea of dependency links one can study the level of interdependence in infrastructures, social networks, and technical systems. Furthermore, NoN-research is related to studies of co-evolution of complex systems [58].

The bidirectional interaction between financial institutions, the financial markets - which are part of a global economy - are illustrated by arrow (A) in figure 3.4. The markets provide information for news agencies (B) which are consumed by people (C). Furthermore, people get information from multiple other sources, such as social

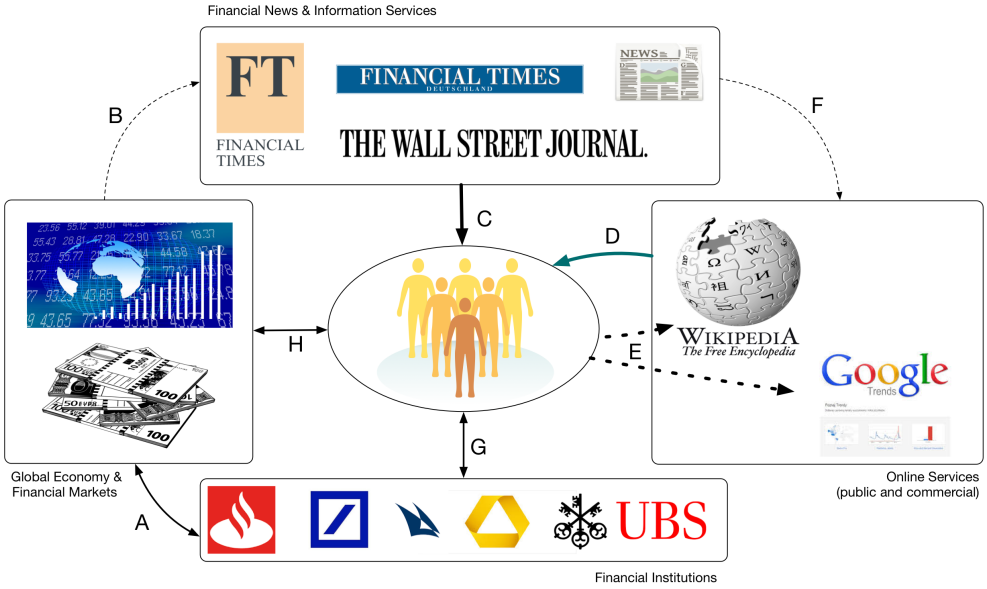

Figure 3.4.: **Networks as representation of complex systems**. Individual aspects of complex systems dynamics require specific representation, such as directed or undirected networks which can exist as static or temporal networks. Interdependence between networks leads to coupled networks in which processes can interfere with each other. Hence, they constitute positive or negative feedback loops. For large-scale systems such as financial markets those effects are well known but often invisible.

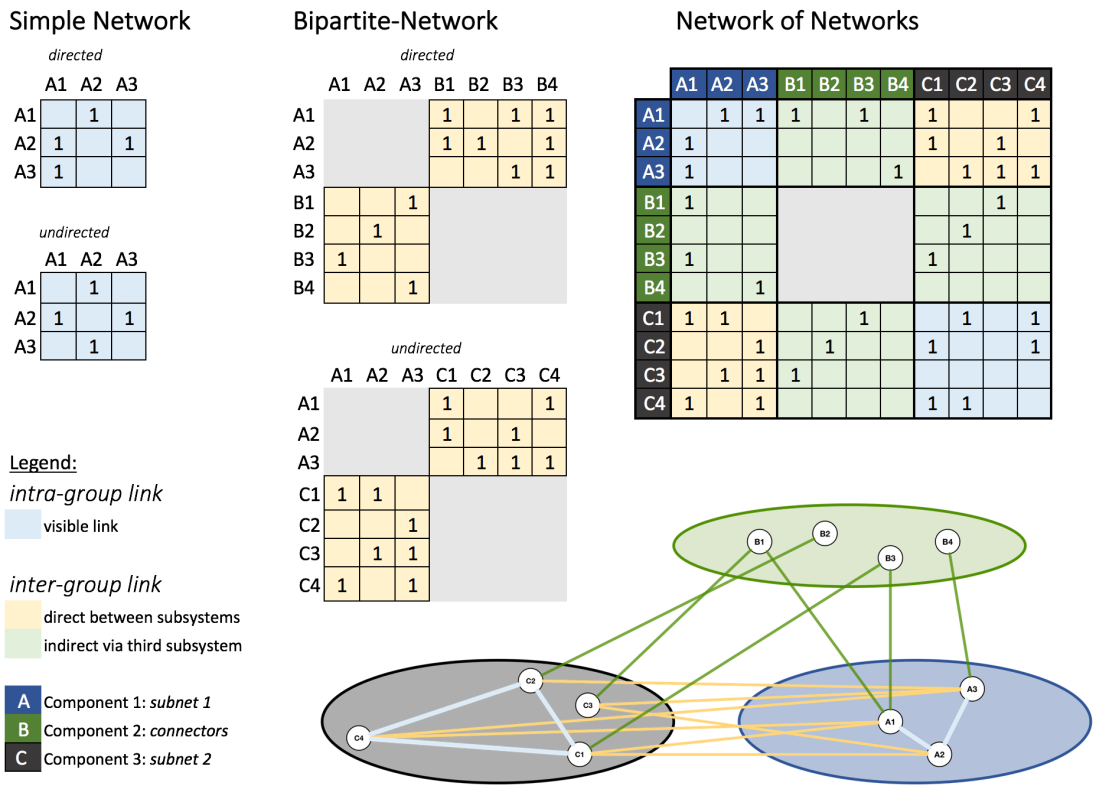

Figure 3.5.: **Networks of Networks (NoN) represent interconnected systems on multiple scales**. NoN consist in general of multiple link and node types. Intra-component links and inter-component links together can lead to feedback conditions. NoN can be composed of multiple existing systems or huge systems can be divided into sub systems in order to re-connect them in a large multi-scale model.

media applications and online services (D). An essential part of social media applications is the participation and contribution by many users (E). Online communities can be influenced by media channels also in a direct way (F). Finally, people interact with the economy in a direct way (H), or via financial institutions (G).

As another example, one can consider the road-network as a first network layer. People which walk along the streets form two more networks: (a) the relation between people is defined by social networks (friendship, family) and (b) there is a communication network, which introduces short-cuts on top of the road network. Even if the real distance between two persons is long and they cannot see each other, they can use mobile communication devices for calling to inform each other about their next movements. This brings both closer together on a logical level. In critical cases, this mobile communication network can cause unusual delays in information flow, because it is overloaded, or it can even collapse completely, and it disappears as a consequence. In other cases, it transfers information between people in real time and allows one to prevent a collapse of the traffic network. In both cases one complete layer of the complex system could be lost, affecting also global properties. As a consequence, properties like the small world property could be lost. This leads to fundamental changes in the dynamic behavior of the entire system and internal structure.

According to Gao *et al.* [54, 59] one has to differentiate two types of dependency links which can be defined by simple conditions, such as: (a) feedback-conditions, and (b) non feedback-conditions. In case of feedback conditions, a failure in one network A can be propagated also via the second network B. If node B_1 depends on A_1, and A_2 depends on B_1, than the node A_2 would be affected by a problem, even if it is not linked directly to A_1.

The advantage of the NoN approach lies in a deeper integration of individual aspects in one large comprehensive model. Even if a system of indifferentiable elements exists in the beginning, one can start with traditional network analysis or k-means clustering to identify clusters of comparable nodes based on topological properties or simply based on node properties and a similarity measure. A segregation of those groups and several link creation strategies as explained in chapter 9 allow a more specific inspection of inherent dynamical properties instead of relying on obvious structural links.

As a practical example Donges *et al.* [27] study the global climate system and model it as a system of interconnected systems or NoN. They combine multiple so far disconnected distinct climatological variables by correlation

links. This way a NoN appears which consists of pairwise interacting sub networks. They do not use the correlation strength as a link weight, but rather a set of new topological measures for local and global properties is introduced. The new local measures are *cross-betweenness centrality*, *cross-clustering coefficient*, and *cross-degree centrality*. The global measures are *cross- edge density*, *cross-transitivity*, and *cross-average path length*.

3.5. Quantitative Analysis for Complex Networks

Interpretation and comparison of network properties and analysis of network dynamics require calculation of characteristic properties as a function of time. Different software packages exist for quantitative graph analysis. Such software packages often have a huge number of algorithms built in, but most of them are limited by available hardware resources and bound to a single workstation. Parallel graph algorithms for large graphs exist as well, but currently, the number of available algorithms is limited.

For this work, we use existing open source graph analysis packages. We identified three appropriate software solutions which provide a rich set of network metrics. They are: (a) Gephi [18], (b) NetworkX [60], and (c) SNAP [61]. Table 3.1 shows four categories of typical network measures used in this work. In addition, Apache Spark comes with the GraphX library [62], which allows large-scale graph processing on Hadoop. Because GraphX provides a BSP[5] based API[6] (compatible to Google's Pregel [63]), more algorithms can be added in the future.

What is an appropriate size of a network for simple analysis? Map-Reduce based algorithms can handle huge graphs without loading all nodes and edges into memory. But even in this case the graph can be too big to be represented or plotted in a meaningful way. Furthermore, iterative algorithms are not suitable for the Hadoop Map-Reduce framework due to the lag of data caching capabilities.

One has to find (a) an appropriate way to scope or reduce the number of nodes and edges, and (b) a representation which emphasizes the aspect one is interested in.

3.5.1. Extraction of Informative Sub Graphs

In case of fully connected networks it is not possible to apply some of the analysis algorithms directly. The degree of each node is exactly the same if link weights are not used. Some algorithms, such as calculation of diameter or betweenness centrality do not take weights into account because only the existence of a link matters in such cases. One has to use modified network measures instead. E.g., a measure called *closeness* - which is the sum of all distances to all other nodes [64] - has been generalized to weighted networks by Newman [65] where Newman used Dijkstra's algorithm [66]. Additionally, Heitzig *et. al.* [67] use also node properties as weights to avoid the bias which may be introduced by heterogeneity which is a consequence of combining multiple types of network nodes.

Using the node and link weights allows corrections according to specific conditions, but in case of a reconstructed network, which usually is created from all pairs or nodes or in other words from all possible links, we need a different approach to transform the fully connected graph into a sparse graph. The obvious approach is filtering by link strength, but depending on the way how those values are calculated, filtering by a fixed threshold is not appropriate (see section 10.4). Instead, the extraction of sub graphs was used in several research projects.

The **Minimum Spanning Tree** (or the maximum spanning tree) provides a sub network of a very specific shape. A tree can not represent the full network, e.g., it has no closed triangles and thus, clustering can not be analyzed in a tree - but one can consider the tree as the backbone of the extracted component. The minimum spanning tree (MST) is a connected, undirected graph extracted from a network of arbitrary structure. The MST connects all nodes with links which lead to a minimal total weighting. One can interpret the link strength as resistance and say, that the MST minimizes the resistance to maximize the flow. If the link strength represents the possible throughput, one would use the inverted normalized link strengths $l_{\text{inv}} = 1 - l_{\text{norm}}$. This is also called the maximum spanning tree. Many different spanning trees are possible in a single graph, but the MST is the one spanning tree with weights less than or equal to the weight of every other. In case of many connected components (which are separated from each other) one gets a minimum spanning forest (MSF). The MSF is a union of all minimum spanning trees. Table 3.2 lists algorithms which were used by Gang-Jin *et al.* [68] to analyze properties of spanning trees extracted from correlation networks.

The **Planar Maximum Filtered Graph** (PMFG) as introduced by Tuminello *et al.* [69] is not restricted to a tree shape - this means also triangles are possible and thus one can calculate clustering of the graph. The method is based on the idea *"that graphs with different degrees of complexity can be constructed by iteratively linking the most strongly connected nodes under the constraint of generating graphs that can be embedded on a surface of a given genus $g = k$* (see also [70]). *The genus is a topologically invariant property of a surface defined as the largest*

[5] BSP: Bulk synchronous parallel. Defines a class of algorithms for graph analysis. The BSP algorithm consists of a sequence of two alternating phases. In phase 1 all nodes can send and receive messages. In phase 2 each node aggregates all received messages and updates its internal state (see also figure 2 in `http://blog.cloudera.com/blog/2014/02/how-to-write-and-run-giraph-jobs-on-hadoop/`).

[6] API: Application programming interface. Defines the routines, input and output types of software and allows implementation of loosely coupled systems.

number of nonisotopic simple closed curves that can be drawn on the surface without separating it." They prove that such graphs have the same hierarchical tree structure associated to the MST but contain a larger amount of information that increases with the genus. This means, that also higher dimensional embedding on hyperbolic surfaces is possible for graphs in order to filter relevant information. Increasing the genus results in increasing the available information. They could show that the relative improvement of the information stored in a graph was highest in case of a planar graph, when the genus assumes the value $k = 0$.

Recently, the **Triangulated Maximally Filtered Graph** (TMFG) was created by Massara *et al.* [71]. They introduced a new filtering method which works well especially for huge graphs. This approximation method allows processing of huge graphs in Big Data environments and extraction of manageable sub graphs, which can be processed in traditional data analysis environments.

3.5.2. Network Measures

The goal of measurements in general is quantification of a specific property. If the same measure can be obtained from two different things and if the type of the measure allows computation of a difference or at least ordering, one can compare the different things with each other regarding this property based on differences and ranks can be assigned. Network measures allow us to quantify network properties and to compare networks. This means we can have two different systems such as network A and network B or even the same system in different states at different times such as $A(t_1)$ and $A(t_2)$. In this way, network dynamics is analyzed. We can identify critical states or transitions in complex systems based on the time evolution of several measures.

The Stanford Network Analysis Platform (SNAP) [61] provides a set of typical network measures as listed in table 3.1. They all are well established and available in a variety of software packages. Others (see table 3.2) are pretty specific and related to particular applications because they were developed to overcome the limitations of previous approaches. In general, quantitative measures are very helpful as soon as standardization and normalization of data allows comparison of different systems in different domains.

Category	**Measure**
simple node and edge statistics	number of nodes and links (z_N, z_L), z_N, z_L in weakly connected component (WCC), z_N, z_L in strongly connected component (SCC)
network size	diameter d, 90-percentile effective diameter d_{90}
clustering	average clustering coefficient $< C_i >$, number of triangles, also number of linked triples z_{triple}, fraction of closed triangles z_{triangle}
community structure	number of communities (number of sub graphs), average community size (simple statistics applied to sub graphs)

Table 3.1.: **Network measures to profile graph data sets using the SNAP software package.** Four groups of network measures are used by Yang and Leskovec [72] to compare properties of 230 networks. Results and raw data are publicly available on the SNAP website [73].

The following graph algorithms or measures are available in Gephi: (1) simple count statistics for nodes and links, calculation of link density; (2) node degree, average degree, average weighted degree; (3) diameter, radius, average path length, (Brandes [38]); (4) modularity, number of communities (Blondel *et al.* [74]); (5) PageRank, (Brin and Page [75]); (6) number of connected components, (Tarjan [76]); (7) average clustering coefficient, number of triangles, (Latapy [77]); (8) eigenvector centrality[7].

Metrics (1), (2), and (7) are also available as dynamic measures in Gephi. GraphX provides implementations for (5), (6), (7), in addition also label propagation for detecting communities in networks [78], and a matrix factorization method called SVD++ [79].

In complex networks - like in other many body systems - one can distinguish group properties from single element properties. Finally, both can be subject of time series analysis. Whenever a property can be measured or calculated as a function of time, we can also calculate the correlation for pairs of such variables which leads to a new network representation for which network properties can be calculated as well, potentially as a function of time, and so on.

Gang-Jin *et al.* [68] analyzed the dynamics of correlation networks from financial markets. They extracted the minimum spanning tree from correlation networks, which were calculated from currency exchange rates. Table 3.2 shows useful network metrics for economic networks and especially MSTs.

[7] A good explanation of the concept can be found in this online tutorial:
`http://djjr-courses.wikidot.com/soc180:eigenvector-centrality`

Name	Application	Definition	Formula
NTL normalized tree length	to analyze the temporal state of the MST, based on its normalized total weight	Length at a given time t when the MST was created, e.g., from cross-correlation of logarithmic returns of stock prices with $d_{i,j}$ as link strength.	$L(t) = \frac{1}{N-1} \sum_{d_{ij}^t \in \mathbf{T}^t} d_{ij}^t$
MOL mean occupation layer	to describe the spread of nodes on the MST and to quantify the changes in the density of the MST	The individual level of a node $\text{lev}(v_c)$ is given by the number of path segments v_c it is away from a chosen root node at time t.	$l(t, v_c) = \frac{1}{N} \sum_{I=1}^{N} \text{lev}(v_i^t)$
(S/M)SSR single- / multi-step survival ratio	to study the short-term stability / robustness and long-term stability of an MST	Survival rate is defined as the fraction of edges found common in two MSTs at different times t_i and t_{i-n} ($n = 1$ for SSSR).	$\sigma_t = \frac{1}{N-1} \lvert E^t \bigcap E^{t-1} \rvert$

Table 3.2.: **Network measures for time-dependent analysis of series of MSTs extracted from correlation networks.** Especially in case of dynamic systems it is important to identify the right properties which change over time and thus allow determining phase transitions of the system. The measures were originally introduced by Onnela *et al.* [80, 81].

It is also possible to handle multiple network layers together in order to calculate a network metric without segregation of the whole system. This takes the fact into account, that nodes can interact in multiple ways via multiple coupling mechanisms. Recently, Halu *et al.* [82] introduced the Multiplex PageRank algorithm to calculate a centrality measure for such nodes in networks with multiple layers. A link-layer represents an individual aspect of a complex system. Multiple link layers form multiplex networks. They are more appropriate in complex systems analysis. Traditionally, network metrics as mentioned in this section can not be applied directly to multi-layer networks. Results of Multiplex PageRank calculation as illustrated in [82] in figure 4 demonstrate how a combination of structural and correlation networks may support a better understanding of time evolution of node properties such as page rank and node centrality.

According to Thelwall *et al.* [83] article relevance measurements derived from social websites are called *altmetrics*. Examples include the well-known *impact factor* and the number of citations of articles in Wikipedia or other SMAs. However, Lozano *et al.* [84] have shown that the relationship between impact factors and a paper's citations get weaker over time. Thelwall *et al.* [83] conclude that metric values for articles published at different times, even within the same year, are often not comparable, and the coverage of investigated altmetrics - except for Twitter - seems to be low. Furthermore, both approaches do not consider the dynamic properties of the document life cycle and usage patterns.

Ciampaglia *et al.* [85] analyzed Wikipedia access time series and introduced a metric called *relative traffic change*. In this way they combine the structural network (to identify a node's neighborhood - see chapter 6 for more details on context networks which were developed in this thesis) with activity time series as measured node properties of Wikipedia pages. They identified a difference in traffic patterns in the presence of link creation and identify two categories of articles: (a) articles which follow the demand ($\Delta V/V < 0$), and (b) articles which precede the demand ($\Delta V/V > 0$). We introduce a comparable analysis method to measure a topic's representation index and to track topic relevance over time in chapter 11.

The next chapter introduces recent research results related to social networks in general. Furthermore it explains the important role of Wikipedia for this work by collecting arguments for calling Wikipedia a complex system.

4. Social Networks

> It is interesting to note how many fundamental terms which social sciences are trying to adopt from physics have as a matter of historical fact originated in the social field.
>
> *(Michael R. Cohen, from: "The Structure of Complex Networks: Theory and Applications" by Ernesto Estrada [86] p.121, 2012)*

A large variety of social network studies was conducted and published during the last decades. Some prominent social networks are the *Zachary's karate club*[1] the *Facebook social graph*, and the *Follower networks* on Twitter and other online communities. Long before the Internet and all its applications existed, the famous Milgram experiment[2] [44] was already an example of a case where the social network was used in research even if it was not accessible directly. Anyway, Milgram could measure path lengths on a social network formed by American citizens by observing the outcome of an experiment conducted on a hidden network structure.

4.1. Application of Methods from Physics in Social Network Research

Typical questions in social network research are: How connected are people and communities? Which short cuts exist in an organization? Can information leak out via hidden connections? Which network structure can improve and which can block efficient information flows? Which elements in the system are the most critical with the most impact on others?

One common approach in social network analysis is based on the structural properties of a system. It is known that network structure can depend on construction methods and data preparation methods. According to Guimera and Sales-Pardo [87] form follows function. In many cases one can find different coexisting and overlapping aspects, which may interfere and thus a *one to one* match between functional networks and the underlying structure can not be expected.

Internet based research relies often on web crawling techniques[3], especially when access to the back end systems is not available. Crawling techniques and extraction of data can introduce strong bias. Thus, the experiment design is crucial, since errors can be introduced on the social and also on the technical level.

Network sizes are in general a limiting factor. Only a few research groups have access to the full data sets like the Facebook social graph or to all Twitter messages. For better efficiency and to enable more researchers to contribute to studies of social aspects in our internet based society one should also focus on methods which work well with sub sets and which can be combined at the end.

Cachia *et al.* [88] discuss the relevance of online social networks (OSN) in general and with focus on future research - which means, prediction of social trends based on properties obtained from social networks is in their focus. They state: OSNs enhance creativity as a result of efficient and easy communication. They also foster collective intelligence by aligning individual thinking towards future goals. Using the inherent information, OSNs can be seen also as expert tools to measure and describe changes in social trends. This goes far beyond the initially mentioned static aspects.

Zhao *et al.* [89] define the entropy of a social network. They analyzed the adaptability of social behavior by comparing dynamical real networks with existing models of social interactions. In this way they found a variable entropy depending on the time of the day during a typical week day. The entropy of a social network is a measure of information encoded in the network's dynamic. Anand and Bianconi used entropy measures for complex networks to implement an information theoretical approach for analyzing complex topologies [90].

In thermodynamics, entropy is related to temperature. So it is not a surprise to find the idea of a "*social temperature*" in literature. K. Kułakowski [91] used the concept of magnetic susceptibility in the Ising model[4]

[1] The Zacary's Karate Club is represented by a social network of friendships between 34 members of a karate club at a US university in the 1970. The data set is public and available here: `http://www-personal.umich.edu/~mejn/netdata`.

[2] The Milgram experiment shows that the distance between two people on earth in average is six but recent results from Facebook show, that the world represented by the social graph is indeed smaller with an average distance of four steps [45].

[3] In this work we use only the next neighbors in the same language like a chosen node and the second neighbor of the nodes in different languages. Such local networks are better to handle. But since we do not want to disconnect the data completely, a comparison of results for different crawl depths allows one to identify the impact of such a cut.

[4] The Ising-Modell is a concept in theoretical physics. It was first studied by Ernst Ising in 1924 (suggested by Wilhelm Lenz) in order to understand ferromagnetism - which can be seen as collective behavior of matter.

to identify a social temperature. His work is related to game theory and interprets the rate of random strategy changes for a set of agents playing two different strategies as a temperature.

The question *"Is the term social temperature just a rhetoric figure ..., or on the contrary, could it be given a precise meaning?"* is raised by Floria *et al.* [92]. They show that *"the formal framework of Equilibrium Statistical Mechanics is, to a large extend, applicable to the description of the asymptotic behavior of strategic evolution. Thus it is providing the key for a formal quantitative meaning of the term social "temperature" in these contexts."*

The "*social force model*" as introduced by Helbing *et al.* [17, 93] leads to a framework to simulate the behavioral process in human crowds and their motion. Crowd dynamics is a consequence of the individual properties of people in this crowd and additional related factors such as shared goals or beliefs. Like the motion of a particle is influenced by superposition of different forces also the motion of people is influenced by many different aspects. Even if those aspects are not forces, in terms of physics, they can be seen as useful factors to describe or simulate the decision making process in simulations of human crowds. An application of the social force model is presented in the appendix (see chapter 17).

Another metaphor with relation to traditional physics is called "*social gravity*". Social gravity was defined by Bannister *et al.* [94]. They developed a force-directed layout algorithm to produce graph drawings by resolving a system of emulated physical forces. In this technique they use social gravity as an additional force to the traditional force-directed layout (presented by Fruchterman and Reingold [95]). This modification allows them to draw trees and forests, as well as more complex social networks in a meaningful way. The core idea of Bannister *et al.* [94] is that social gravity assigns a kind of mass to network nodes. Because this new property is proportional to the network centrality of the node, a structural property - derived from the social network structure - is included into the calculation of the total force on that node.

Social networks can be compared with biological and technical networks. Newman and Park [96] found that social networks have non trivial clustering with relation to transitivity. Furthermore, they found positive correlation in the node degree, known as assortative mixing. They explain those differences as a consequence of the clustering in social networks, which is much higher than clustering that is found in natural networks.

Another interesting feature identified in social networks is caused by feedback loops. Feedback loops are typical properties of a complex system. Kawamoto and Hatano [97] studied Twitter communication networks and found a correlation in the re-tweet rates. As a consequence of these correlations, an explosive diffusion of information occurs. They observed a changed average re-tweet rate together with increased fluctuations which both can be seen as a reason for the observed explosive diffusion.

4.2. Computational Social Science

The continuously increasing impact of computers on society can not be overseen. A comparable influence comes also from modern communication technology (internet messengers, mobile phones, Facebook, Twitter), which initially all were computer based, but not limited to desktop PCs any longer. High speed Internet and highly flexible interconnected mobile devices allow many different use cases in commercial contexts or even life style related. Nowadays, online content can be consumed more or less anytime[5].

In the western society, online platforms are all around us and using them does not require an explicit context switch any more. This means, in order to use a social app (short for social media application) one just has to take the mobile device out of the pocket and no matter what one was doing, such a system can interrupt the personal activity stream in a disruptive way. But also positive effects can be observed, e.g., one can share positive emotions and feelings also easily without a significant disturbance of the own mood. Kramer *et al.* [98] conducted an important large-scale experiment in real life. This experiment was based on real world data, obtained during planed manipulations of the underlying system. They used data from 689,003 Facebook accounts to study the transfer properties of emotions. Emotional contagion was found to be able to bring people into the same emotional state without their awareness and even without a direct interaction with other people.

Dodds *et al.* [99] conducted data analysis on 4.6 billion expressions from online communication of 63 million people using Twitter. They created a metric to measure happiness of people and refined the quality of the results by combining the data-driven approach with a traditional survey on 10,000 users. The tool they use is called *word-shift-graph*[6]. This method is based on the concept, that usage patterns of words are related to the emotional state of the people which communicate. Both projects (Kramer *et al.* and Dodds *et al.*) are comparable. They are representative examples for an increasing trend regarding the usage of huge data sets obtained from public spaces in order to support social science. This requires also more intense computational treatment and leads to the increased importance of data science techniques in the field of traditional social science.

But from a technical and conceptual perspective both research projects are very different. Kramer *et al.* influenced the system they study actively. They reduced the number of positive or negative expressions during

[5] Spatially, the access to such technologies is still limited, depending on the continent and political situation in the region. Africa has very weak infrastructure and blocked services in China are a consequence of the censorship.

[6] The word-shift-graphs compare the positive and negative rank changes of words used in communication about a topic. Cody *et al.* [100] used it, e.g., to conduct a poll on opinions related to climate change based on Twitter data.

communication without a notice to users. On the other hand, a purely data-driven approach as presented by Dodds *et al.* had no direct impact on people. This is why Kramer *et al.* triggered an active discussion about ethical aspects of such interfering experiments. One can clearly see, that a deeper integration of social science and technology has huge advantages. Statistical results can be much more significant and methods more robust if more data is available. Nowadays, data collection can be much easier because many details are available as side product of traditional system operations, e.g., included in server logs. But, one must not forget the risks and potential damage on society and on individual persons. The impact of such analysis procedures and even the impact of results is not well understood at this time.

4.3. Content-Based vs. Communication-Based Social Networks

Wikipedia has been in the focus of scientific research projects for more than a decade. Already in 2005, five years after the Wikipedia project was started by Jimmy Wales, Holloway *et al.* [101] published a study about the semantic structure of Wikipedia. They used a technique called *category mapping*. Such a category map shows a network in which category pages are nodes and links between them are derived from articles by a co-occurrence analysis. Two categories are considered to be equal, if both are assigned to many common articles together.

Boyack *et al.* [102] present 8 different similarity measures to calculate similarity links for scientific journal articles. They differentiate between co-citation and inter-citation. If journal **A** contains citations to journal **B** and **C** we can say, **A** defines a co-citation link or at least **A** contributes to the co-citation link strength between **B** and **C**. Inter citation links are the obvious links. Analysis of co-citation links requires deeper analysis of the document corpus.

Wikipedia structure was analyzed in both mentioned studies, but not based on article content and not with a strong focus on different languages. They also ignored the influence of usage patterns on structure forming processes.

Since Wikipedia offers usage statistics we can present new results about the dynamical aspects of Wikipedia here in this work. Details about the technical background, especially about the aggregation procedures applied to the server log data to generate hourly access statistics on a per page level for all Wikipedia pages, can be found in Holloway *et al.* [101] and Reinoso *et al.* [103].

Reinoso *et al.* [103] provide useful insights regarding Wikipedia usage. They had found the daily cycles in access activity and concluded that Wikipedia is typically used during the working hours. We could identify different usage patterns on a weekly time scale, which show significant differences for working days and weekends depending on the topic (see section 7.2.1, figure 7.5 and figure 7.6).

Furthermore, they compared the access and edit activity and found, that less than 7 % of Wikipedia usage (page requests) is related to contributions. Having this in mind one can say, that Wikipedia is primarily a content driven social platform. On the other hand there are the online communication platforms such as Facebook and Twitter.

A classification of social networks was done by Kim *et al.* [104]. According to their study, online services can be classified by multiple criteria, e.g., regarding dominating properties, or core functionality. Such functionality can be: (a) e-commerce, (b) content publishing, (c) content sharing, (d) social interaction, (e) personal trading, and (f) messaging services.

In all cases, at least a personal profile or a descriptor of any type of item, such as products, services, or books form the content base. This content can be interlinked. Content sharing requires also a network of interacting people. Some users offer content, others search for it and request it. Their exchange can be legal like trading of products or illegal such as sharing of copyrighted material for free.

A totally new user experience comes from an intense social interaction in content publishing platforms which allow early communication between readers and authors of a book about a trending topic, such as open source software and technology. Publishers like O'Reilly offer early access programs. This way, the community can influence the creation of the content.

Many lessons can be learned from social networks offered on the WWW. Personal activities, such as trading on online market places like eBay and messaging via Facebook and Twitter would not work without the combination of the two dimensions: *content* and *interaction along social connections*. One cannot isolate those aspects any more!

"What is Twitter, a Social Network or a News Media?" is the title of an article published by Kwak *et al.* [105]. They collected 41.7 million user profiles, 1.47 billion social relations, and 4262 trending topics from 106 million tweets. In comparison to other social networks, they found a deviation in the degree distribution. The follower network has a non-power-law distribution. Measuring the reciprocity of the network reveals that most of the users just consume twitter messages. They use Twitter as a source for messages rather than being active in social interactions. Furthermore, Kwak *et al.* [105] found a non-power-law follower distribution in a topological analysis of the entire Twitter site. The follower graph has a short effective diameter and low reciprocity. This marks a deviation from known characteristics of human social networks. Furthermore they found that the majority (over 85%) of topics are headline news or persistent news in nature.

An analysis of three different node ranking measures shows that static and dynamic aspects differ: Kwak *et al.* report that PageRank and a ranking based on the in-degree show comparable results, but the node ranking based on re-tweet activity differs significantly. Especially because of the special relation between followers or fans of famous Twitter accounts, the likelihood of re-tweeting seems to be higher. They also found, that user participation is different for different topics. There exist core members which produce content for special topics, e.g., political topics. In other topics, the number of participating users grows over time; this indicates a real increase of interest in this topic. Furthermore, they compare the freshness of topics in Google Trends and Twitter. 95% of trending topics are new on Google each day, but only 72% are new on Twitter. This means, persistence is higher in Twitter, probably because of the internal structure of the social community. Currently, this is only speculation, because a comparison of both systems on a structural level has not yet been done.

This thesis can not solve the problem, but we contribute techniques which allow creation of hybrid analysis procedures to study the time-dependent properties of multi-faceted systems on large scales. E.g., we have found that contribution to and usage of Wikipedia are two different processes with different properties regarding short and long-term correlations. The content creation process shows correlations only on very short time scales while the information consumption process contains long-term correlations. For more details see our publication: "*Fluctuations in Wikipedia access-rate and edit-event data*" [8] which is also presented in chapter 13.

4.4. Wikipedia: A Complex System of Connected Networks

Why have we used Wikipedia data in this work? Since Wikipedia is a very large and well known global online system, it can be seen as global proxy for user interest in particular topics. It is available in more than 230 languages and all kinds of mobile devices can access the service, which is hosted by the Wikimedia Foundation. Wikipedia usage is free. This means, we do not have to care about access limitations because of social status. Because Wikipedia is available since 2001 on traditional PCs and also on mobile devices we can assume, that there is no disturbing effect from recent short term marketing hypes, related to individual devices or communication applications. Examples for systems with an implicit bias are the Google Play store and Apple iTunes which are restricted to specific devices only. Social media systems can be restricted to a specific audience because of technical reasons or because of a very strong focus on a too narrow group of users. Nowadays, Facebook is widely accepted by all kinds of users from all generations. Contrary to this, e.g., "Stud.IP" forms online communities focused on students in German universities. Such restrictions do not exist for Wikipedia. But there are indirect restrictions, especially for political and economical reasons. Even if Wikipedia usage is in general for free, not all regions on earth are well connected via the Internet[7]. According to such factors one has to be aware of the fact, that not all topics are represented equally, nor do all topics exist in all languages. We developed a method which allows us to take all information about a topic, which is available in any language, into account. With this, we can compare how broad and comprehensive the representation of a particular topic in Wikipedia really is (see section 11.2).

Finally, I think it is important to notice that Wikipedia is not driven by particular business needs or an interest in earning money with the system. No advertisement is influencing what a user can read nor are there additional recommendations which also change or influence user behavior. The Google search keyword history is another very often used proxy to measure user interest. Google Trends enables a comparable kind of traceability of user attention, but with less transparency, since the raw data used in the system is not public.

Is Wikipedia a social network? Inspired by this question we analyzed available publications about Wikipedia. Our conclusion is: Wikipedia can be seen as a social network. Users, especially editors are related to each other during collaboration on an article. This can be a positive and inspiring relation or even disruptive. Kaltenbrunner *et al.* [106] study the editorial process of Wikipedia articles, which is considered to be never ending - Wikipedia articles are never complete. Additional explicit social interaction happen on the accompanying talk pages. The comment activity per page is in general higher than edit activity but both show comparable properties. A power law is reported for the number of activity peaks per article, peak-lengths, and time between two consecutive peaks. A negative form of interaction between Wikipedia users was studied by Yasseri *et al.* in [107]. They analyzed a phenomenon called *edit-wars*. The edit-wars are an example of destructive social activity.

Viégas *et al.* [108] introduce the *history flow* visualization mechanism. This tool highlights collaboration patterns and allows a visual analysis of article histories in a convenient data exploration procedure, e.g., edit-wars are visible as 'zigzag' patterns (see figure 6 in [108]).

With all those results in mind we can argue that Wikipedia is also a complex system. In the following section we review the criteria list, presented in chapter 2: **(a)** *Many coexisting links or dependencies in complex systems lead to feedback loops:* Wikipedia shows a complex link structure which involves different types of links. Traditional links between two pages in the same language are enriched by so called inter-wiki links. Inter-wiki links connect pages about the same topic in different languages. Furthermore, membership in a category can also be interpreted as a link. The pages and categories (which in general are also pages but of a specific type) form a bipartite network. And as shown in the previous chapter, bipartite networks can be transformed into networks of one node type only.

[7] Accessibility of the Internet in 2015 is reported in the "World Internet Users and 2015 Population Stats" report `http://www.internetworldstats.com/stats.htm` and for 2016 in `http://www.internetlivestats.com/internet-users/`.

(b) *Interference, or resonance effects are, e.g., related to self- organized structure formation. This way, complex systems exhibit emergent phenomena.* We can find the exponential growth of activity which indicates interest in particular topics. Activity patterns of linked pages can be very similar because of overlapping topics and user interest. Category links are like short cuts and support these overlapping aspects even more.

(c) *The multiple ways of coupling (especially the strong links and short distances) lead to cascading failures which may have catastrophic consequences on the overall system behavior.* Destructive elements exist also in Wikipedia. Vandalism and edit-wars are two cases. Potentially this can also lead to cascades of failures and uncontrolled deletion or manipulation of page content. But in general, due to the technical properties of the Mediawiki software, the system can recover from such a condition.

(d) *As a result of coupling multiple systems on different scales coexist. Complex systems are nested.* The complete Wikipedia system consists of many particular Wikis, one per language, and inside each Wiki, the content is organized hierarchically in categories. Since the page network shows the scale-free structure one can clearly conclude, that also Wikipedia consists of coupled systems of multiple scales.

(e) *Nonlinear interactions* cause the exponential increase in user activity. This can be seen as another example of an emergent phenomenon caused by the densely interlinked Wikipedia pages.

Furthermore, we said: *complex systems may be open.* This is clearly the case for Wikipedia. New pages and links are added regularly even if the overall size of the system is already really huge.

We close this chapter with some more facts about Wikipedia's structural properties. Kamps and Koolen [109] report that the Wikipedia link structure is comparable with the WWW, but Wikipedia pages are more densely linked. For Wikipedia, the outlinks and inlinks behave the same and both are good indicators for relevance of an article. Interesting is the difference in their local link structures. Because of the higher density of local links in Wikipedia - probably as a consequence of categorization of content - such a local context is more applicable as a good indicator for relevance ranking of pages. This also motivates our approach of local neighborhood networks - which ignores the global link structure of the entire system.

Capocci *et al.* [110] have studied the structural properties of Wikipedia already in 2006. They found a close analogy to the structure of the World Wide Web but major differences in the growth process of Wikipedia. They investigated the portion of links which belong to the specific areas in the so called *bow-tie representation* of a network (see fig. 1 in [110]) and identified a substantial lack of correlation between the in-degree of nodes and the average in-degree of the related upstream neighbors. They also identified a clear community structure in the English and Portuguese Wikipedias.

After an introduction of concepts for complex systems research, network theory, and a review of research results related to Wikipedia the mathematical methods employed in this work are described in the next chapter.

5. Mathematical and Computational Methods

> Prediction is very difficult, especially if it's about the future.
>
> *(Nils Bohr, Nobel laureate in Physics, 1922)*

Starting with practical considerations, related to data acquisition and data cleaning methods, some uni-variate and bi-variate time series analysis procedures are described in this chapter, which closes part I. Later we apply these methods in particular case studies. The chapter ends with a short explanation of used statistical tests and a discussion of surrogate data generation methods. A combination of several computational methods is required for network reconstruction and characterization, as explained in part II.

5.1. Describing Real World Phenomena with Time Series

In order to analyze the behavior of things or phenomena one needs an appropriate representation of them. Such a representation can be just an idea which allows at least to think about it or one which can be visualized for better understanding and communication about it. In many cases a verbal description using all the richness of language and even drawings were a preferred way to represent and share knowledge, e.g., when Alexander Humboldt explored Central and South America in the 18-th century when no camera existed. Scientists of this era had to paint and draw (see, e.g., the work of Georg Forstner[1]) or they wrote prose as, e.g., Johann Wolfgang Goethe[2].

Nowadays, as a consequence of the technical revolutions, it is much simpler and common to create a digital representation of the real world which can be analyzed by a manifold of measurement procedures and mathematical treatments. Furthermore, many properties of individual entities can be recorded over time. If a continuous recording is not possible we use a discrete approach based on *snapshot series* (e.g., one image per time step) or simply ordered series of data points. In order to study the dynamics of a system we must guarantee the right order of all observations. This means, the time of the specific measurement must be known as accurate as possible. All information regarding time intervals and resolution is an example for implicit usage of metadata.

Time series analysis means: a series of chronologically ordered data points (measured values) has to be combined with all information required for contextualization. A series of values for which a start time t_0 and a constant time offset interval δt is known is called equidistant time series. The number of available data which can efficiently be handled and the affordable time resolution define natural limits for data analysis and information retrieval procedures.

5.1.1. Continuous vs. Discrete Time Series

Many properties of nature are described by analytic functions. Such a mathematical representation is helpful primarily for theoretical investigation and furthermore for analytic models. In general, an analytic model is a set of formulas, which describe the behavior of a system as a function of time or any other parameter. Since many different influencing parameters can be included, even if they can not be known exactly upfront, they can be estimated based on regression on measured data. Finally, the desired models allow predictions about the future behavior of a system. In general, a time-dependent property exists in continuous time.

For practical reasons, time series are discretized. Discretization of continuous values happens automatically during the measurement process and is predominantly influenced by the measurement device. Because the time-dependent analytic function is evaluated only at discrete times we loose information during the discretization procedure. Evaluating the function means here, that either a value $y_{\mathrm{calc}}(t)$ is calculated for time t or a measurement $y_{\mathrm{measure}}(t)$ is available at time t. If no data is available, usually zero is recorded, which is a fundamental problem if zero is also a valid value which could be measured. In this case it is not possible to distinguish, if the value exists and is zero or if a value is not available at this time. The measurement procedure can include post processing steps, such as re-sampling. In case of simulations, one has to discretize the data because digital devices such as computers can not handle continuous values. Although analog computers have several advantages, e.g., they can be used to solve ordinary differential equations (see [111]), they are not very common in recent data analysis and large-scale simulation applications.

[1] His drawings are presented in the National History Museum in London and also published in several books. More Details about G. Forstner can be found on Wikipedia.

[2] Examples are available online: `http://www.zeno.org/Literatur/M/Goethe,+Johann+Wolfgang/Naturwissenschaftliche+Schriften`

During the discretization procedure the indefinite number of possible times is replaced by a definite number of discrete values. The discretization is usually done in time or space, or both. Usually, a time series consists of a list of (measured) values, i.e., the time series data, and additional metadata. Metadata enables interpretation of the values in the right context.

Aris *et al.* [112] investigated advantages and disadvantages of several representations of event time series. Their work is relevant especially for large time series data sets, since the representation has a major impact on the overall performance of analysis procedures.

Mörtl *et al.* [113] designed a continuous dynamical process utilizing both continuous phases and discrete events in a unifying view. Events are important for segmenting complex trajectories into primitives. Such primitives are introduced as anchoring points for enhanced synchronization modes. They study the effectiveness of the designed behavior by objective measures of phase and event synchronization. More details about event time series follow in section 5.1.3 later in this chapter.

5.1.2. Modeling Time Series

A simple theoretical model of a time series - taken from the book *"The Analysis of Time Series: An Introduction"* [114] - is described by:

$$y(t) = A(t) + B(t) + C(t) + D(t) \tag{5.1}$$

where $A(t)$ is a seasonal variation of a specific period, $B(t)$ describes trend variations, $C(t)$ describes cyclical variations which correspond to recurring factors, and $D(t)$ covers all random variations, such as random extreme events which are not already covered by one of the three previous terms or noise of a particular type. Practical applications are based on scalar data, measured at discrete times (see above).

A time series model is useful to represent a natural process by calculating the values $y(t)$ for any given t. Future values of a variable can be calculated more or less precisely, depending on the type of the underlying process and the knowledge about the system. In deterministic systems, the future position of an object can be calculated by equations of motion. But as soon as an unknown event occurs or a previously not considered factor influences the system in addition to all already covered influences, the results of such a prediction are not precise or even useless.

In general, applications of models derived from data, are only valid in the context of several, often limiting, assumptions, e.g., existence of a stationary closed system. Terms $A(t), B(t)$, and $C(t)$ describe deterministic processes. The third term, $D(t)$ is part of a second class of processes. Random processes are found in nature very often. *Brownian Motion*, and *Random Walk* are important examples of random processes and have been studied intensively.

An alternative to the representation of a phenomenon in the time domain, is the frequency domain. A transformation between both complementary domains is given by the Fourier transform $\mathcal{F}$ and inverse Fourier transform $\mathcal{F}^{-1}$. Many applications use Fourier transform combined with filters for segregation of a system into different subsystems. The length T of a time series - the time range between the first and the last data point - and also the resolution Δt - which is the distance between consecutive data points - influence the minimum and maximum frequencies measurable in a time series, $f_{\max} = \frac{1}{2\Delta t}$, $f_{\min} = \frac{1}{T}$. In general, some information about the underlying process gets lost during the measurement because the length of the time series, the resolution, and the bandwidth of the Fourier spectrum are limited by technical and practical reasons.

As mentioned before, spectral filters are applied in order to isolate or separate individual coexisting aspects of processes. Such aspects can be decomposed by band filters, e.g., to separate individual bands of EEG time series (for more details see, e.g., section *Method Summary* in Bashan *et al.* [22]).

The frequency ranges of the relevant bands depend heavily on the domain one studies. In our case, the Wikipedia edit activity is recorded with a time resolution of one ms and for access activity hourly data is available. For a better understanding of the edit activity, the shorter intervals are very helpful, but a direct comparison of both aspects is not possible in this way. On the other hand, weekly and daily activity cycles can be found in Wikipedia using this data.

Even in the case of data about financial markets we have to deal with lots of limitations. Publicly available data is limited to daily resolution. This makes the data in general comparable with data obtained from Google Trends, which has the same time resolution. But relying on daily closing prices is very dangerous. It also leads to wrong assumptions about the market, or to wrong conclusions (see page 323 in "Fraktale und Finanzen" by Mandelbrot and Hudson).

According to [115] the *Piecewise Linear Representation* (PLR) is perhaps the most frequently used representation of time series for mining time series databases. The procedure, which creates such a piecewise linear representation is called *segmentation*. Especially in the context of similarity analysis it is a good practice to represent a time series by a set of motives[3]. Such motives can be simply linear approximations of the time series or even polynoms of a higher order. The DFA method (introduced in section 5.3.2) uses also segmentation, but instead of reducing the data to a small number of parameters per segment, the DFA works on residuals after removing the trends from each segment. Thus, DFA can be seen as complementary approach to plain motive analysis.

[3]Motives can be of any shape, not only linear approximations of the function.

Eq. 5.1 combines deterministic and stochastic processes. A purely stochastic approach is the decomposition of the time series into a permanent and a transitory component, also known as the *Beveridge Nelson Decomposition*. This method was introduced by Beveridge and Nelson [116] in order to study business life cycles based on economic time series. They state: *"... a large number of studies has shown that many economic time series are well represented by the class of homogeneous non-stationary 'ARIMA' processes. In such a process, the first differences are a stationary process of autoregressive-moving average form."* In many cases, a transform to natural logs is required, before the first differences are stationary [116], see also section 5.2.5. The benefit of this model is that at any given time only values from the past are required. This avoids extrapolation problems associated with two-sided filtering methods such as centered moving averages [116]. Since the decomposition depends only on the past, the calculation can be done in real time.

5.1.3. Event Time Series

For this work, we had to distinguish between discrete time series (evenly spaced) and unevenly spaced event time series (ETS) based on data availability. As long as events are not really discrete, a lower sampling rate can be used, but according to the Nyquist–Shannon sampling theorem information is lost if expected duration, and sampling rate do not match. Due to sampling it is also possible to loose the precise time of the event.

For social media data this seems not to be a problem but in technical use cases and for intra-day trading the exact timing is of high importance. Using ETS solves this problem. Each value is stored in a list with an exact timestamp. This also reduces required storage especially if only a few sporadic events have to be handled. But event time series are not efficient for continuous values.

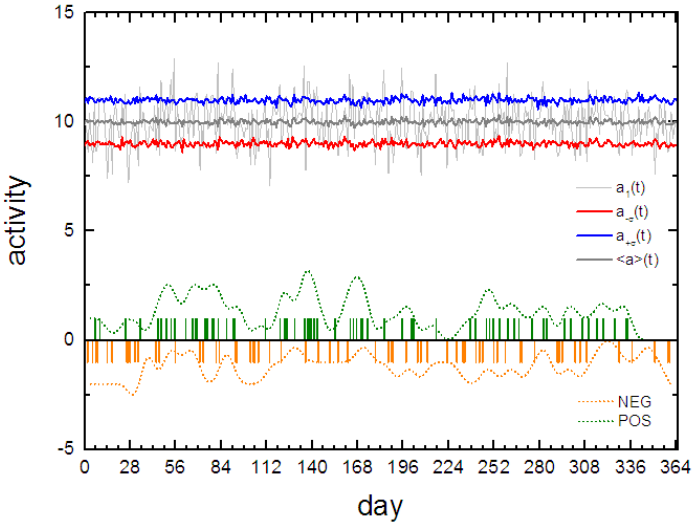

Figure 5.1.: **Transformation of discrete equidistant time series to ETS**. A node property (e.g., access activity $a_1(t)$, light gray) represents one element from an ensemble of 100 elements ($\langle a(t)\rangle$, dark gray). All ensemble members are random processes with Gaussian distribution ($\mu = 10.0$ and $\sigma = 1.0$). Events are generated if a threshold is exceeded. Dynamic thresholds are $\pm\sigma$ in this case (blue and red curves).

A detailed investigation of efficiency of different event time series is presented by Aris *et. al.* in their article *Representing Unevenly-Spaced Time Series Data for Visualization and Interactive Exploration* [112]. They discuss the challenges for unevenly-spaced time series data and compare the following four methods: (a) sampled events, (b) aggregated sampled events, (c) event index, and (d) interleaved event index.

They describe the advantages and disadvantages, choices for algorithms and parameters, and compare the different methods regarding performance. We consider this document as a guideline for further improvements of the Hadoop.TS software package [11].

ETS can have a natural origin, e.g., different things happen at distinct points in time. ETS can be obtained by transforming a continuous time series as shown in figure 5.1, e.g., if peak detection algorithms are applied in order to find events based on patterns and logical rules.

Such events have to be characterized by specific parameters (threshold, peak form, motive, and time scale) and their time stamps. The time stamp describes when the event occurs and thus it can be the beginning, the center or the end of a particular pattern or even the time of a characteristic property such as a minimum or maximum value.

Threshold based transformation of continuous time series to ETS: The time, when a continuous variable exceeds the threshold defines an event. The direction can also be used to provide further information, such as categories of events (see olive and orange spike trains in figure 5.1). For an ensemble of time series we use the following mathematical expression: $e(t)\exists : a_i(t) > n\cdot\sigma(a)$ to define individual events based on the group properties. Figure 5.1 shows one example out of a set of 50 time series in light gray together with the ensemble average in dark gray. With $n = 1$ the upper and lower boundaries are given by $\langle a_1(t)\rangle \pm \sigma_a$ respectively.

The number of events divided by total time (length of the interval) is called event density. This allows an equivalent usage of "density" and "event count".

One has to be careful, if different data sets with different aggregation properties such as start and end time, length, and bin with from different sources should be combined. In case of different filter methods the data sets

would be incompatible. Furthermore it is recommended to verify the calibration functions for every metric on each new data set.

Dynamic thresholds are shown as blue and red curves in figure 5.1. Such critical boundaries are derived from the probability distribution function for time interval i. They are defined by $(u, l)_i = \langle a(i) \rangle \pm n \cdot \sigma_a(i)$. Resulting ETS contain the time stamp together with a constant value of one (see olive and orange spike trains) to indicate only the occurrence of an event. A scalar value is useful to to express also the criticality of the event, not just its existence. An aggregation time $t_a = 7$ defines the time interval during which individual events are aggregated in order to create a continuous signal from event based data. The result is a smoother equidistant time series which represents activity as raw data but now based on events (olive and orange dotted lines) instead of the full data set.

5.2. Data Cleaning and Preparation

Data sets are not always as clean as expected. Due to technical problems some values can be wrong or absent. Domain specific effects - such as externally influenced patterns (also called exogenous influences) - can lead to variations in the data, which do not reflect the real properties of the underlying process, one wants to study. E.g., if we are interested in the dynamical properties of attention paid to web resources it is not enough to measure click rates only. E.g., a jump in activity can be caused by a real increase of interest in a topic, but also by additional access channels which just connect more users that have already been interested in that topic long time before. This illustrates that it is highly relevant to understand the context of the system one studies.

Time series pre-processing aims on providing the right data - which means the right time series data, grouped according to a particular research question - in an appropriate representation for further analysis procedures.

If data points are missing, one can try to fill the gaps based on simple assumptions, such as using average values based on a comparable time interval, or simply by interpolation on existing data. In general, such preparation operations are not part of the analysis procedure, rather they belong to the measurement process. If the measurement procedure can not be controlled, the measured results are not reliable and additional steps can not lead to trustworthy results. Thus it is very important to use only reproducible operations during the data collection and pre-processing phase and in addition to keep all raw data for later checks and validation.

If data comes from different domains it is important to identify the right transformations which allow a comparison. Two common approaches are "zero-transformation" and "first differences". For time series modeling algorithms such as the AR, MA, or ARIMA processes it is also important to provide stationary time series.

5.2.1. Standardization

Time series can only be compared to each other if properties like mean value, standard deviation, or even both are in the same range. In order to achieve this, a processing step called normalization, also known as "zero-transformation", is applied. Via a linear transformation of all samples of the time series a normalized time series $y_{\text{norm}}(t)$ is calculated: $y_{\text{norm}}(t) = (y(t) - \mu)/\sigma$ with $\mu = \langle y \rangle$ and $\sigma = \sqrt{\langle y^2 \rangle - \langle y \rangle^2}$. The result $y_{\text{norm}}(t)$ has mean zero ($\mu = 0$) and unit standard deviation ($\sigma = 1$).

Other relevant standardization approaches are Gaussian, uniform and linear standardization. For the Gaussian and the uniform standardization a rank ordering function from the sample marginal distribution to a Gaussian distribution (uniform distribution respectively) is applied. In case of a linear standardization the value range is computed via a linear transformation so that the minimum is transformed to 0 and the maximum to 1.

5.2.2. Detrending

Especially if linear, polynomial, or even periodic trends can be identified clearly, they should be removed from the original data before a deeper analysis with a focus on non-deterministic aspects starts.

Knowledge about such trends is valuable because they allow a classification of the system elements which are related to them (see section 7.2.1 for more examples). After such trends have been removed one can separately analyze signals and noise or better to say the deterministic and stochastic part. Figure 5.2 shows three access-rate time series for Wikipedia pages (a,c,e) and their weekly trends (b,d,f). Since all pages are from the same language we can not see a time delay for the nightly access minimum. The influence of extreme events is visible also in the weekly trend patterns, in (d) on Sunday, and in (f) on Wednesday and Thursday.

Cyclic patterns have already been found by Yasseri *et al.* [117] in Wikipedia edit-event time series. Our report about the Swedish Wikipedia project [10] shows such activity cycles also for Wikipedia access-rate time series. We found two typical patterns: (a) weekly, and (b) daily patterns. Both are superposed. The daily pattern is caused by human life style or sleep patterns. Seasonal effects, content specific usage patterns, and cultural influences are the reasons for the overlapping patterns on multiple time scales.

Furthermore, a procedure called *"whitening"* is used to remove trends from time series (see Friston *et al.* [118]). First, a model is extracted from the data, and than, based on the parameters which are used to describe the trend all the time series are cleaned by subtracting the trends. This approach can only be used if all time series are

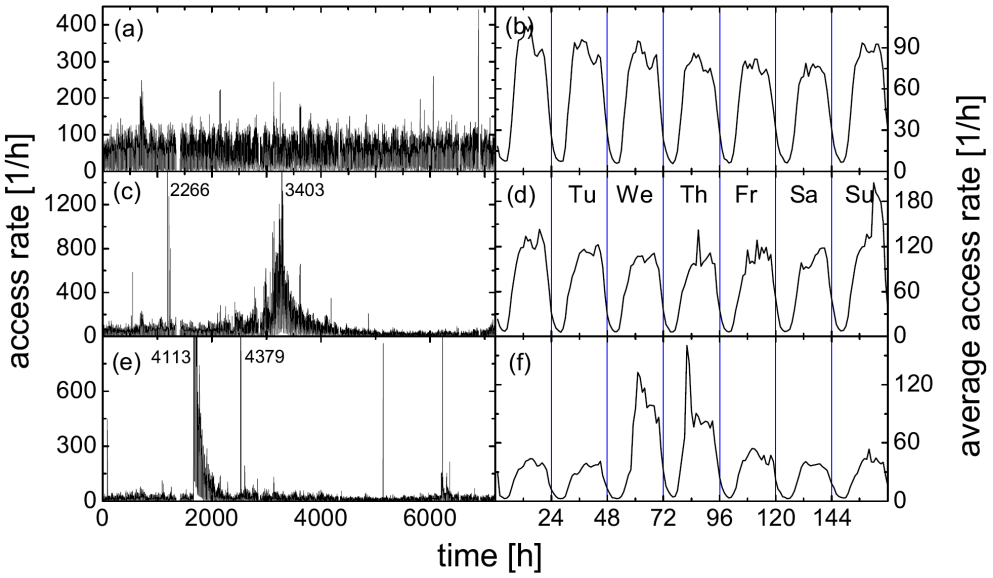

Figure 5.2.: **Examples of Wikipedia access-rate time series and weekly trends.** (a,c,e) show Wikipedia access-rate and (b,d,f) weekly trends obtained for individual time series of three selected articles with (a,b) rather stationary access-rates (topic 'Illuminati (book)'), (c,d) an apparently endogenous burst of activity (peak on May 7, 2009, topic 'Heidelberg'), and (e,f) an exogenous burst of activity (topic 'Amoklauf Erfurt' (shooting rampage); it peaks on March 11, 2009, as another shooting rampage occurred in Winneden). The left parts show the complete hourly access-rate time series (from January 1-st, 2009, till October 21-st, 2009; i. e. for 42 weeks = 294 days = 7056 hours) with numbers in the plot giving the height of peaks truncated to show baseline fluctuations. The gap around $t = 1200$h is a systematical disruption and was found in all records. Parts (b,d,f) show weekly average access-rates with hourly resolution. Note, that the peaks from the left panels (c,e) overlap with daily cycles (d,f) on specific days (Sunday in *(d)* and Wednesday to Thursday in *(f)*).

expected to show the same fundamental properties which are represented by the selected model. Friston *et al.* [118] discuss the whitening procedure and the related bias which can be introduced by it in more detail. They show that: "*(i) whitening strategies can result in profound bias ...*" and that *"(ii) band-pass filtering, and implicitly smoothing, has an important role in protecting against inferential bias."*

In case of Wikipedia we found, that trends are very different between pages from different topics and languages. Multiple time zones and a variety of usage contexts lead to different weekly trends. Because of this we decided to remove such trends on a per-node level.

The second new approach we introduced in this work is *contextual detrending* (see figures 7.3. and 7.4. in chapter 7). This approach is a normalization technique based on the local neighborhood defined by content semantics and local link structure.

5.2.3. Time Shifts and Gaps

Data sets from different domains can contain different features which can cause multiple side effects. In our case study in chapter 15 we use data form Wikipedia and from stock markets. While the Wikipedia access and edit history is available for all days - except during periods of unpredictable system failures - the stock market activity is limited to trading days. Periodically, during the weekends there is no trading activity, and bank holidays also interrupt this activity. Ignoring such days would lead to information loss and wrong results.

One can expect a different behavior of people before or after such special non-trading days. If such days are simply ignored, we would have to ignore also the available data from the second domain, e.g., hypothetically we can think about the following scenario: private traders (non professional trading persons) use Wikipedia and other online systems to collect information about stock markets before they buy or sell. During the weekends or on official holidays, they probably have more time available for such information consumption activities. If this is true, we would miss the increased activity which is related to changes in the trading data during the next time period. This indicates how important the right alignment of time series from different domains is. Our approach is to fill the gap with the average value of the two surrounding values before and after the gap.

5.2.4. Stationarity of Time Series

A time series is called stationary if there is (a) no systematic change in the mean value (no trend), (b) no systematic change in variance, and (c) no periodicity. All individual sections of a time series (called episodes) show comparable properties if the time series is stationary.

Stationarity also depends on the length of the episodes. Short episodes can be interpreted as quasi stationary, since no trends and no periodicity exist on that selected scale. But especially for short windows the statistical significance of results can be the limiting factor. The nature of the process has a strong impact on stationarity on longer time ranges.

The average and standard deviation of a time series allow identification of purely random processes. In this case, both are constant. In case of a random walk, which is a non-stationary process with time-dependent average value and standard deviation, the series of first differences is stationary. Such transformations are also common in financial data analysis. The next section explains in more detail why so called *log returns* are used instead of raw returns or raw prices for portfolio analysis.

5.2.5. Preparation of Financial Time Series Data

Levels of stock prices are very different. They range from less than \$10 to \$1000 within one index (a group of stocks representing a market) during one year. Absolute values and absolute differences can not be compared well in this case. In order to be able to analyze relationships between such time series we need normalization. The preferred way to normalize the data is using relative changes. Hence, let us define the *return* r_i at time i: where p_i and p_j are the prices at times i and $j \equiv (i-1)$: $r_i = \frac{p_i - p_j}{p_j}$. Because the prices are not stationary, returns are used instead.

Furthermore, the assumption of a log normal distribution of prices means that $\log(1+r_i)$ is normally distributed. If this assumption is true we can apply cross-correlation analysis to series of log returns.

Figure 5.3.a shows stock prices for two German companies, Adidas, and Deutsche Bank - both included in the stock index DAX.

In general one can not assume an ideal normal distribution of log-returns but rather a stretched exponential distribution or even a power-law distribution with tails [119]. The reason is the existence of extreme returns which follow a power law and also a dependency on the aggregation length (see [120]).

In order to test for a log Gaussian distribution we fit a parabola in double logarithmic scale (see figure 5.3.b). Differences identify the tails of a potentially existing power law distribution. Those differences are related to non stationarity. They are not caused by trends in this case. Because of this we are not able to eliminate them by detrending methods. Such non stationarity should not have heavy influences if cross-correlation is applied to relatively short sliding windows, but in case of very long time series, the result would be dominated by the higher values. For increased time series length one would also not find a converging stable result because of fat tailed distributions.

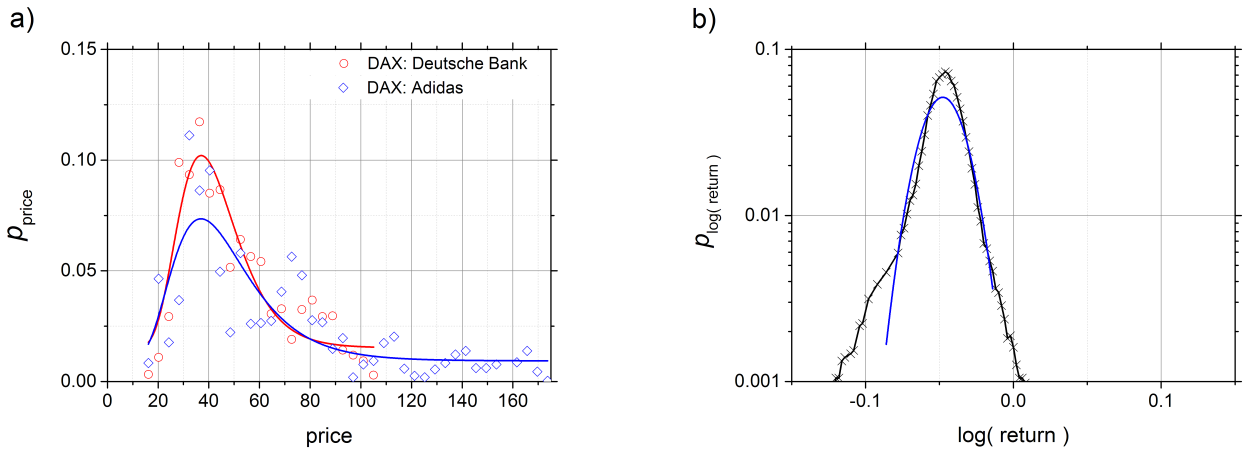

Figure 5.3.: **PDF of financial time series:** Although daily stock prices p_t are available, log-returns $\log(r_i)$ are used, rather than price p_t or raw returns r_t. The return r_t at time t is defined as the relative price change $\frac{p_t - p_{t'}}{p_{t'}}$ since the last time $t' = t-1$. This leads to an implicit normalization and allows a direct comparison of all components of a stock index, even if their price level, trading volume and contribution to the index differ. (a) Probability of prices p_t with log-normal fits. (b) Distribution of log-returns in log-log plot (black curve) and a parabola (blue curve) for comparison with a Gaussian distribution.

5.3. Univariate Time Series Analysis

This section covers time series analysis methods applied to single time series, called *univariate time series analysis*. Univariate analysis provides properties of individual objects. Multivariate analysis for link reconstruction is introduced in the next section and the formalism for network reconstruction is explained in chapter 9.

Our initial approach for describing and analyzing complex systems was primarily based on time series data which describe individual properties of single elements as a function of time regardless if there are correlations or dependencies between these objects. A time series contains data obtained from one individual thing. If links between such entities physically exist than one can also measure the properties of those links as a function of time. Examples of properties which describe links are the density of traffic on roads in urban road networks or the delay probability for flights in air traffic networks which both can vary over time. In some cases, the link properties are not directly accessible and have to be calculated from pairs or triples of time series as described in chapter 9.

Univariate analysis provides measures which can be assigned to a node or a link of a network as a weight. Especially if individual node properties should be taken into account weighted networks and thus also weighted measures are important. Weighted network measures address variability and dynamics of system elements and system structure. Wiedermann *et al.* [121] provide a detailed discussion about their new weighted network measures in the context of climate network studies.

In order to study the dependency between function and structure of complex systems it is important, first to create the appropriate time series representation and second, to select the right metric or measure for correlation or dependency analysis. Depending on the kind of network one studies, it is helpful to investigate the properties of individual elements by univariate analysis methods before a detailed topological analysis starts. This allows one to identify categories of comparable nodes regardless of internal structure and topology.

Some analysis methods can only be applied to, e.g., stationary time series. If such specific properties are required, additional transformations have to be applied to the raw data. If a process is stationary or not can be found out by analyzing the average values, variances, or auto-correlation. Time series with inherent trends can be detrended or analyzed via methods which include detrending (like detrended fluctuation analysis). A common approach is simply to fit a linear or polynomial function which is then subtracted from raw data series to prepare for residual analysis (see whitening in section 5.2.2). If no simple model can be found it is very handy to detrend by periodical averages, such as daily, weekly or monthly averages as described in the previous sections.

5.3.1. Autocorrelation

The autocorrelation C of a stochastic process describes the correlation between values of the process at different times t, as a function of the time lag s. It is important to note that sometimes, the term is used interchangeably with autocovariance, which is confusing since both are not the same. Autocorrelation is the result of normalizing the autocovariance function.

Let x be some repeatable process, and t be some point in time after the start of that process where t is an integer for a discrete-time process or a real number for a continuous-time process. Then $x_j(t)$ is the realization produced by a given instance j of that process at time t.

The autocorrelation function is calculated for a range of time lags s by calculating the Pearson correlation for the original time series with a shifted time series. The analysis begins with a subtraction of the average value $\Delta x_j(t) = x_j(t) - \bar{x}_j$ with $\bar{x}_j = \langle x_j(t) \rangle = \frac{1}{L} \sum_{t=1}^{L} x_j(t)$. Here, L is the length of the considered jth time series $(x_j(t))$. Then the autocorrelation function is calculated for various time delays s (see, e. g. [122]),

$$C(s) = \frac{1}{\langle \Delta x_j(t)^2 \rangle (L-s)} \sum_{t=1}^{L-s} \Delta x_j(t) \Delta x_j(t+s) \tag{5.2}$$

If the $\Delta x_j(t)$ are uncorrelated, $C(s)$ is fluctuating around zero for $s > 0$. For the relevant case of long-term correlations, $C(s)$ decays as a power law for large s characterized by a correlation exponent γ,

$$C(s) \sim s^{-\gamma}, \quad 0 < \gamma < 1. \tag{5.3}$$

A direct calculation of $C(s)$ is often hindered by unreliable behavior of $C(s)$ for large s due to finite-size effects (finite L) and non-stationarities in the data (i. e. a time-dependent, not well-defined average $\langle x_j(t) \rangle$ that changes with the considered length L). Such trends in the time series cause misleading results, e.g., a very slow decay. Such a slow decay is either an indicator for existing autocorrelations or for non stationary time series. Hence, this method works well only for short time lags, and thus it can not be used to identify long-term correlations.

5.3.2. Detrended Fluctuation Analysis (DFA)

The DFA method was introduced by Peng *et al.* [28] in order to overcome these obstacles. As an alternative to autocorrelation analysis the DFA method reveals long-term correlations in noisy, non-stationary time series. It has

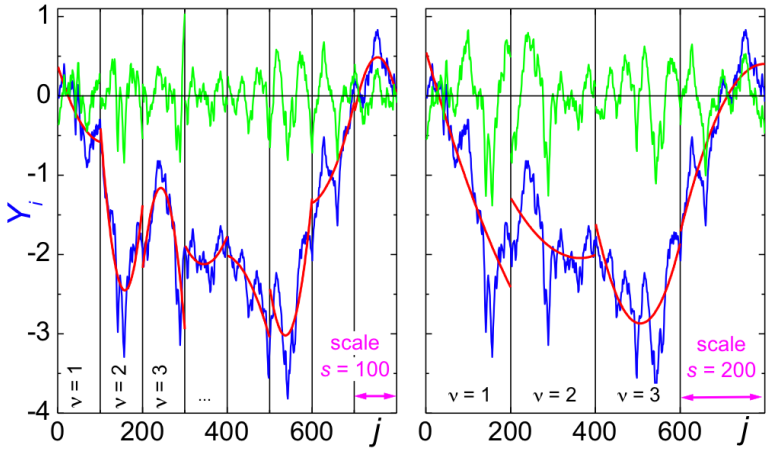

Figure 5.4.: **Illustration of the detrending procedure in the detrended fluctuation analysis.** The figure is based on figure 1 in [29].

become a widely used technique for the detection of long-term correlations in different fields including medical, climate, and economic data analysis.

For more detailed discussions of the method and its properties see [29, 122]. In general, the DFA procedure consists of the following five steps (see also figure 5.4):

1. calculate $Y_j(i) = \sum_{t=1}^{i}[x_j(t) - \langle x_j(t)\rangle]$, $i = 1, \ldots, L$, the so-called 'profile' (blue curve),
2. divide $Y_j(i)$ into $L_s = \mathrm{int}(L/s)$ non-overlapping segments of equal length s (black vertical lines define intervals),
3. calculate the local trend for each segment by a least-square fit to the data, where linear, quadratic (red curves), cubic, or higher order polynomials (conventionally called DFA1, DFA2, DFA3, ...) [32] are used in the fitting procedure,
4. determine the variance $F_s^2(\nu)$ of the differences between profile and fit in each time segment ν of s data points (green curve),
5. calculate the average of $F_s^2(\nu)$ over all segments ν and take the square root to obtain the fluctuation function $F(s)$.

Multiple iterations with segments of different s are necessary to determine the dependency of $F(s)$ on the time scale s. For long time series this is a time consuming procedure which fits well to the distributed approach supported by our software package. Usually, $F(s)$ increases with increasing s. If data $x_j(t)$ are long-term power-law

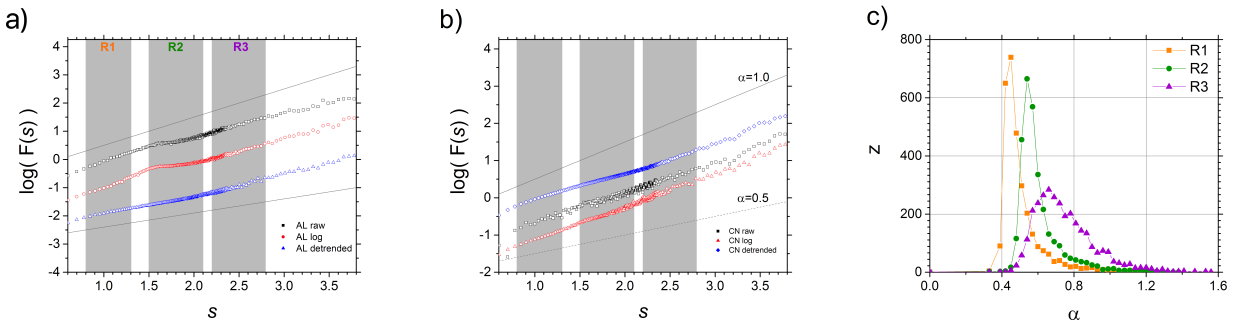

Figure 5.5.: **(a,b) illustrate the impact of transformation and filtering of raw time series on** $F(s)$. The black curve shows the fluctuation function $F(s)$ for raw access activity time series for Wikipedia page *Formula One*. The red curve shows $F(s)$ for the logarithm of raw time series and the blue curve shows $F(s)$ for detrended time series (weekly averages are removed) for a group of nodes in (a) and a single node in (b). (c) shows a comparison of the fluctuation coefficient α for different scales (R1, R2, and R3) as marked in (a) and (b) for the detrended time series.

correlated according to Eq. (5.3), $F(s)$ increases, for large values of s, as a power-law [28, 29, 122]

$$F(s) \sim s^{\alpha}, \quad \alpha = 1 - \gamma/2. \tag{5.4}$$

The fluctuation exponent α is calculated by a linear fit applied to a plot of $F(s)$ as a function of s on double logarithmic scales (see figure 5.5.a and 5.5.b).

For long-term correlated time series one can find $\alpha > 0.5$, and in the case of $\alpha = 0.5$ the data is uncorrelated. Especially for sparse time series, like event time series, one cannot apply DFA. This is why return interval statistics is applied.

5.3.3. Return Interval Statistics (RIS)

Return interval statistics (RIS) represents an alternative approach to identify correlation on multiple scales in time series. RIS is not as popular as the DFA method to study properties of non-stationary processes. Especially for sparse event time series it is possible to identify long range correlations by RIS. Long-term memory effects in dynamic systems are identified based on the analysis of return intervals between extreme events that exceed a given threshold. In case of a sparse event series, each occurrence of an event is defined by the time stamp directly, and in case of continuous time series we use a parameter q to define a threshold as shown in figure 5.1. and figure 5.6.

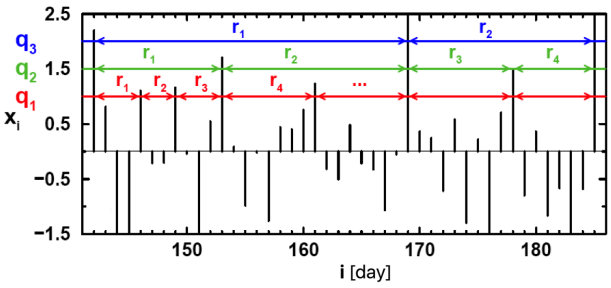

Figure 5.6.: **For three different thresholds q_i the statistical properties of the inter event times r_i are analyzed in RIS.** The figure was taken from Eichner *et al.* [126]

Depending on the properties of the underlying system the distribution of inter-event times can follow a power-law distribution, a Poisson distribution, a stretched exponential distribution or even a bimodal distribution like it was shown recently by an analysis of telecommunication data of human interaction events [127]. To describe the recurrence of events exceeding a certain threshold q, i. e., $x_j(t) > q$, one investigates the statistics of the return time intervals $r = t_2 - t_1 | x_j(t_1) > q \wedge x_j(t_2) > q \wedge x_j(t) \leq q | t_1 < t < t_2$ between such events at times t_1 and t_2. In uncorrelated time series ('white noise'), the return intervals are also uncorrelated and distributed according to the Poisson distribution,

$$P_q(r) = (1/R_q) \exp(-r/R_q), \tag{5.5}$$

where R_q is the mean return interval $\langle r \rangle$ for the given threshold q. For long-term correlated data, on the other hand, a stretched exponential distribution

$$P_q(r) = \frac{a_\gamma}{R_q} \exp[-b_\gamma (r/R_q)^\gamma] \tag{5.6}$$

has been observed [22, 23, 24] where the exponent γ is the correlation exponent from Eq. (5.6), and the parameters a_γ and b_γ are independent of q [24, 126]. In order to compare time series with different average inter-event times R_q the normalized distributions $P_q(r)R_q$ of return intervals r between events exceeding the different thresholds q have to be used.

RIS and DFA are related to each other. The next section shows how the exponents γ and α can be used for a comparison of the correlation properties of continuous and sparse time series, see also figure 5.7.

5.4. Multivariate Time Series Analysis

Multivariate time series analysis is applied to pairs or tuples of time series, e.g., to calculate link strengths for correlation networks. Such correlation networks are also called functional networks and describe functional or dynamic aspects of complex systems.

Many of the approaches used in this work are widely accepted. Anyway, it is very important to remember their specific advantages and also their limitations. Cross correlation (CC) or Pearson correlation should not be used as a universal approach of link reconstruction for functional networks without deeper investigation of raw data or calibration. Analysis of the features of the used data and a reasonable large number of time series pairs are required to identify significant functional correlation between or inside complex systems. Only linear dependencies can be identified using CC. Non stationarity, extreme events and cyclic patterns have also a negative impact on the results and can make them useless. In such cases - in general in case of non linear dependencies - it is useful to apply rank

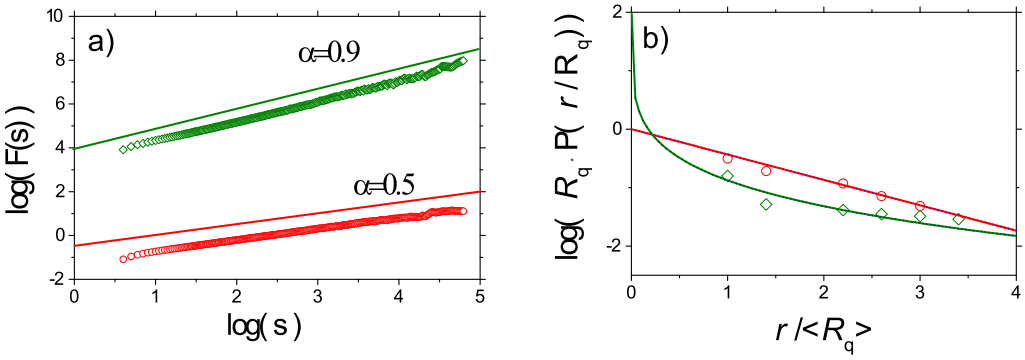

Figure 5.7.: **Comparison of RIS and DFA.** For two surrogate data sets (uncorrelated: red circles, correlated with $\gamma = 0.2$ according to Eq. (5.3): olive diamonds) results of both methods of long-term correlation analysis are shown: (a) DFA of order two, and (b) RIS. Lines indicate the theoretical behavior according to Eqs. (5.4)-(5.6) - red line: $\alpha = 0.5$ in (a) and simple exponential in (b), olive line: $\alpha = 0.9$ in (a) and stretched exponential in (b).

correlation, such as Spearman, and Kendall correlation. Another promising approach is the Event-Synchronization method. The shear variety of possible measures and possible combinations requires a generic analysis framework as described in the next part (the main part) of this work. Hence, a more specific discussion about link strength calculation and significance tests is provided in chapters 9 and 10.

In general, time series pre-processing is done to obtain stationary time series for further analysis (see section 5.2). Such transformed time series still represent the original process but not completely. In general, time series are never a complete representation of a process but always one with less dimensions compared to the original process. This is because information, e.g., trends are removed or because both, the sampling rate, and the length of the time series are limited for technical and practical reasons.

After detrending, analysis results will be related to the stochastic part of the process. If trends are extracted, they can be analyzed independently (see section 7.2). Trends can also be correlated. Separation of the deterministic part (driven by trends) and the stochastic part provides two different views but one has to be careful in this case, since not only stochastic components remain after detrending. If one aspect is removed from the raw data than the remaining data is only stationary, if no other aspect (no other trend on a different scale) coexists.

Especially if seasonal aspects and cultural aspects overlap, one can identify such trends on multiple scales, e.g., in the access-rate time series for the Wikipedia page about Formula One we could clearly identify overlapping yearly, weekly, and daily patterns.

Pereda *et al.* [131] use multivariate time series analysis in neurophysiology with the aim of studying the relationship between simultaneously recorded signals. They build on recent advances of information theory and nonlinear dynamical systems theory. This means various types of synchronization between time series enable them to assess the existence of nonlinear interdependence between signals. The concepts they evaluate are: phase synchronization, generalized synchronization, and event synchronization (see section 5.4.3).

In general one has to bear in mind that one does not know many details about the internal structure of the system. Thus, it would be wrong to expect only linear dependencies between the elements under consideration. Radebach [132] uses a nonlinear approach, called mutual information (MI) instead of Pearson correlation to define the association between nodes in climate networks. This also allows one to overcome the limitations of CC.

It is important to note that statistical methods, e.g. Pearson correlation and mutual information analysis (see section 5.4.4), in general reveal inter-dependencies. The correlation value obtained from Pearson correlation has a sign, but this sign is not related to the direction of the relationship between two time series. Contrary to this, event synchronization gives a synchronization strength Q and direction information q from which directed networks can be reconstructed.

5.4.1. Cross-Correlation

In case of a linear system, the input and output variables, represented as time series, are related to each other by a linear model $y(t) = m \cdot x(t) + n$. If such a linear relation exists between the two signals it can be analyzed by Pearson correlation analysis. The Pearson correlation coefficient was introduced by Pearson in 1846. Hauke and Kossowski [133] compare cross-correlation and rank correlation on the same sets of data. As mentioned already in the previous section, some good reasons exist to be careful in using this approach.

Here in our analysis framework we use Pearson correlation analysis in order to measure the similarity of two time series. A linear dependency with no time delay would lead to a high correlation value.

The co-variance function $R_{xy}(\tau)$ is defined as a convolution, which is a mathematical operation on two functions $x(t)$ and $y(t)$ with $\langle x\rangle = \langle y\rangle = 0$. For each value of τ (see s in Eq. 5.2 and Eq. 5.3) one obtains a number which expresses the overlap of both functions $x(t)$ and $y(t)$. The value of τ which corresponds to the maximum value of $R_{xy}(\tau)$ is a typical measure for a time delay between two signals.

$$R_{xy}(\tau) = (x * y)(\tau) = \int_{-\infty}^{\infty} x(t) \cdot y(t+\tau) dt \tag{5.7}$$

For comparison of two time series, and if no time delay is assumed, one simply calculates the co-variance value $R_{xy}(\tau = 0)$. This allows a quantitative interpretation of equality of the shapes of both time series. In case of periodic signals of the same frequency, it would be possible to measure a phase difference. The cross-correlation function for two continuous time series is defined as:

$$F_{xy}(\tau) = \frac{R_{xy}(\tau)}{\sigma_x \, \sigma_y} \tag{5.8}$$

In case of $x(t) = y(t)$ (two identical time series) the cross-correlation function is given by the auto-correlation function: $F_{xy}(\tau) = C(\tau)$.

In case of discrete time series the cross-correlation function is defined as:

$$F_{xy}(\tau) = \frac{1}{\sigma_{xy}(L-\tau)} \sum_{t=1}^{L-\tau} \Delta x(t) \Delta y(t+\tau), \tag{5.9}$$

with $\sigma_{xy} = \sqrt{\langle \Delta x(t)^2 \rangle \langle \Delta y(t)^2 \rangle}$. The parameter τ determines the time delay between both time series, and again $\Delta x(t) = x(t) - \bar{x}$ and $\Delta y(t) = y(t) - \bar{y}$.

It is important to note, that a high or low cross-correlation value does not allow a conclusion about causal relations.

Because the signals can vary heavily which causes different variances a normalization is applied in the cross-correlation function. To study the time dependence of the cross-correlation in very long time series one can apply sliding window techniques. Usually, overlapping time windows are used to get smoother results.

Furthermore, we implemented a slightly different approach when studying Wikipedia access data. First we split the time series into an ordered set of N non-overlapping episodes with length l. For this case we define the modified cross-correlation function $\varphi_T(0)$ as (see also [134]):

$$\varphi_T(\tau = 0) = \frac{\sum_{t=1h+24h\cdot T}^{24h(1+T)} \big(x(t) - \overline{x}\big) \cdot \big(y(t) - \overline{y}\big)}{\sqrt{\sum_{t=1h+24h\cdot T}^{24h(1+T)} \big(x(t) - \overline{x}\big)^2} \sqrt{\sum_{t=1h+24h\cdot T}^{24h(1+T)} \big(y(t) - \overline{y}\big)^2}} \tag{5.10}$$

T is the index for the day and $\tau = 0$ means that no time delay between the two signals is considered. For daily analysis $l = 24$ hours and the number of elements in the set is $N = 7$ for one week or $N = 30$ for one month. Instead of simply averaging the cross-correlation coefficients for all episodes in each set, the median value is taken from the sorted list of coefficients in the particular group to reduce the effect of outliers. More details can be found in [134].

5.4.2. General-, Spearman-, and Kendall-Correlation-Coefficient

In many cases, especially if not enough data is available and if the value pairs are not taken from a bi-variate normal distribution it is not possible to identify the correlation between two time series in a reliable meaningful way by using Pearson correlation. This is why two alternatives, Spearman rank correlation, and Kendall rank correlation, are important for correlation analysis in large complex systems.

Spearman correlation (SC) is a non parametric measure of rank correlation. This means instead of calculating the statistical dependence between the values of two time series, the rankings of values are compared. In this way SC is like Pearson correlation of the rank series. As a consequence, SC can be used for data series with non linear dependencies and for categorical values as well. High positive or negative values of the Spearman correlation coefficient indicate a monotonically increasing or decreasing dependency between the two variables.

Another specific correlation measure is called Kendall-Rank correlation. The Kendall correlation only compares, if the values of discrete observations are equal or not. This means, it can not be applied to continuous variables or only after an appropriate discretization. But finally, both, Spearman- and Kendall-Correlation are special cases of a *general correlation coefficient*, defined by Kendall in 1944.

Here we use the Pearson correlation (and variations of it) as similarity measure and for calculating functional links from time series data which define networks for several time scales at discrete points in time.

The role of rank correlation in network dynamics: Using traditional network analysis algorithms we assign to each network node a value which describes its importance - called rank of a node. Now, using the rank correlation methods we can identify how stable the system is over time. The entire system has an internal structure, represented by the links. Links can change very fast and a changed degree distribution function reveals a changed internal structure. But could we conclude the opposite as well? Is a stable degree distribution an indicator for a static network? This is not the case. Links can fluctuate internally, and the overall structure is the same over many time steps, even if multiple nodes contribute differently. By tracking the rank of nodes beside the degree distribution we have a second representation for the entire system.

Finally we are able to describe the time evolution of the entire system by two time series: one for rank correlation coefficients, and a second for parameters which describe the degree distribution at each point in time where the network was obtained from the activity time series of the corresponding time interval. Even in cases of a systematic change in the environment around a network the internal structure and the node ranks should not change significantly if the network is not affected by the external influence. Changing node ranks indicate a response of the system to any existing stimulus.

5.4.3. Event Synchronization

In general, correlation analysis is used to quantify how well two processes are aligned to each other in the time domain. Depending on the nature of the processes one has to study individual pairs of events or one can use simplified representations, such as periodical functions. Hence synchronization can be detected if for two processes coincidence in timing or phase can be measured.

Event series are usually not used to describe periodic processes. However, *event synchronization* (ES) compares the time delays between pairs of events from two different processes, especially if no periodicity exists in both. ES is a simple and efficient algorithm, which operates on pairs of sparse event time series[4]. Quiroga *et al.* [135] introduced ES in 2002 to study neural spike trains in EEG signals. Malik *et al.* applied this method to data about monsoonal rainfall [136, 137]. Another approach for synchronization analysis is given by Mörtl *et al.* [113]. They use events as anchoring points to segment trajectories in order to improve synchronization modes. In this way, they create a unified view. This view consists of continuous phases and discrete events.

One of our sub projects was done in order to investigate properties and features of this promising method before we use it to reconstruct correlation or similarity networks. The first experiment simulates a transition between identical and independent time series, the second one covers the case of different event densities. Results of conducted experiments will be discussed in section 10.5.4.

Motivation for event synchronization for Wikipedia analysis: Collaborative processes like the Wikipedia edit process are the results of self-organization. The content creation process has a different nature if compared with the crowd-based information access process, such as usage of Wikipedia pages by users and online or mobile applications. One reason is, that both processes have a different audience. Changes to pages are not done so often, in many cases less than once per month. Also the number of Wikipedia editors is not as big as the number of readers or indirect users like automatic tools which have Wikipedia clients build in. Because of this, the edit event time series are rather sparse and methods for dense or continuous time series like cross-correlation or rank-correlation can not be applied.

To track the evolution and to compare the properties of Wikipedia pages at several discrete times, we implemented a data structure called *event time series* (ETS). Such an ETS object contains the indexes or the time stamps. Figure 5.8.a and 5.8.c show examples for low and very high event densities. Both, Wikipedia edit-event time series and Wikipedia access-rate time series are the results of aggregation procedures. ETS can be visualized in different ways, e.g., as shown in figure 5.8. The number of detected events during each measurement interval or at each time step is shown in the top row. Because a useful data set inspection is not possible with this representation a second plot is provided. The bottom row shows the cumulative event number (also called event index) as a function of time. This means that the maximum value equals the total event number. Aris *et al.* [112] use a comparable approach called *event index* and *interleaved event index*. During creation of ETS from complete event collections some information is lost due to aggregation. Note that one can not reconstruct the original event collections from ETS data structures. If multiple events occur at the same time or during the narrow time window (which expresses the time resolution) we count the number of individual events in each window (see figure 5.9.a and figure 5.1). Only the time stamp or the index of windows for which events have been registered are stored. The time resolution used in common software solutions is milliseconds because this is also the default resolution of the timer in common computer systems. In our study we use hourly resolution for practical reasons.

In case of high event density or high access-rates the series look more like discrete continuous time series. A threshold based transformation, as illustrated in figure 5.9.a, can be applied to any time-dependent function to create sparse event time series for faster processing.

[4] Aris *et al.* ([112]) discuss the advantages and disadvantages of four different representations of unevenly-spaced time series, especially in the context of visualization. Sampling and aggregation reduce the required memory but introduce errors. Time differences are not represented in case of time indexes, and the time axis is partially stretched by interleaved event indexes.

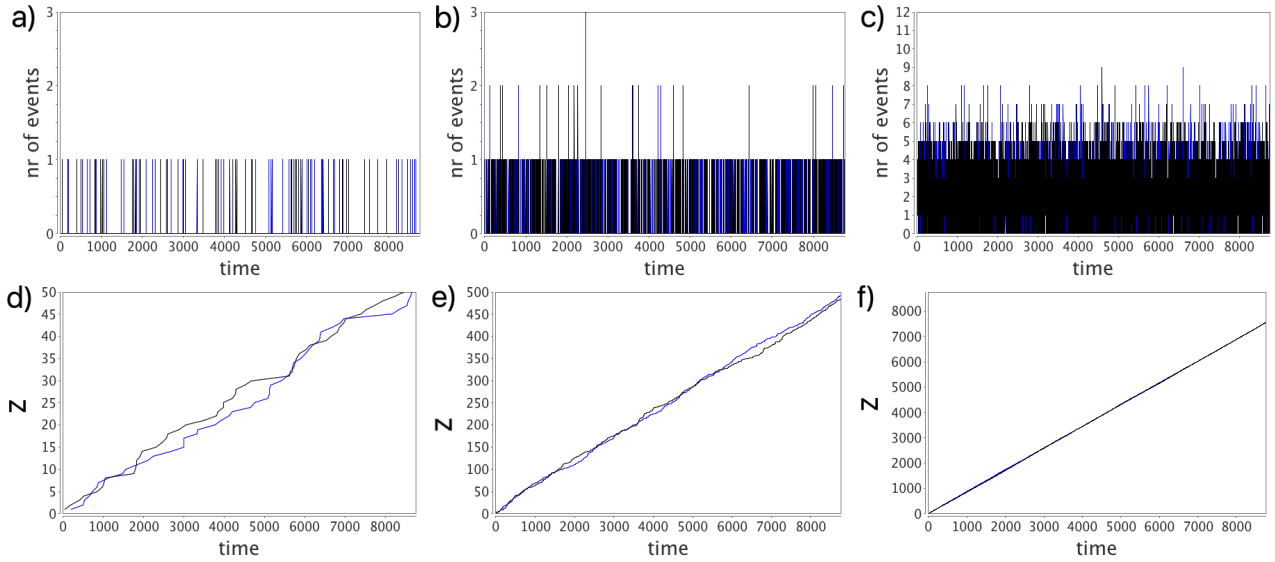

Figure 5.8.: **Visual comparison of two representations of event series**. Visualization of two exemplary time series (in different colors) with low (a,d), medium (b,e), and high event density (c,f) support decisions about necessary pre-processing methods and possibilities for application of specific analysis algorithms. (d,e,f) show the cumulative interval count for time intervals with 1 or more events (called z) as a function of time. For sparse ($\rho = 0.006/h$) event series the curve deviates stronger from a straight line (d) than for the ETS with 10 times higher density $\rho = 0.06/h$ (e). The ETS with high density ($\rho = 0.6/h$) (f) can be handled like a discrete time series of a continuous variable. In order to apply ES analysis one has to provide a sparse ETS by using an event filter procedure and an appropriate threshold as shown in figure 5.9.a.

The values of such dense series can be filtered with simple filter rules, e.g., if for a given pair of two consequent values the first one is below and the second one above a certain threshold q one can interpret this as an event which happened at time $t_1 + (t_2 - t_1)/2$. This filter approach is also used to define the inter-event times dependent on the threshold q for RIS (see section 5.3.3). Instead of using a threshold one can interpret the change of the sign as an event as well. ES analysis can easily be applied to data series if time stamps are already given in form of indexes. Otherwise interpolation is required to define the time stamp of an event. Validation of results always requires a transformation back into a human readable time representation. We found, that a reasonable way to handle the time stamps is working with a time offset, defined as the time since a well defined time t_0. An example is shown in figure 3.c in [4]. There we present the measured data with a numerical equidistant index, which represents the week since January 1-st, 2004.

The Algorithm

According to Quiroga *et al.* [135] i and j are two event time series with events occurring at times t_l^i and t_m^j. s^i and s^j are the total numbers of events in these series. The indexes l and m range from $1...s^i$ and $1...s^j$. In order to find a quantitative measure which describes, if two series are synchronous based on co-occurrence of events. We calculate how often an event in time series i precedes an event in time series j, or how often an event in time series j precedes an event in time series i.

A time lag τ_{lm}^{ij} is used to quantify closeness or co-occurrence of two events. Figure 5.9.b illustrates the definition of τ_{lm}^{ij} which is calculated as:

$$\tau_{lm}^{ij} = \frac{1}{2} \min \left(t_{l+1}^i - t_l^i, t_l^i - t_{l-1}^i, t_{m+1}^j - t_m^j, t_m^j - t_{m-1}^j \right) \tag{5.11}$$

For different application contexts one can consider different definitions of τ. Using a constant global time lag τ_g is an alternative, suggested by Quiroga *et al.* [135]. It can be defined as a minimum or average of all τ_{lm}^{ij}, keeping in mind that the parameter is meant to avoid double-counting and thus should be sufficiently small.

Depending on the calculated difference between t_l^i and t_m^j, especially if it is less than τ_{lm}^{ij}, both events are considered synchronous. We calculate the contributions for all event pairs J_{lm}^{ij} as:

$$J_{lm}^{ij} = \begin{cases} 1 & \text{if} \quad 0 < t_l^i - t_m^j < \tau_{lm}^{ij} \\ \frac{1}{2} & \text{if} \quad t_l^i - t_m^j = \tau_{lm}^{ij} \\ 0 & \text{else} \end{cases} \tag{5.12}$$

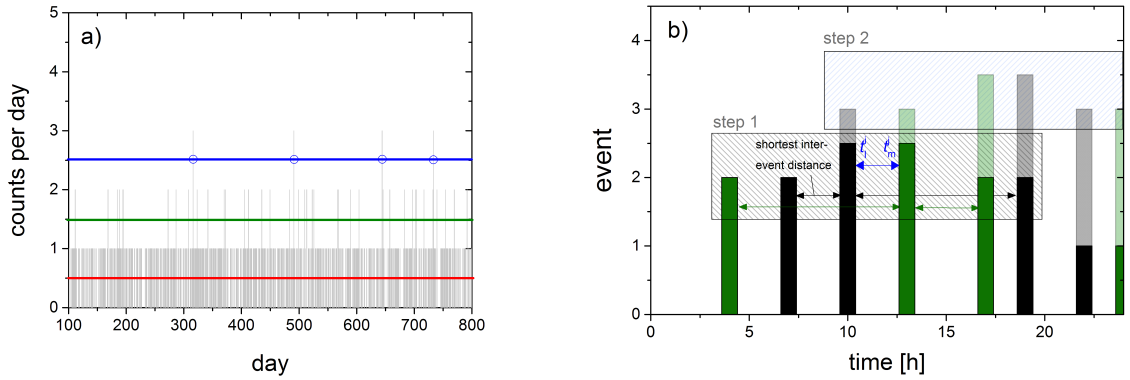

Figure 5.9.: **Creation of event time series (ETS)**. Sparse event time series are results of an aggregation of spontaneous events or of a transformation. (a) shows the threshold based transformation of a continuous time series into an event time series. Spike trains as explained in Quiroga *et al.* [135] and variables used in Eq. (5.11) are illustrated in (b). Events of time series i are shown in black and for series j in green.

Because t_l^i must occur either before or after t_m^j we can conclude that $J_{lm}^{ij} + J_{ml}^{ji}$ must be 0 or 1. This takes into account that all event pairs can be either asynchronous or synchronous.

For all pairs their individual contribution to synchronicity or asynchronicity J_{lm}^{ij} is aggregated as $c(i|j)$:

$$c(i|j) = \sum_{l=1}^{s^i} \sum_{m=1}^{s^j} J_{lm}^{ij} \tag{5.13}$$

We repeat this procedure to calculate also all contributions $c(j|i)$ to cover both possible directions. The first result is Q_{ij}. It represents the strength of event synchronization and is calculated as the sum of $c(i|j) + c(j|i)$, while subtracting $c(i|j) - c(j|i)$ gives us the value q_{ij}, which represents the directional strength of the coupling between both series:

$$Q_{ij} = \frac{c(i|j) + c(j|i)}{\sqrt{s^i s^j}} \tag{5.14a}$$

$$q_{ij} = \frac{c(i|j) - c(j|i)}{\sqrt{s^i s^j}} \tag{5.14b}$$

To normalize the values we divide both by $\sqrt{s^i s^j}$, then the event synchronization strength Q_{ij} ranges from 0 to 1 where the maximum can only be reached if $s^i = s^j$. The delay value ranges from $q_{ij} = -Q_{ij}$ to $q_{ij} = +Q_{ij}$. If $Q_{ij} = 1$ both series are completely synchronized because for every event in time series i, there is a synchronous event in j. $Q_{ij} = 0$ means that the series are completely desynchronized. The delay describes which of the series is leading, e.g., if $q_{ij} = -Q_{ij}$ for all pairs of synchronous events the events in the first time series i precede the events in series j. If $q_{ij} = +Q_{ij}$ the second series is leading, and in case of $q_{ij} = 0$, none of the series can be considered to be leading or influencing the other time series.

5.4.4. Mutual Information and Mutual Correlation

Non-linear dependencies between two time series $x(t)$ and $y(t)$ - if such exist - can not be be determined by Pearson correlation analysis. In case of existing non-linear dependencies Pearson correlation would indicate no correlation or a strongly decreased correlation value, especially for long time series. One needs a different approach, e.g., the concept of *mutual information* (MI). MI is based on the Shannon entropy H calculated for the joint probability density function $P(x, y)$ and the individual probability density functions $P(x)$ and $P(y)$. Kraskov *et al.* [138] describe the mathematical properties of MI and Lizier [139] provides an implementation of several time series algorithms with information theoretical motivation including MI.

The strong prerequisites of Pearson correlation (bi-variate normal distributed values) is not required for MI. Instead, for each pair of time series the values are binned into B ranges called bins. MI calculation can be applied as a sliding window technique for episodes of different length. It gives a value between 0, for independent time series and $log(B)$ for time series which completely share all information (or between 0 and 1 in case of normalized MI). MI for a particular time window contains no information about a direction or a delay. For that kind of

analysis, a different approach is used, e.g., the *transfer entropy* (TE) because this approach also takes the history of signals into account. TE is also included in the JIDT software package (see Lizier [139]). *Conditional mutual information* (CMI) allows one to measure the influence of a third element on the relation between two elements. CMI was introduced by Wyner [140] already in 1978. CMI is conceptually related to the measure of partial correlations between pairs of stocks in the presence of a third variable, e.g., the stock index to which both belong to as introduced by Kenett *et al.* [23].

5.5. Statistical Tests for Probability Density Distributions

This section summarizes the properties of used statistical tests. We need such statistical tests to identify if two ensembles of random variables are of the same distribution. This is called a two sided test, because two sets of measured data are involved. A single sided test identifies if a measured distribution is significantly different from a given distribution.

The Kolmogorov–Smirnov test (also known as K–S test or KS test) is used to compare probability distributions. Usually one uses a sample probability distribution – from measured data - and a reference probability distribution – which can be also obtained from measurements or simply be calculated from a known formula which describes the probability distribution function. Because the test is nonparametric, it is not required to calculate the probability distribution functions first. The core idea of the test is the Kolmogorov–Smirnov statistics, which quantifies a distance between the two distribution functions (sample and reference, see also [141]). The null hypothesis is that both distributions are drawn from the same or the reference distribution respectively. There is no restriction to a particular shape of the PDFs which are tested, but the KS test works with continuous, one-dimensional probability distributions only. According to Wikipedia: *"The two-sample K–S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples."*

A more specific test for testing for normality is the Shapiro–Wilk test. In this case, the null-hypothesis is that the population is normally distributed. One has to choose an α level (also called significance level) to compare a calculated p-value with. The null hypothesis is rejected if this p-value is smaller than the previously chosen alpha level. This means now, that there is evidence that the used sample was not normally distributed. The null hypothesis - the data was normally distributed - cannot be rejected if the p-value is greater than the previously chosen significance level. The Shapiro-Wilk test is biased by the sample size (see [142] p.143) this means one might find statistically significant results for any large samples. Practical implementations are often limited to 5000 values. This has an implication on the selected window size or episode length. One has to have in mind that, e.g., stock market prices are not normally distributed. In this case an additional transformation is required (see figure 5.3).

If we have to test really huge samples - for example for a goodness of fit test - it is possible to use a modified Kolmogorov–Smirnov test for testing for normality of the distribution (see also [143]). One has to standardize the samples (subtract mean value and divide by variance) and the result of this transformation can be compared with a well defined reference distribution (the standard normal distribution in this case). To save some computational resources one can also modify the reference distribution by setting the parameters mean and variance according to the values derived from the sample. Because it is known that such a modification changes the null distribution of the test statistic one can use existing tables [144] (see pp.117-123).

Furthermore it is important to remember that a p-value is not a measure to quantify how well a distribution could be compared with a normal distribution. It is also not possible to find out, which one of a set of distributions fits better. In order to achieve this, one has to vary the alpha level so that for the one distribution the null hypothesis has to be rejected but not for the other. For this significance level the one is a better choice than the other, but this again is a binary decision and not telling anything about the distance between both.

Based on Monte Carlo simulations Razali *et al.* [145] have found that the Shapiro–Wilk test performs best followed by the Anderson–Darling test in a comparison of Kolmogorov–Smirnov, Shapiro–Wilk, Anderson–Darling, and Lilliefors tests.

5.6. Surrogate Data for Functional Tests and Significance Tests

Surrogate data can be entirely artificial or it can be derived from real data. Usually, randomized data series are created based on specific assumptions. Transformations, which change specific properties of the data in a well controlled way, to allow a systematic comparison of analysis methods applied to the surrogate data and the real data are also common. Thus, surrogate data is typically used for cross-validation and plausibility tests in addition to analysis of measured real world data.

In this work we use two different types of surrogate data. The first type is based on random number or semi random number generators (RNG). The RNG reproduces properties of the time series as obtained from a real world system. In order to be compatible with the real data, one has to assure, that surrogate data shows similar

properties as the real data. For this purpose one has to measure characteristic properties using univariate analysis and calibration procedures. Relevant properties are mean value, variance, distribution of values, auto-correlation and long-term correlations. Another important property is fractality but in this work the fractal nature of time series was not analyzed. In general a model of the required time series is useful, but since time series modeling is not trivial we prefer the simpler approach, based on the above mentioned parameters which can be extracted from measured data.

A second type of surrogate data is generated directly from real data by randomization, either by shuffling the values or by modification of the Fourier transform of the original time series. Simple shuffling is usually repeated in multiple runs. The assumption is hereby that shuffling destroys all kinds of correlations but the distribution of values, mean, and variance are conserved. In this way, a comparison of analysis results from real data with results from shuffled surrogate data allows identification of non-random effects in measured data sets which are caused by short-term and long-term correlations within and between the time series.

Theiler *et al.* [146] tested for nonlinearity in time series using surrogate data. They specify a linear process as null hypothesis first. Using the *phase-randomization method*[5] they generate surrogate data series. Since these series are produced such that they are consistent with the null hypothesis they calculate a discriminating statistic for the original series and for all surrogate data series. They reject the null hypothesis - and detect nonlinearity in the data series - if the calculated values for the original and the surrogate data set are significantly different.

5.6.1. Modeling Properties of Random Numbers

For simple calibration experiments, reference models are used to create time series with comparable features or properties for statistical tests. The analysis algorithms are then applied to such semi-random data in order to measure the influence of patterns on specific features measured by certain algorithms. Common patterns are: (a) sine-waves with a given phase ϕ, amplitude a, and frequency f, (b) single spikes of a given height a, or (c) plateaus and increasing or decreasing trends between fixed levels $[a_{min}, a_{max}]$ of given width w.

In chapter 10.5 we investigate the impact of single peaks and long-term correlations in individual time series on the correlation properties for pairs of time series. Based on such a specific calibration we can define optimal significance levels depending on the specific data used.

A set of random numbers can be used as surrogate data for statistical significance tests. Important properties include the parameters which describe the distribution of the values, and the correlation properties. This means, short-term correlation measured by the autocorrelation function as well as long-term correlations, which are detected via DFA, have to be produced by the random series generator.

Hence, random numbers are produced or generated differently for several applications. Sometimes one wants real randomness, e.g., for application in cryptography. Pseudo random numbers are useful for engineering since they can be reproduced if the same conditions exist, e.g., if the same seed value is used together with the exact same implementation on the same platform.

First, a random number generator has to create a large number of uncorrelated uniformly distributed scalar random values which is not trivial. How random the numbers really are, depends on several properties. In order to reflect fairness in a game with fair dices, one needs evenly distributed integer values, in the range 1 to 6. This means that the probability of each of the six outcomes is equal to $\frac{1}{6}$. Auto-correlations should not exist, because in the presence of auto-correlation a higher probability for a particular number exists, depending on the history. Uncorrelated values are not influenced by the history.

Especially if large systems should be simulated over a long time period one has to use a random number generator with a long period instead of traditional random number generators built in into every computer system or programming language. In 1998 Matsumoto and Nashimura [149] published the first random generator with a period $(2^{19937} - 1)$ which is longer than the estimated number of electron spin changes since the beginning of the universe (10^{6000} vs. 10^{120}) [150].

Uniformly distributed random numbers are used to generate new samples of uncorrelated random numbers with any possible distribution (see fig. 5.10). Shaping the distribution function in this way influences the mean value but not the correlation properties (see fig. 5.10.b and 5.10.c). In order to change the correlation, the frequency spectrum has to be manipulated. A manipulation of the Fourier coefficients introduces auto-correlations. Details about the Fourier Filtering Method (FFM) are explained in section 5.6.2.

The following methods are typical means to modify properties of random number series: **(1)** *Random shuffling* destroys correlations while the distribution of values (the shape of the PDF), mean and variance are not changed. **(2)** *Amplitude-transformation* by a frequency dependent factor (see also section 5.6.2) creates long-range-correlation, and changes the distribution, mean and variance slightly. **(3)** *Phase-randomization* destroy the fractal properties of the time series.

[5] According to [147] (last section of chapter 2, p.8) the phase randomization procedure creates surrogate data with the same correlation properties as the original signal. Following their procedure one performs a Fourier transform on the original time series, which preserves the Fourier amplitudes but randomizes the Fourier phases. Finally, one performs an inverse Fourier transform to create the surrogate data series (see p. 48 in [147].

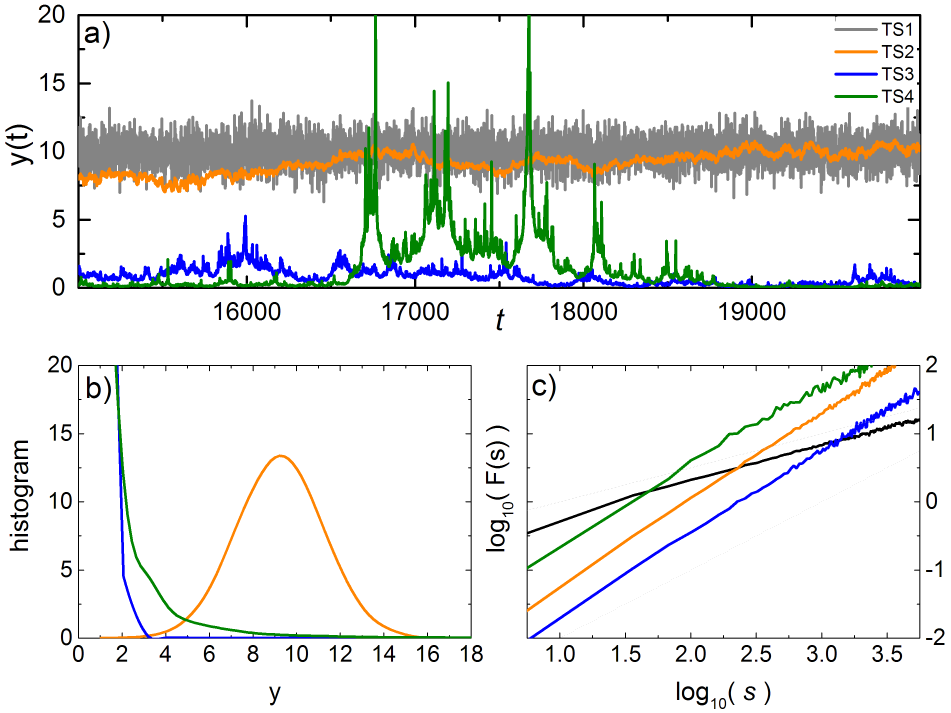

Figure 5.10.: **Example time series:** white noise (gray), long-term correlated series (orange, green, and blue) are shown in (a). Histogram of value probability density functions (PDF) for sample data sets from random number generators (RNG) implemented in the open source software packages Apache Commons Math [148] and Hadoop.TS [11] are shown in (b). By using FFM (see section 5.6.2) also the distribution of values is influenced which has to be corrected by rank-ordering. Shuffling has no influence on the distribution of values. In case a sliding window technique is applied the distribution of values per time window is influenced by global shuffling, especially if time series are non-stationary.

5.6.2. Generation of Long-term Correlated Random Numbers

The Fourier filtering method (FFM) is a very common method for generation of sequences of random numbers with power-law correlations.

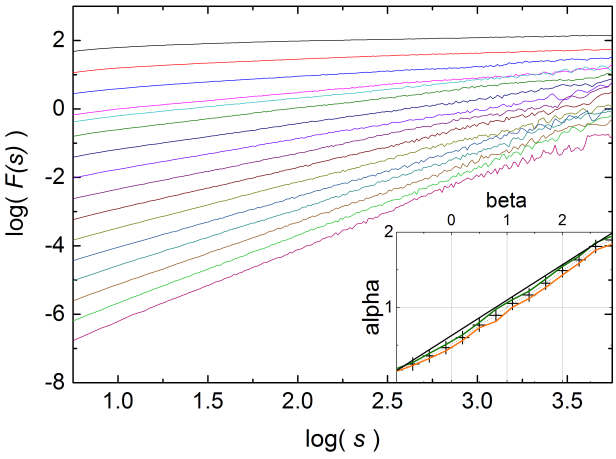

Figure 5.11.: Fluctuation functions for series of random numbers created with the Fourier filtering method (FFM). The fluctuation coefficient α was calculated by a linear fit in the log-log plot of the fluctuation function $F(s)$ in the range $0.8 < \log(s) < 2.5$ and is shown in the inset. The black line shows the theoretical relation $\alpha = \frac{1+\beta}{2}$. The black crosses show the average from 100 iterations and the green and orange curves illustrate the $\pm\sigma$ band.

According to Makse *et al.* [151] this method has the disadvantage of presenting a finite cutoff in the range over which the variables are actually correlated. Thus the FFM is not suitable for the study of scaling properties in the limit of large systems.

A modified Fourier filtering method was published by Makse *et al.* in 1995. Due to the modification, the cutoff in the range of correlations is removed and the actual correlations extend to the whole system in the modified method [151].

Figure 5.10.b shows PDFs for four random number series (TS1,TS2,TS3,TS4) generated by a RNG from the Apache Commons Math package [148]. Because such values are uncorrelated the slope in figure 5.10.c is $\alpha = 0.5$ for TS1. TS2, TS3, and TS4 were modified by our implementation of the Fourier filter method [11]. The distribution of values has been changed for TS3, and TS4. The important difference is that the fluctuation coefficient which indicate long-range correlations is now larger than 0.5 as shown in figure 5.10.c. In [152] Schreiber and Schmitz proposed an iterative method to correct deviations in spectrum and value distribution. The surrogate data is filtered towards the correct Fourier amplitudes and rank-ordered[6] afterwards to the correct distribution in an alternating iterative approach which stops after a finite number of steps. The remaining difference can be interpreted as the accuracy of the method. Such long-term correlated random numbers are used in this work for functional testing and validation of the DFA implementation.

5.6.3. Creation of Networks with a Given Degree Distribution

So far, all considerations regarding surrogate data generation were related to time series. Network generation was briefly mentioned in chapter 3. Several software packages for network generation based on well known network models exist. I close this section with a specific network creation method, introduced by Chung and Lu [153]. Their approach utilizes the idea of a hidden variable to shape the degree distribution of a random network. First, to each node a random value d drawn from a distribution P_{init} is assigned. In a second step, a selector function $f(d_i, d_j)$ is used to define the expected topological structure of the resulting network. But one has to be careful, because only if the degree distribution function decays fast such networks are uncorrelated (see [39] p.35).

[6] Rank ordering allows a reorganization of values in a time series in order to introduce long-term correlations. After phase Fourier filtering, not only the phases are modified also the amplitudes of the inverse Fourier transform are different from the original distribution. The positions of the values determine correlation properties. Thus, we have to bring the highest value of the original series to the position of the highest value in the transformed series. The second highest value of the original series has to be placed at the same position as the second highest value in the manipulated series an so forth. In this way, we introduce the same correlations but do not change the original distribution. Instead of aligning the values directly, the ranks of the values are used to reorder the original series.

Part II.

Method Development and Enhancement

This part consists of seven chapters which cover improvements of existing methods. Furthermore, some novel methods for interdisciplinary complex systems research are described. The data sets for this part were primarily selected for illustrative purposes. Application of the methods in a particular scientific context is presented in part III and as an appendix in part IV of this work.

Overview

> Experience without theory is blind, but theory without experience is mere intellectual play.
>
> *(Immanuel Kant)*

Graphs and networks are common representations of complex systems. Statistical properties of an ensemble of nodes are calculated from individual properties of many network nodes. Structural graph properties are a special kind of properties of the whole ensemble of nodes and links. They cannot be measured or calculated from individual elements. Such structural properties can be presented as individual node's properties, such as node degree or membership in a clique or cluster. In order to get values for those metrics, a graph analysis procedure has to be applied. A graph can be seen as an individual entity on its own, consisting of nodes with time-dependent properties, which for itself define time-dependent link properties. Finally, individual link groups describe special aspects of real world phenomena.

An important task in complex systems research is modeling of dynamic properties. Our approach is based on the correlation between node properties, network structure, and link properties. Not all properties can be measured at once nor directly. Technical and economical limitations are addressed by recent technological improvements. Inexpensive distributed storage, linked data sets, and massive parallel processing provide the technical frameworks for next generation network analysis.

Properties of network nodes, especially time-resolved properties captured in time series are used to calculate correlation based links between such nodes. The purpose of such correlation analysis is recreation or detection of hidden links between nodes. Such links can be both, cause and result of interactions between nodes. One way to model such processes is to form link layers, one layer per process. In this step, a decomposition of the system is done. One may loose information during this step, but for simplification it is necessary. Later on, after, the correlation networks are known, a combination of several layers is used to draw a complex network again. One has to decide if a horizontal or a vertical cut should be applied. As horizontal cut we consider creation of layers which consist of links of the same type. Vertical cuts contain different link types around a well defined, usually small number of nodes. A comparison of types of networks shows, if the decomposition has a strong influence or if it could even be neglected. Structure-induces stress (SIS, see chapter 12) is a novel approach to quantify the impact of functional networks on the underlying network structure.

Network links can directly be extracted from existing data sets, if such links are explicitly defined. Such links are often described as hyperlinks in HTML documents, citations in research articles or books, or even semantic annotations[7]. Furthermore, citation networks[8], co-authorship, co-occurrence of terms, or the grammatical structure of human language are useful sources for extraction of structural information.

A second type of links has to be reconstructed from node properties available as time series data. Because this connection is not explicit, one has to start with an assumption or a hypothesis, which describes why a correlation between both entities exists and how this dependency can be expressed by a function, and thus, how data analysis can reveal this hidden link. An indirect connection between two systems (sub-systems, or elements) can be defined by an external system which has an influence on both, either instantaneous or with a time delay.

A very simple approach is the calculation of a correlation matrix for all possible pairs of network nodes within an ensemble. Pearson correlation is a common technique, but it cannot be used in the case of sparse time series

[7] Semantic annotations are available, e.g., in online documents in one of the formats called *microformats* or JSON-LD.
[8] A citation map as presented in [154] shows also the geographical embedding of scientific work.

or if the values in a time series are not from a Gaussian distribution. In such cases, event-synchronization is an alternative approach.

In many cases, one is interested in a comparison of the underlying structural network (the explicit network) and the functional network (the implicit one). The full system, e.g. all messages from all people using a particular communication service, such as E-Mail, Twitter, Facebook, or others is not available publicly. Even if accessing all that data would be allowed, the amount of data is too much and it is not possible to handle all that data with traditional techniques.

In case of our Wikipedia studies, we would have to process the time series of about $8 \cdot 10^6$ pages. A time series bucket with access-rate data on hourly resolution for one year would require about 261 GB. It is not required to have all that data in memory at once. Network links are calculated from time series pairs. But how much memory would be required to store the full correlation matrix with a delay of $\tau = \pm 14$ days, which is two weeks? In this case each link would require 116 bytes. The un-directed network has around $32 \cdot 10^{12}$ links and requires ≈ 172 PB to store the network dynamics on weekly resolution for a whole year.

Our approach aims at reducing the amount of data during individual analysis steps while comparability is always given. As long as the impact of a simplification can be estimated, one can decide if the introduced error is finally acceptable or not.

6. Embedded Context Graphs

> Divide each difficulty into as many parts as is feasible and necessary to resolve it.
>
> *(Rene Descartes)*

Decomposition of interconnected systems helps to simplify but at the same time it causes information loss. Researchers have to analyze carefully, how the system in the research focus is related to other systems and how internal components are related to or connected with each other. In the case of the World Wide Web, which is formed by hyper-linked pages, it was easy in the beginning. The number of such pages was still countable and maintainable, but soon the amount of content was too large. Nowadays several online systems compete. They offer comparable functionality to the same still growing audience. Studies on large social media applications reflect segregation effects as observed in social communities. The probability of becoming a member of an online community is affected by age, location, or the number of friends, which use a certain system already. The book *Networks, Crowds, and Markets* by David Easley and Jon Kleinberg [155] describes in chapter 4 the *interplay between selection and social influence*. These concepts are applied to interacting social entities - to human beings, but they can be generalized to social content networks and hybrid systems as well. In this way, the content, created by social communities, reflects properties of the community and can be used as a stub for analysis. This is one of the main reasons for using Wikipedia data in this work.

6.1. Define Subgraphs in Interconnected Networks

How can we define the focus and the neighborhood of a complex social network? A simple approach is presented by Dorogovtsev *et al.* [39]. Figure 6.1.a is an illustration taken from [39] (figure 6). It shows the embedding of nodes in a directed graph. In general, if edges are undirected, such a network consists of a giant weakly connected component (GWCC), which is also called *percolating cluster* and several disconnected components (DC). In case of directed networks the GWCC consists of: (1) a giant strongly connected component (GSCC); (2) the giant out-component (GOUT); (3) the giant in-component (GIN); and (4) the tendrils (TE).

The GSCC is the set of nodes which is reachable from every node by a directed path. GOUT is the set of nodes approachable from the GSCC by a directed path and includes GSCC. GIN contains all nodes from which the GSCC is approachable and includes GSCC. TE form the rest of the GSCC. These are also disconnected nodes which have no access to the GSCC and are also not reachable from it. In this definition GSCC is the interception of GIN and GOUT. This allows a formal notation of the system N as in [39]. We write:

$$N = GWCC + DC \tag{6.1}$$

and

$$GWCC = GIN + GOUT - GSCC + TE. \tag{6.2}$$

As of May 1999, the entire Web, containing $203 \cdot 10^6$ pages, consisted of the GWCC, $186 \cdot 10^6$ pages (91% of the

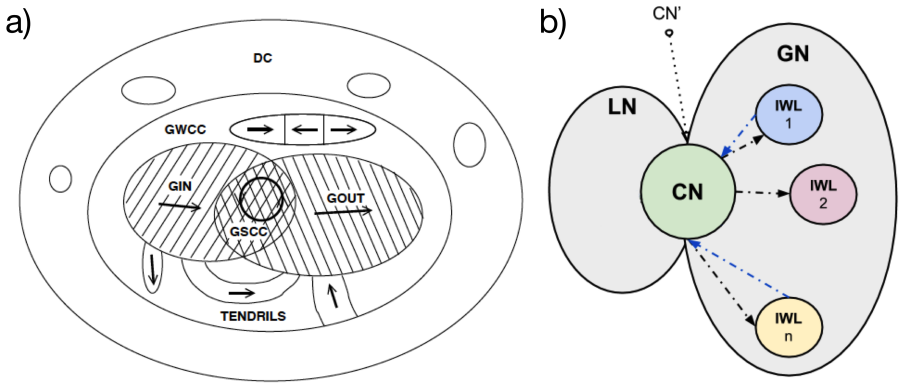

Figure 6.1.: **Context and neighborhood of a semantic concept.** A general description of components in a complete network (taken from [39]) is shown in (a). Not all resources of large systems can be accessed at once (because of technical limitations or due to the usage of parallel algorithms), a localized approach as presented in (b) is required. In (a) GSCC contains all nodes which are reachable from every node by a directed path. GOUT contains all nodes approachable from the GSCC. GIN contains all nodes from which the GSCC is approachable. GIN and GOUT include GSCC, and TE contains the rest of the GWCC, especially all the disconnected nodes which have no access to the GSCC and are not reachable from it. (b) Shows a systematic approach to define local networks, which set an analysis scope. The central node CN is a single page about a company, or a project, or it is a *'list-page'*, which bundles links to groups of pages, e.g. the page about a stock market index. CN has links to and from the local neighborhood (straight arrows) defined by pages in the same language (group LN). Inter-wiki links usually connect CN to many IWL pages in different languages (dash-dot arrows) covering the same topic or semantic concept. Those pages contain links to their local neighborhood in the same language (here not shown as arrows) which all contribute to the group GN. This way, all pages around all IWL pages define the global neighborhood GN. Some concepts are linked via redirect links or from so called disambiguation pages by specific links. The sketch shows those as CN' because of their strong relation to the semantic core.

total number of pages), and the DC, $17 \cdot 10^6$ pages. In turn, the GWCC included: the GSCC, $56 \cdot 10^6$ pages, the GIN, $99 \cdot 10^6$ pages, the GOUT, $99 \cdot 10^6$ pages, and the TE, $44 \cdot 10^6$ pages.

The size of the Internet, ten years later, in 2009, was estimated by M. Zillman [156] in *'The Deep Web'* report. It covers about 1 trillion pages of information located through the World Wide Web in various files and formats that the current search engines on the Internet either cannot find or have difficulty accessing. According to [156] search engines can find only about 20 billion pages at this time (in 2008).

Nowadays, a definition of the size of the internet would even be harder. What is the *Internet*? Is it the set of machines in the background and the cables or radio links between them, or is it the interlinked content, or both? Is this already a comprehensive definition? The modern Internet is everywhere, as it functions as the base for communication, which happens on top of the physical network using information entities like documents and messages. A new term is getting more and more attention: *The Internet of Things* (IoT). Historically, the ARPANET is the ancestor of the Internet, it represents the early version of the technical backbone. The World Wide Web (WWW) can be seen as a layer of structured content which was made available by connected servers. Social networks like Twitter, LinkedIn, Facebook, or LiveJournal evolved in the context of heavy social interaction between internet users who like to share content also privately. Nowadays, the content of such communication is also persisted and connected with user profiles within and across those systems. This leads to heavily interlinked content of multiple types, ranging from static pages, via comments and discussions to so called instant messages which do not automatically disappear after reading like in a traditional chat. The results are hybrid networks, formed by content and communication networks, with dynamic user interactions on top. New social communities emerge and wrap around the content and finally around the technical devices. Communication between people addresses real life aspects, but more and more also existing digital content. Finally, the communication itself becomes part of the content as soon as messages are stored, shared and interlinked. Beside social communities, more and more additional devices, which are not part of the technical infrastructure of the internet, are added to the network. Such devices provide even more data and contribute to communication, as they can generate messages like simple delivery confirmation or even alerts in critical conditions. Such interdependent networks are already present everywhere, but their impact on society and individuals is not well understood at this point in time.

Because over-simplified network representations are not appropriate for complex systems research, we need a clear definition of context even if the networks are describing different domains.

6.2. Definition of Local Neighborhood Networks

For practical reasons we have to develop an approach, one which allows us to collect a reasonable amount of data starting in a well defined environment - the scope or the focus of the study - and additionally to include a representative amount of data from its neighborhood, even if this is incomplete.

We have collected four exemplary data sets in order to develop our methodology and to demonstrate and systematically evaluate its performance. These test data sets consist of pre-processed log data and local network structures for several Wikipedia articles.

The primary structure of the Wikipedia network was retrieved with the Mediawiki-Explorer package, which uses a Java based implementation of the Media-Wiki-API client [157].

Data sets used in this part include a few selected central nodes (CNs, see Table 15.1) as well as their local and global contexts as illustrated in Fig. 6.2). The local context is defined by all articles (nodes) directly linked to the CN in the same Wikipedia, i.e., the same language version. This local neighborhood group of articles is denoted by LN. Each inter-wiki-link (IWL) connects the CN to a node addressing the same topic in another language (group IWL). The IWL group defines a global representation of the chosen semantic concept in all languages. We note that inter-wiki-links are not necessarily bi-directional, so that a CN regarding the same topic in another language may have a slightly different IWL group. Our IWL groups are always defined by the inter-wiki-links of the CN.

Finally, all articles directly linked to articles in the IWL group form the global neighborhood, a group denoted by GN. Note that the GN does not include LN, but it includes IWL and CN. This scheme, illustrated in Fig. 6.2, allows aggregation of data regarding specific topics in any language. At the same time, it is possible to separate the whole content stored for a single topic (or term) for each individual language to enable a language dependent analysis. However, it is not important if some of the selected pages are also linked to each other within a group or across the group boundaries.

Here, our interest is in the role of Wikipedia as a public crowd based source for news in the context of market activity, and we are especially focused on the reader's side. The number of article downloads reflects the state of a larger part of the society which can hardly be influenced by a dominating opinion of a single publisher using a shiny picture or provocative headlines on page one of a newspaper. Readers select articles intentionally and they are not flooded by topics which just sell well. Moreover, Wikipedia is not a commercial system nor is it influenced by advertisement like in the case of Google search.

The creation of the research context means in our case that keywords or Wikipedia pages have to be selected according to a given research topic. A market study has to cover entities, related to the market, such as participants, competitors, products, and other related subjects. Our approach uses existing implicit semantic relationships between Wikipedia pages to discover such term neighborhoods automatically as illustrated in figure 6.1.b. Based on these local neighborhood networks we have a set of domain specific topics and the pages which they are embedded in. Using the embedding as a kind of a background signal allows us to normalize the directly measured values in a context sensitive and time-resolved way.

6.2.1. Single Concepts vs. Groups and Categories

The local neighborhood of a single concept forms a tree as long as only next neighbors are taken into account. In many cases this would not provide a sufficient amount of data for statistical analysis. Therefore, we can extend the data collection procedure in two directions. First, a deeper crawl is possible, this means, we collect data for a higher recursion depth by following more links. Second, we can select more CN pages, especially if the same concept has different names or facets represented by multiple pages. In this case each facet contributes to the overall amount of information. Figure 6.2.a shows the group definition as a set diagram without explicit links. Links are used here for group definition only. Elements do not have to have links to all elements in the same group. All possible links between groups are highlighted in the adjacency matrix in Fig. 6.2.b, which is also a simplification of the network, shown in Fig. 6.2.c. The network as shown here is only a representation of one single aspect, derived directly from the data. Such a structural network describes a particular state of the system at a given time. It does not reflect the dynamics of the system at the same time.

The primary goal in this chapter is to define the *'core'* and the *'hull'* (see Fig. 6.2.b), which represent the research scope. In the first approach, as mentioned before, we use only one node as CN. The semantic core is formed by CN and SCN (also labeled as IWL since it contains inter-wiki links to all languages) in this case. The semantic neighborhood is LN, GN (all languages). The result is a tree, as long no back-links to pages are used. In order to handle a bidirectional embedding in the neighborhood, all in-links would have to be considered as illustrated in Fig. 6.1.a. The second approach is based on a list page chosen as CN. A category page can be used as well. The approach does not start with the central node but with a *'Meta page'*. This has an impact on the data collection procedure. IWL contains now the Meta pages in other languages. LN and GN are not the neighborhood but all core topics, and they form the set of CN pages like shown before. This approach requires an additional step for data preparation but allows access to many underrepresented topics. An alternative approach for local network aggregation is based on the concepts of the linked data web. DBPedia is called *"a nucleus for a web of open data"*[158]. The major part of the semi- and unstructured data from Wikipedia is available in a

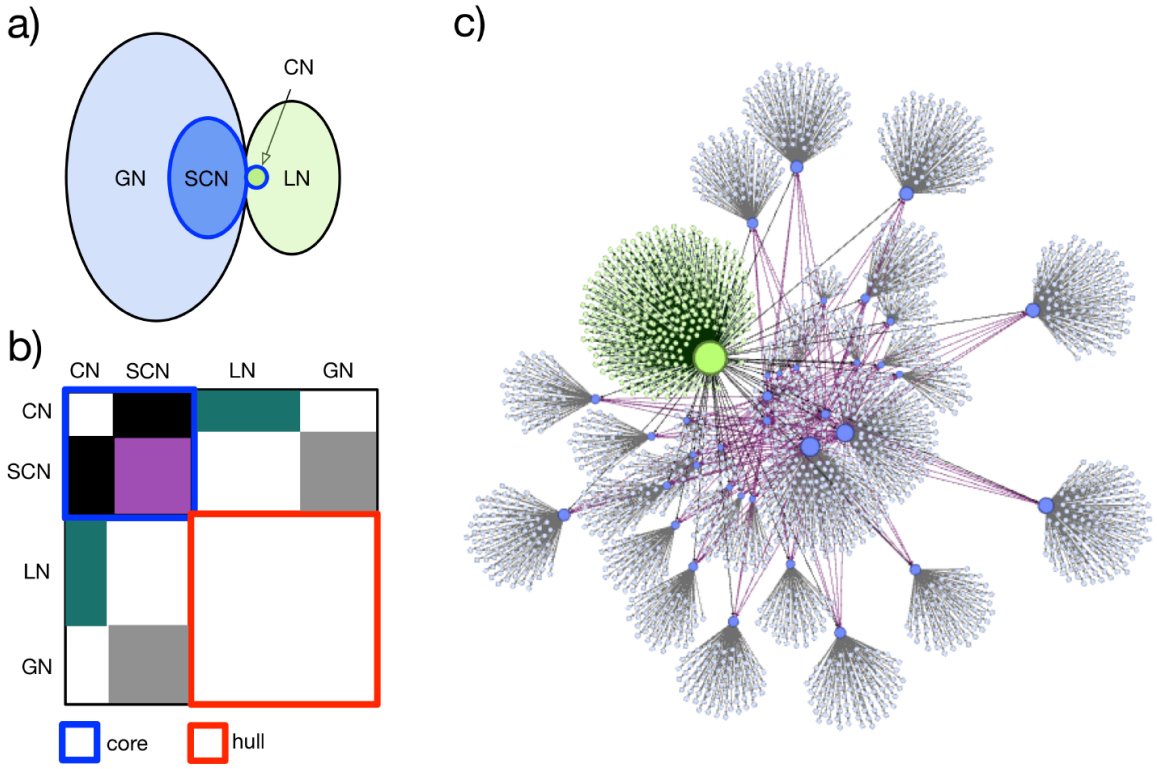

Figure 6.2.: **Selection and preparation of contextual data sets**. (a) For each topic (central node, CN), we study all directly linked nodes in the same language (local neighborhood, LN), all nodes regarding the same topic in other languages called *semantic core, SCN* (linked by inter-wiki links), and all nodes linked to nodes in SCN group (global neighborhood, GN). (b) The CN and the SCN group (IWL respectively) form the core of the local network for a semantic concept, while both neighborhoods form its hull. White colored node pairs (see matrix in (b) are ignored because they express long range relations between groups which are considered to be less relevant. (c) Network representation of all nodes regarding the central node 1.1 (from table 15.1). Note: IWL and SCN are synonyms for the same group, where SCN is more general and IWL refers to a particular contextualization approach.

transformed highly structured representation [159, 160]. This allows arbitrary queries in SPARQL[1], which are more flexible than queries in SQL-Databases. Partial import into local databases, and data discovery procedures support research as well as a possibility of logical reasoning based on variable ontologies. But currently, to our knowledge, no access history is available for DBPedia.

Summary

Selection of a context and a neighborhood allows comparison of otherwise non-comparable systems, based on topological network measures. Beside descriptive statistics of node properties and edge properties we calculate network measures using a reduced system size. Especially comparisons of systems from different domains and analysis of the same system over time are important for applications in science and industry. The static network structure can now be seen as a kind of *ground state* of a system. Different systems can have comparable or very different structures represented as networks. Especially, if more layers of functional networks are added, it is necessary to have a reference layer. Each functional layer can have an impact on the underlying structural network or both can be consistent.

[1]SPARQL: **S**imple **P**rotocol **a**nd **R**DF **Q**uery **L**anguage; defines a standard query language and data access protocol for use with the Resource Description Framework (RDF) data model.

7. Data Selection and Study Design

> The saddest aspect of life right now is that gathers knowledge faster than society gathers wisdom.
>
> *(Isaac Asimov)*

The dynamic content network of interlinked Wikipedia pages is created and influenced by the social network of interconnected Wikipedia users. One has to consider multiple networks because Wikipedia is not just one large multilingual system but rather a combination of independent sub projects grouped by language[1]. Additional processes - such as server and data maintenance tasks - have also an impact on system availability and usage. The growth of the network, its individual pages, and the growth of page clusters are embedded into an ongoing restructuring process. Although, we do not study the growth process of the network in detail, rather than the interactions on top of the network at a given point in time, it is very important to notice that the system in the focus of our study is not in an equilibrium.

The Wikipedia page network has a well defined inherent structure. This complex structure of static links (hyperlinks between pages) and user-to-page relations forms an evolving network in which links are added and removed over time. Both, the edit- and the access-processes coexist and influence each other. It seems to be intuitive, that network growth, content usage and contribution of new content are highly interdependent.

One of the goals of this work is a systematic and quantitative comparison of two distinct but coupled processes on top of a complex system. The first process is the interactive modification of Wikipedia pages by Wikipedia user communities or individuals. The second process is information retrieval, either sporadic consumption, or systematic research by humans, or access by automatic systems, such as web crawlers or mobile applications (see Wikipedia page with title *'Tools/Alternative browsing'* [163]). Because automatic page retrieval influences the outcome of data analysis procedures it is important to describe the raw data carefully. This allows identification end elimination of hidden biases caused by a variety of reasons. Therefore, we introduce a new approach for contextual data preparation. Based on the idea of semantic context networks and the herewith defined idea of a semantic neighborhood we developed a generic procedure applicable to all types of studies, focused on Wikipedia or even more general, time series analysis on data collected from dynamic evolving systems.

7.1. How to Select the Right Time Series?

Even if the system is non-stationary we have to prepare the data in a way, which allows us to treat the system like a stationary one. This assumption can be correct if the selected time windows are not too long. Monthly data sets give us 720 data points on hourly resolution or 30 data points on daily resolution, which is fine for correlation analysis. Especially the analysis of a daily patterns (see [10]) or weekly patterns in access-rate time series allows a classification of network nodes according to typical usage patterns.

The data preparation procedure includes pre-aggregation and grouping according to the inherent structure of the local neighborhood network which is selected for the individual study. Thus, we need an auto-adaptive approach, especially because we cannot know much about the data quality in the beginning. Besides the raw input data, which is used in the following analysis procedures, also the contextual metadata is retrieved and conserved in our shared knowledge base. This method enables traceability and allows context sensitive interpretation of results, e.g., outliers and systematic trends can be identified and eliminated this way.

In this chapter we present typical properties of the data set. Data was extracted from Wikipedia server log files and from the life Wikipedia system. During this step we prepared time series buckets (TSB). The page content was also collected. The full page text is available for further analysis as a contextual corpus dump (CCD).

This enables context sensitive group based analysis. Both techniques, contextual corpus dump and time series buckets also allow fast random access to individual pages and time series for any given page in an offline environment. This chapter shows results of a data set inspection. We begin with descriptive statistics of some example groups.

Four different local neighborhood networks for English Wikipedia pages have been selected in order to illustrate important time series and structural network properties, such as daily and weekly patterns beside the degree distribution of underlying networks. One can now choose from raw data, trends, and detrended data. Comparison

[1] According to `http://meta.wikimedia.org/wiki/List_of_Wikipedias` more than 280 different Wikipedia projects exist beside more specific content category types such as, e.g., Wikibooks [161] or Wiktionary [162]

of, e.g., daily and weekly trends can be used to classify nodes of the network. In particular, the page about *'Formula One'* was selected because of the well known repeating patterns. For *'Influenza'* we assume a seasonal effect as well, but on a different time scale. After the outbreak of Ebola in West Africa in 2014 we also choose the related Wikipedia page for a direct comparison with the page about Influenza, which are both from the same context. Finally, the page about *'Econophysics'* was select because of personal interest the fact, that the topic is not yet so well known.

In previous studies [10, 164, 134, 8, 165] individual time series have been analyzed independently from each other. Now, in this work, also the context of the pages is taken into account. Further details about context definitions are provided in section 6.2. This allows a systematic contextual comparison of comprehensive node groups covering several topics, even if a clear separation of topics is not possible.

Data set profiling - as a general procedure combined with early visual inspection (see figure 7.1) and simple quantification of data set properties - is crucial, especially if data was obtained from large-scale systems and multiple sources. Data set profiling allows early validation and plausibility tests as soon as intermediate results are available.

Table 7.1 shows two data set profiles. Of special interest is the number of existing pages in the local neighborhood network and the number of available access- and edit-rate time series. We inspect the data coverage rate r_{dc}, beside the average access and edit activity per group. For economical and technical reasons it is not always possible to collect all data about all pages. This makes appropriate data set profiles even more important.

Page Network		**Access Activity**			**Edit Activity**		
Formula One							
Group	z_{pages}	z_{ats}	a^{*}_{access}	r_{dc}	z_{ets}	a^{**}_{edits}	r_{dc}
2 CN	1	1	3800	100 %	1	623	100 %
2 IWL	91	3	621	3 %	91	43	100 %
2 LN	1128	70	1955	6 %	1125	58	99 %
2 GN	17293	1066	437	6 %	16907	16	97 %
Influenza							
Group	z_{pages}	z_{ats}	a^{*}_{access}	r_{dc}	z_{ets}	a^{**}_{edits}	r_{dc}
4 CN	1	1	3407	100 %	1	335	100 %
4 IWL	107	12	516	11 %	110	13	103 %
4 LN	684	203	1777	29 %	772	87	113 %
4 GN	7781	1005	522	12 %	7634	14	98 %

Table 7.1.: **Data set Profiles**. Two example data sets from Wikipedia pages about popular topics have been selected to illustrate time series and network properties which are relevant for this work. Group sizes (z_{pages}), number of available time series per group (z_{ats} and z_{ets}), average daily access (a^{*}_{access}), annually edit activity per node (a^{**}_{edits}), and data coverage (r_{dc}) are compared for core (groups CN, and IWL) and hull (groups LN, and GN) from Local Neighborhood Networks. The dataset names are also the page names of the selected central nodes taken from English Wikipedia.

Column z_{ats} and z_{ets} in table 7.1 show the number of available time series for access-activity and edit-activity. For all existing pages it is possible to load the edit history. The coverage ratio for edit event series is usually 100%, sometimes it is less, if pages were removed since the network structure was collected or more, depending on the crawl mode (see groups 4.IWL and 4.LN in column r_{dc} in table 7.1). In case of a life crawl, we can detect the differences in the page network (and thus more nodes than previously identified lead to $r_{dc} > 100\%$ as in groups 4.IWL and 4.LN). During static crawl mode only the edit history of the previously collected page names is retrieved. In this case the coverage cannot be above 100%. If no edit activity was detected during a given time range, we could not distinguish between no activity or a non-existent page. If the growth dynamic is taken into account during a time-dependent study, one should carefully distinguish both cases because a non-existent page is not the same as an inactive page. If the page creation date is after the end of the period of interest, then this page should be considered as non existent. In our case, the data coverage ratio is calculated for access- and edit-activity. The average number of access-events per node and day is shown in column a^{*}_{access}. Column a^{**}_{edits} contains the average number of edit events per node per year.

These values allow an estimation of the available amount of data and the collection cost (time and resources), and it helps to estimate the expected significance of results. If the coverage ratio is very low, as in the case of the page about *Formula One*, one should interpret the results carefully. The reason for this low coverage can be a technical problem during data collection. As data transformation and aggregation is done via self made systems, additional validation and plausibility tests are required. If no technical problem can be found, there is still the chance of existence of a hidden bias. Our analysis approach can be applied also if the groups are not complete. In general, the central node is the one, which was selected because of its relation to a specific topic. The neighborhood is used for auto-adaptive normalization. The more time series we can get the better the accuracy would be. Even if only three rows would be available, results can be meaningful.

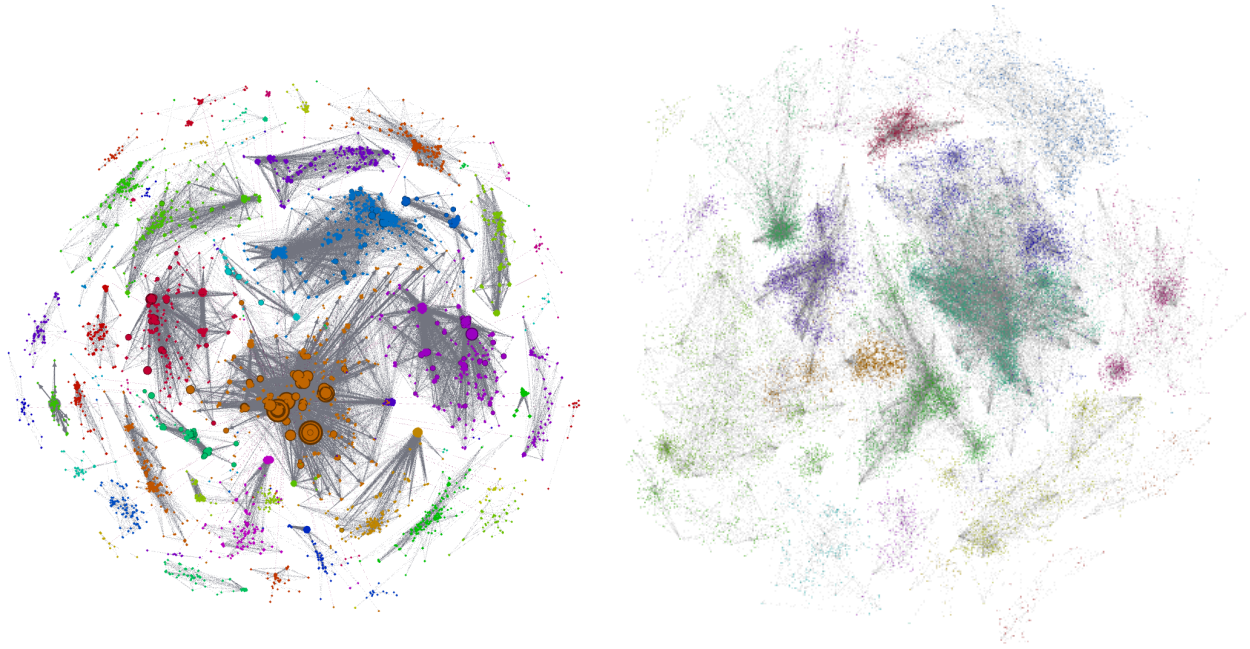

Figure 7.1.: **Local Neighborhood Networks for two different Wikipedia pages** starting at the central node *'Influenza'* and *'Econophysics'* (from English Wikipedia) the neighborhood network can be seen as a set of linked trees (left) or a clustered network (right). These networks allow a context sensitive interpretation of user's interest in represented topics. They are used to expose the structural properties of the groups used in contextual analysis and for comparison with dynamics related data in contextualized time series dashboards (TSD) as shown in figure 7.3). Furthermore, a comparison with the text based representation index reveals potentially existing bias (see figure 11.1).

The selection of appropriate central nodes is crucial for purely data-driven analysis. The most common problem during early study design was the selection of a Wikipedia page which was only a redirect page. In many cases, no access count data exists for such pages. One has to follow the redirect link and the page name has to be replaced. Also disambiguation pages should not be selected as a starting point for crawling, except, one wants to study properties of such pages. Table 7.1 also highlights some typical problems found in the raw data set. From four initially selected central nodes, only the two about *Formula One* and *Influenza* are shown here. We use those also for demonstration of the newly developed analysis methods. Figure 7.1.a highlights properties of individual nodes. The global properties of the neighborhood, especially the density and size of clusters is in the focus of figure 7.1.b.

7.1.1. Properties of Selected Local Neighborhood Networks

A local neighborhood network (LNN) allows segregation of essential information without losing all information included in the structural embedding. Depending on n, one cares only about the next neighbors. Such a simple local neighborhood graph (n=1) has a tree structure. A broad embedding including multilingual aspects is available for larger n and if inter wiki-links (IWL) are considered.

The left panel of figure 7.1 shows multiple trees which form the multilingual neighborhood of the page *'Econophysics'*. An obvious network structure appears for $n = 2$ (see right panel of figure 7.1).

For studies about topics, which cannot be defined by one term, such as political topics, economical aspects, art, culture, or financial markets, one should include more than one central node. In chapter 15 we show such a market study. 42 different Wikipedia pages were selected as central nodes. All are representative for the emerging Big Data market or at least historically related to the topic. Combining all those trees in one graph also produces a network structure. Cluster analysis reveals several modules comparable with figure 7.1. Such modules can either represent clusters of highly interlinked pages within a given semantic space or more likely pages from the same language. Figure 7.1 shows color-coded modularity classes for pages from multiple languages. This clearly illustrates, how important a proper validation of the selected data is. An early inspection of the network structure allows validation and plausibility checks before expensive data extraction and network analysis algorithms are applied. Note, such clusters form sub-graphs and depend on the method of data extraction.

We can conclude: (1) Data collection separated by language works best for trees ($n = 1$). The focus of such a study is preferable the lingual space. **(2)** Content and topics influence the clusters in case of deeper crawling with $n > 1$. In this case much more neighbors are taken into account. Efficiency of this method can decrease very fast.

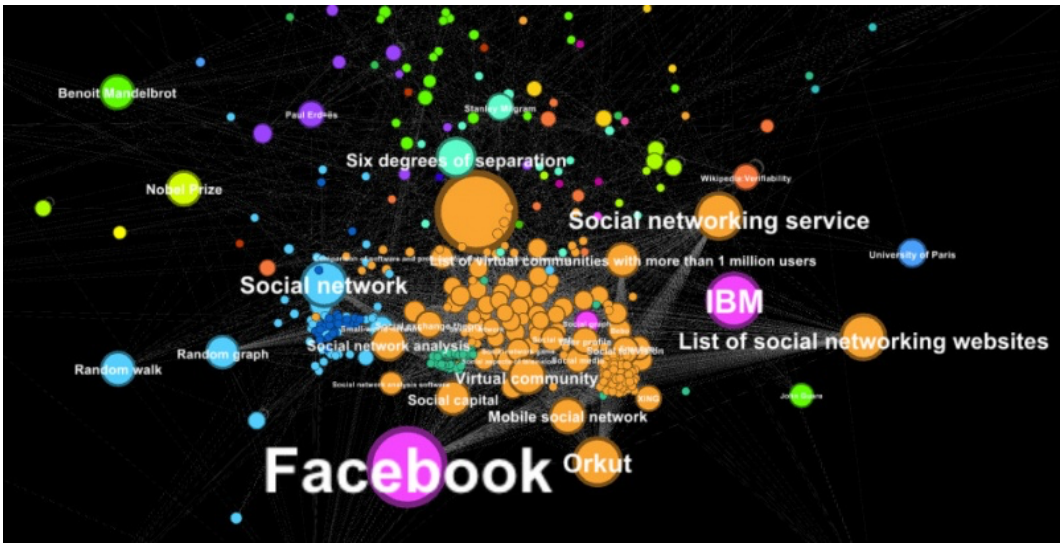

Figure 7.2.: **Representation and embedding of the *'Milgram Experiment'* in Wikipedia.** A Local Neighborhood Network (LNN) is formed around the page of each individual semantic concept or topic. LNNs provide useful and machine readable metadata for Wikipedia pages. Such a contextual embedding allows quantitative and structural analysis. Especially a comparison of different topics and their representation in different languages as a function of time is important for improvements in dynamic network models, which describe intercultural phenomena. In interdisciplinary communication, especially in global systems like Wikipedia, it is important to include all user groups, which are represented by Wikipedia pages in different languages. Population size and language usage differ by topic and culture. This hidden bias has to be identified before it can be taken into account during the interpretation of analysis results obtained from automatic procedures, like log-file analysis.

(3) With $n > 3$ the network will be very large in most cases. Instead of a deep crawl, a broad selection of central nodes should be considered as this allows creation of meaningful interconnected neighborhood networks even with a flat crawl procedure.

Figure 7.2 shows such a second level neighborhood for the Wikipedia page about the famous *'Milgram Experiment'*. The network is not just a set of simple trees any more. A systematic comparison of topics regarding their embedding is possible. We show common network measures, available in the network visualization software Gephi [18] in table 7.2.

A topic oriented analysis would be useful, if we are able to measure a characteristic property which allows one to distinguish a broad topic from a group of very specific pages. Network visualization gives a first impression very fast, but in some cases the networks are too large. In this case we have to care about a numerical representation, automatically computed from raw data. I suggest to count the number of triangles and to calculate the diameter of the LNN representing a particular topic. An efficient detailed analysis of thousands of such LNN graphs is possible, based on our preliminary results, but because of limited time and computational resources we could not apply machine learning algorithms for automatic classification of the LNN graphs obtained from Wikipedia.

For some examples we collected the LNN graphs, access-rate time series, and edit-event time series. The time series dashboards are presented in the following sections.

7.1.2. Contextual Time Series Dashboards

Contextualization based on local neighborhoods, as introduced previously, allows grouping without '*a priori*' information, especially in a non stationary environment. The structure at a given point in time is taken into account for contextualization. Even if more nodes will exist later or if nodes will be removed or split this approach works stable. Because Wikipedia evolves over time - this also means more users might be attracted - a higher connectivity in the functional network or a change of the topology can be a consequence. It is really important to know the system's structure and how it evolves over time. With this in mind, a comparison of averaged time series data allows a further contextualization. One can compare the activity pattern for different groups rather than the absolute access activity to individual nodes. Individual events (represented by bursts) can be very important because they exist in multiple series. They can also be considered as outliers or results of a technical problem inside the system. Figure 7.3 shows a *contextual time series dashboard* for two different LNN of very different topics. The access activities for both pages show the same pattern during the highlighted period (see grey bar). This is a clear indicator for a systematic error in the data or in the system, which produces the data. One can

only identify such systematic errors if background and context information is available and used.

Furthermore, figure 7.3 allows a direct comparison of the patterns, such as peaks (see label A), plateaus of constant interest (see label B), or increasing trends which are overlaid by oscillations (see label C). We use the time-dependent relevance index (TRRI, see Eq. 11.5 and Eq. 11.6 in chapter 11) for a direct comparison of local and global neighborhoods. As shown in figure 7.3.d, we can identify local maxima above the seasonal trends and the comparison of the seasonal trends is much easier with this relative measure.

In general, the edit activity (see figure 7.4) is far less than the access activity to Wikipedia pages. In all cases we found, that the edit activity for the central node (CN, green crosses) is above or comparable to the edit activity in the local neighborhood (LN, black curve) on a comparable level for all four selected groups in figure 7.4. A lower activity in group IWL indicates less interest in the topic in other languages compared to the chosen language of the core node (a,c, and d). For group 2 we found a high edit activity also for other languages, especially at the beginning and the end of the period, where the blue symbols (IWL) are close to the black line, which represents the average activity in the local neighborhood.

7.2. Peaks, Hidden Bias, and Trends

How can one distinguish between a unique peak and a repeated pattern? This question seems to be trivial, but it is not. A peak can appear regularly on the same day of the year. The frequency of the phenomenon and the selected time scale affect the distinction between unique events and repeated patterns.

7.2.1. Weekly Patterns on Hourly and Daily Resolution

If the length of the time series is well defined, then we can find an appropriate time scale for which we calculate the average values, e.g., for all Mondays, for all Tuesdays, and so on in order to calculate the weekly average values for hourly (as shown in figure 7.5) and for daily resolution in figure 7.6. Such weekly patterns reveal a typical fluctuation in the order of 50% of the maximum but on Sundays the activity is more than twice as high as the average activity on normal days. A peak in the weekly patterns can be caused by recurring events or even by one single large event (compare with the single peak in the activity time series for the Wikipedia page *'Illuminati'* in [8] figure 1). Not only the weekly activity pattern is influenced, during the burst, also the variance is significantly higher. In case of recurring peaks triggered by recurring events, like Formula One races on Sundays (but not all Sundays) the variance would be smaller. The ratio between absolute value and variance can be used as a measure to distinguish periodic re-occurrences and sporadic events. A comparison of average and mean value allows also to identify the presence of outliers. In such a case, both values would be different, but in case of a Gaussian distribution without any extreme values the mean and median of the distribution would be equal.

Wikipedia access time series show seasonal trends on multiple time scales. Daily access patterns are dominantly caused by the day-night cycle. Some illustrative examples for cycles in access-rate time series are shown in [10] and for edit activity in [117]. A strong effect can be observed for pages written in languages with a very strong linguistic localization. This periodic effect is much weaker for pages in English language because of its wide spread

Metric	*Influenza*	*Formula One*	*Econophysics*	*Ebola*	**Reference**
k, topology filter ($k > 1$) - no leaves					
Number of Nodes	21831 (6.66)	131593 (18.31)	13155 (19.04)	34337 (9.15)	n.a.
Number of Edges	281466 (42.03)	1651125 (71.44)	54276 (49.25)	296473 (41.64)	n.a.
k, no filter					
Number of Nodes	327793	718529	69086	375450	n.a.
Number of Edges	669600	2311124	110207	711950	n.a.
Average Degree	12.893	12.547	4.126	8.634	n.a.
Nr. of Weakly Con. Comp	8	3	1	2	Tarjan *et al.* [76]
Nr. of Strongly Con. Comp	18591	121146	12618	30762	
Modularity	0.854	0.945	0.833	0.925	Blondel *et al.* [74]
Number of Communities	70	74	22	81	
Average Clustering Coeff.	0.434	0.323	0.357	0.509	
Average Path Length	3.82	2.70	3.62	3.06	U. Brandes [38]
Number of Shortest Paths	16095492	59851540	2182375	6866597	
Diameter	10	9	8	13	

Table 7.2.: **Social Network Profiles (SNP)**. Comparison of complex networks requires a comprehensive view. Such a view is provided by an SNP. We calculate SNPs in Gephi. A better integration into analysis workflows is possible with libraries such as *JUNG2* [166], the *Gephi toolkit* [18], or *snap.py* [167]. Definitions and implementation details about the listed algorithms are available in the cited articles (see column references).

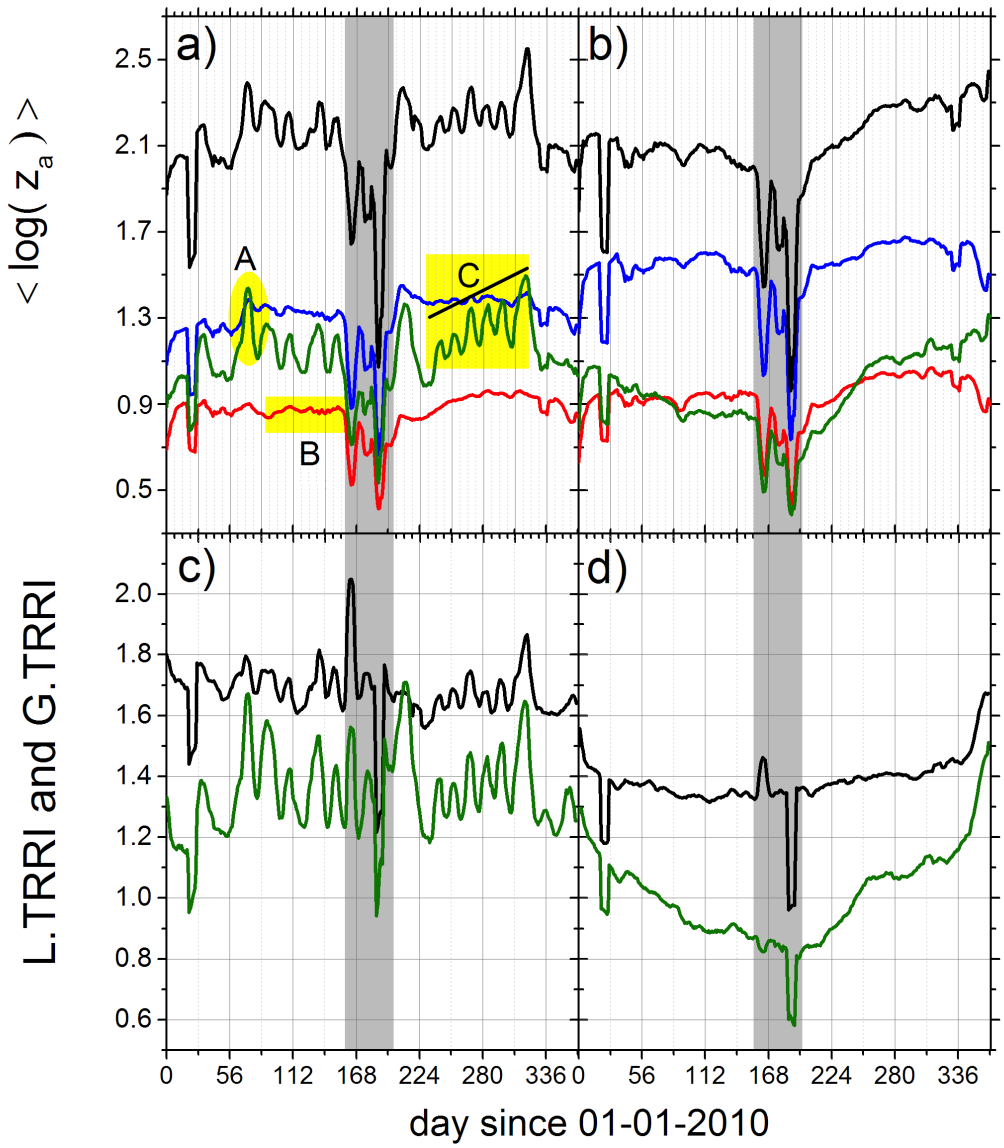

Figure 7.3.: **Time Series Dashboard (TSD) for Wikipedia access activity**. The logarithm of the hourly access activity in semantic neighborhoods of Wikipedia pages *'Formula One'* (a,c) and *'Influenza'* (b,d) is shown for the year 2010. In (a) and (b) the black curve is for the central node (CN), green for IWL, blue for the local neighborhood (LN), and red for the global neighborhood (GN). The gray area highlights a time window, during which errors in the data acquisition procedure were identified. In (c,d) the Time Resolved Relevance Index (TRRI, see chapter 11) shows the effect of the detrending approach. Local context (L.TRRI, black) is the English Wikipedia page whereas the global context is defined by all non English pages (G.TRRI, green). Such context information is essential for interpretation of user's interest in represented topics without ignoring the properties of the influencing neighborhood.

and its global relevance. The day-night cycle can be modeled by a periodic function with a characteristic scaling factor s_d for each day of the week. Because $\sin(\omega t)$ has negative values, $\sin^2$ should be considered.

Other classes of seasonal patterns are less or not periodic and therefore cannot be easily modeled by a simple function. Depending on the occurrence of events in the real world, such patterns can be represented by a series of strong peaks. Such peaks are found dominantly on a specific day of the week for some weeks of the year only, e.g., on Sundays as shown for the Wikipedia page *'Formula One'* (see top row in figure 7.5). A longer seasonal trend with a maximum during the winter period and a minimum during the summer time can be shown for the page about *'Influenza'* (see section 7.2.2).

Figure 7.5 shows typical patterns on hourly resolution. At daily resolution (see figure 7.6) the schematic differences between the two topics become more obvious. For figure 7.6 the data was re-arranged, so that the first day is Monday. This allows a more intuitive interpretation according to a calendar week.

One has to be careful here. In figure 7.5 a time window is aligned with the raw data, which starts on the 1-st of January of 2010. This was a Friday. Both representations have to be used in an appropriate context. For a

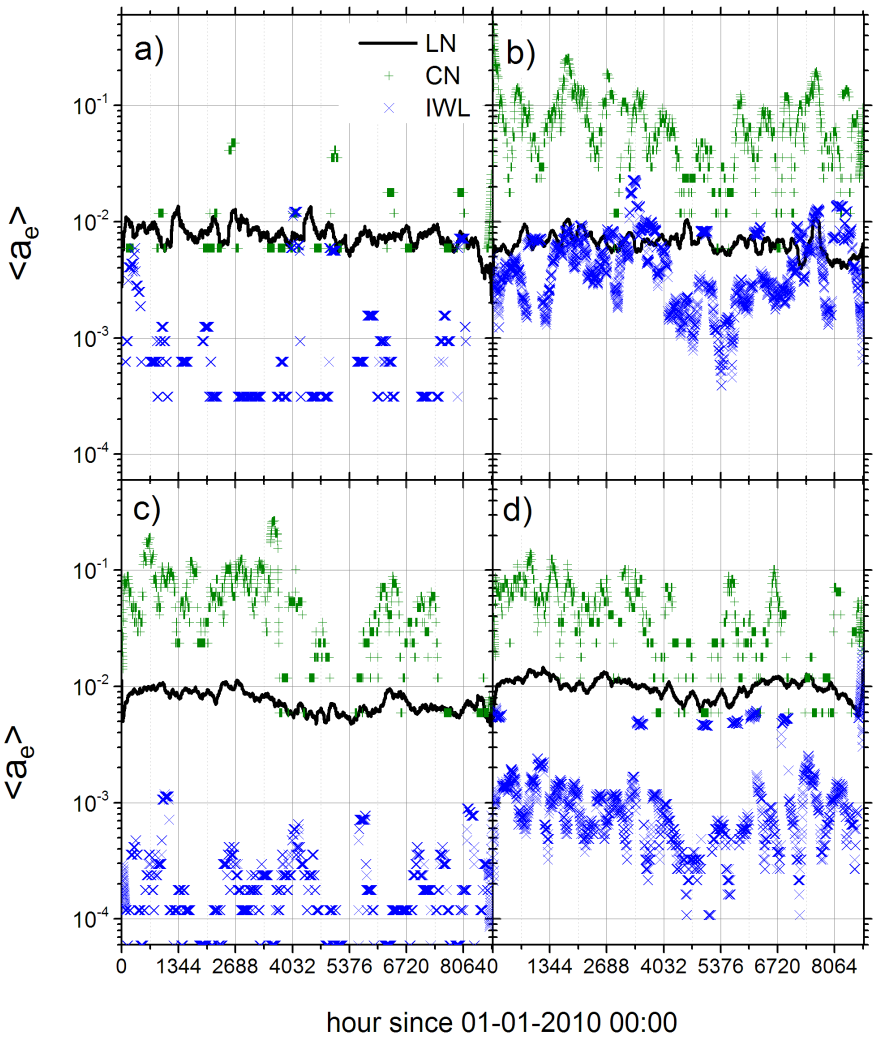

Figure 7.4.: **Time Series Dashboard (TSD) for Wikipedia Edit activity**. Comparison of the direct neighborhood of Wikipedia pages *'Econophysics'* (a), *'Formula One'* (b), *'Influenza'* (c), and *'Ebola virus disease'* (d). In (a) we found a very low edit activity, which is plausible, as the topic Econophysics is a niche topic and rather young. The edit activity in established topics as in Formula One is high also for the pages in other languages as shown in (b) by the blue and black curve. Note, the activity in group IWL (blue) is close to the activity of group LN (black). (c) and (d) are examples with a much lower interest from people who don't use English (blue). The difference between CN (green) and IWL (blue) is up to three orders of magnitude temporarily.

general interpretation, the "re organized" data seems to be preferable.

One can see that on weekends the access activity significantly differs. It decreases in the core and hull network around the page 'Influenza'. In case of 'Formula One', we can see a difference between the core and hull. The core, formed by the page and all international representations of the same (CN and IWL) shows increased activity on weekends, due to the races, while the neighborhood in English language and also all international neighborhoods show decreased activity on weekends. A normalization to the maximum value can highlight this structure even better.

This data allows a classification of the pages. The weekly pattern reveals if a page is continuously in line with the neighborhood (as in case of 'Influenza') or if a specific time exist, during which the access activity differs from that found in the neighborhood (as in case of 'Formula One'). The top row inf figure 7.6 highlights the different behavior in different colors. This property can easily be described with a signed number. The bottom row shows no such difference, so we treat this as a neutral node, while the other one shows a clear polarizing behavior. Such a node property cannot be measured from just one single time series. One needs the structural properties of the neighborhood graph to define the appropriate node groups for which the averaging procedure is then applied. This kind of node property, if assigned to the central nodes of a page network, can influence the layout of the graph if the layout procedure takes this polarity into account. Beside '*Social Gravity*' (see Bannister *et al.* [94]) it seems to be reasonable to use also the new concept of '*neighborhood polarization*' - a macroscopic analogy to spin - which describes a node's orientation within a force-directed layout. This leads to a measure called *structure induced*

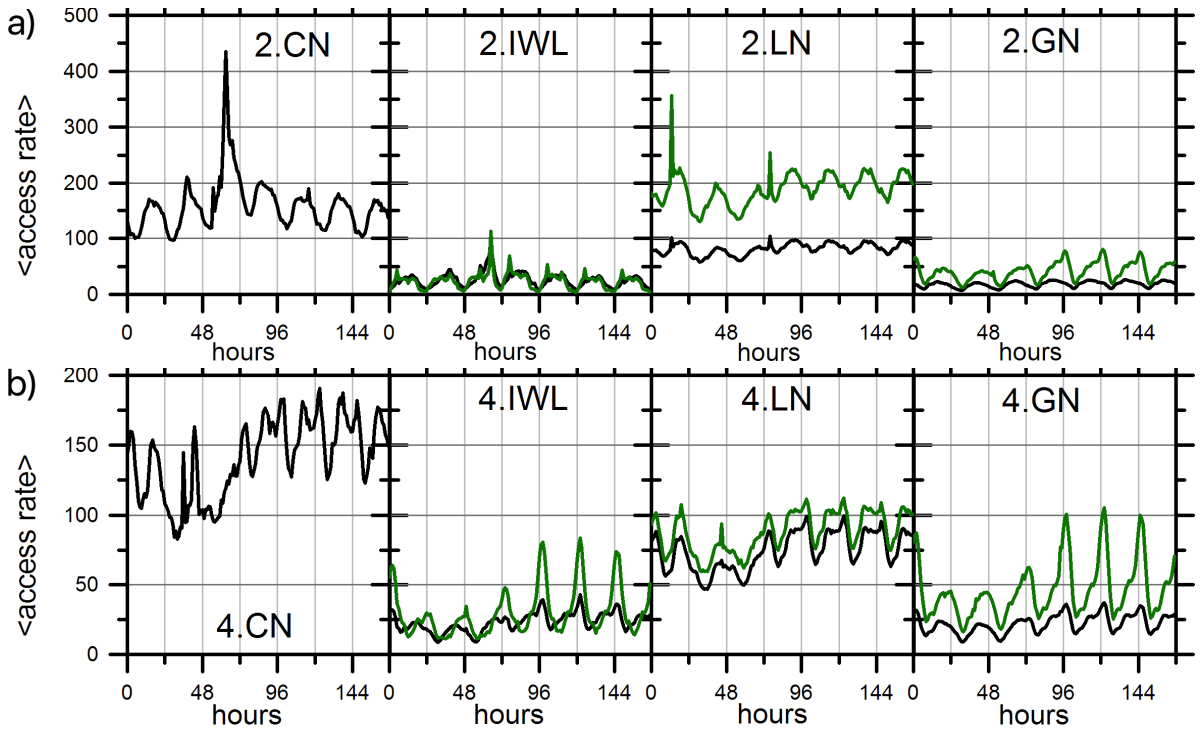

Figure 7.5.: **Weekly trends in access rates at hourly resolution**. Characteristic patterns in Wikipedia access-rate time series (black curve) show a strong day-night cycle especially for pages in languages which are not spoken globally. Strong peaks in the weekly averages can be caused by individual extreme events on one single day or by recurring events on the same day of the week. The two causes can be distinguished by analyzing the standard deviation (green curve)

stress. This novel approach will be introduced in chapter 12 (see figure 12.1).

Table 7.3 shows results of correlation analysis for a comparison of weekly trends. Such a quantitative comparison of weekly trend patterns is fast and efficient. It allows pairwise node classification and early detection of anomalies. Node pairs with comparable properties behave the same way. This leads to high correlation values which express a strong similarity (see right column in table 7.3). A different type of pages, which are not in line with their neighborhood, can easily be distinguished from those, because of their low correlation values (see left column in table 7.3).

7.2.2. Seasonality in Detrended Time Series

Identification of weekly trends (see previous section) is useful for two reasons. First, they allow a classification of nodes. Furthermore, removing these weekly trends, as shown in figure 7.7, works as a smoothing and normalization procedure and leads to time series which can be compared directly with each other.

The top row in figure 7.7 shows the averages of the logarithm of access-rate data for one year for the central node CN (black) and different neighborhoods (IWL: olive, LN: blue, and GN: red). The bottom row presents the results for the local and global time-resolved relevance index (see Eq. 11.5 and Eq. 11.6). This data allows a better comparison between the time series on a daily basis. For example, the differences between L.TRRI (black) and G.TRRI (olive) or just the sign of the difference can be used to define another node property. We call this property *activity polarization*. A node has a positive polarization if L.TRRI > G.TRRI (see also red area in figures 7.7.a and 7.7.b) and a negative polarization otherwise (see blue area in figures 7.7.a and 7.7.b). The node activity can

Groups	*Formula One* R_{Pearson}	*Influenza* R_{Pearson}
CN - IWL	0.97	0.99
CN - LN	-0.36	0.99
CN - GN	-0.19	0.98
LN - GN	0.93	0.99

Table 7.3.: **Comparison of weekly trends with Pearson correlation**. A quantitative comparison of weekly trend patterns is a fast and efficient approach for node classification. Nodes which behave like their neighborhood (right column) can be distinguished from those with opposite trends (left column).

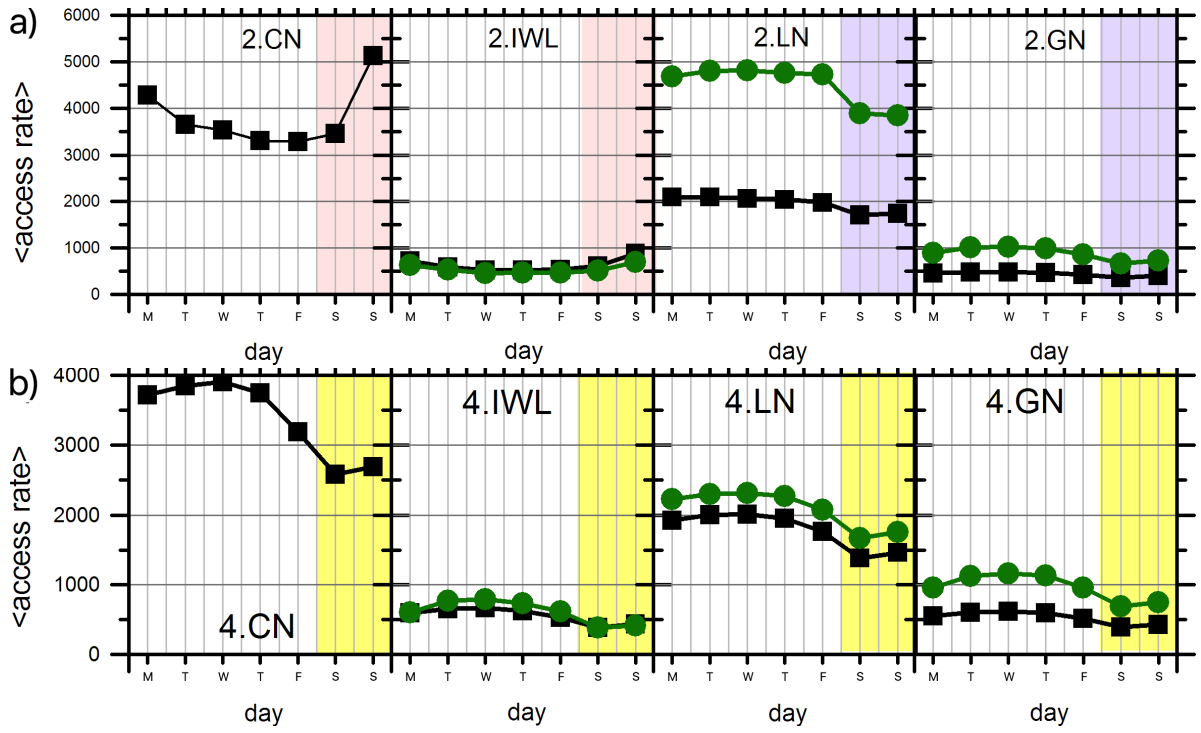

Figure 7.6.: **Weekly trends in access rates at daily resolution**. Characteristic patterns in Wikipedia access-rate time series show differences between core and hull of the local neighborhood network (LNN). A systematic comparison is done using Pearson correlation analysis and a visual inspection. The top row shows an increased interest in the topic *'Formula One'* on weekends, due to the date of races while the interest in related pages is decreased at the same time. The difference is highlighted in different colors. Red marks increased activity (blue for decreased) on weekends compared to weekdays. A coherent behavior was found for the page *'Influenza'* and the surrounding LNN. Weekly access patterns are the same for core and hull in this case and therefore highlighted in the same color.

now be interpreted as stable, if the polarization does not change during a time range. In figure 7.7.d no significant changes are detected during summer time (except for two spikes). The fluctuating polarization as shown in 7.7.c is characteristic for instable nodes. Bursts, triggered by real world events - in this case the Formula One races - dominate the overall activity of the page. This kind of page classification helps interpretation of results obtained from further analysis as explained in the following chapters.

7.3. Activity Correlation in Coupled Processes

For each individual time series we study correlations between access activity and edit activity. Because Pearson correlation is very sensitive and results can be misleading especially if the values are not from a Gaussian distribution we also apply the Spearman rank correlation and Kendall rank correlation to the data points obtained from the previously defined time series groups.

Group	R_{Pearson}	R_{Spearman}	τ_{Kendall}
Formula One			
core	0.92 −	0.40 −	0.33 −
hull	0.62 ∗∗	0.65 ∗∗	0.47 ∗∗
Influenza			
core	0.95 ∗∗	0.77 +	0.62 +
hull	0.67 ∗∗	0.61 ∗∗	0.45 ∗∗

Table 7.4.: **Correlation between access-activity and edit-rate time series**. Pearson correlation, Spearman rank correlation, and Kendall rank correlation (see section 5.4.2) are applied to core and hull of selected local neighborhood networks. The symbols indicate if correlation is significant (∗∗ $p < 0.001$; + $p < 0.01$; − not significant) the absolute values should be used with care.

We evaluate the following null-hypothesis: a correlation between the edit-activity and the access-activity of Wikipedia pages exists. According to the results in table 7.4 we cannot reject the null-hypothesis in both examples.

The correlation is less significant in case of the core due to the small number of data points. Pearson correlation should not be used in this situation, it would even indicate the opposite result, which is a stronger correlation. Local neighborhood networks can be compared based on this measure over a period of time. Thus, a time-resolved classification, depending on the correlation properties is possible, but only if a sufficient number of nodes exist in the selected groups.

Table 7.4 illustrates the differences between three applied correlation methods. Based on Pearson correlation we would conclude, that for core and hull a high correlation exist. Based on rank correlation a different result appears, a difference in the significance level between correlations in core and hull is visible for both examples for both rank correlation algorithms. The absolute values of the correlations should be used very carefully. Since the origin of the data is not well known, and no controlled measurement, but rather a very open data gathering approach was used, it is recommended to do a qualitative analysis only, based on a comparison of significance levels rather than absolute correlation values.

Summary

This chapter introduced the idea of node properties, named *neighborhood polarization* and *activity polarization* which are assigned to network nodes based on context sensitive time series analysis. Instead of the absolute activity of a node, the relative activity and a comparison of single node properties with group properties were evaluated. Results are presented in contextual time series dashboards.

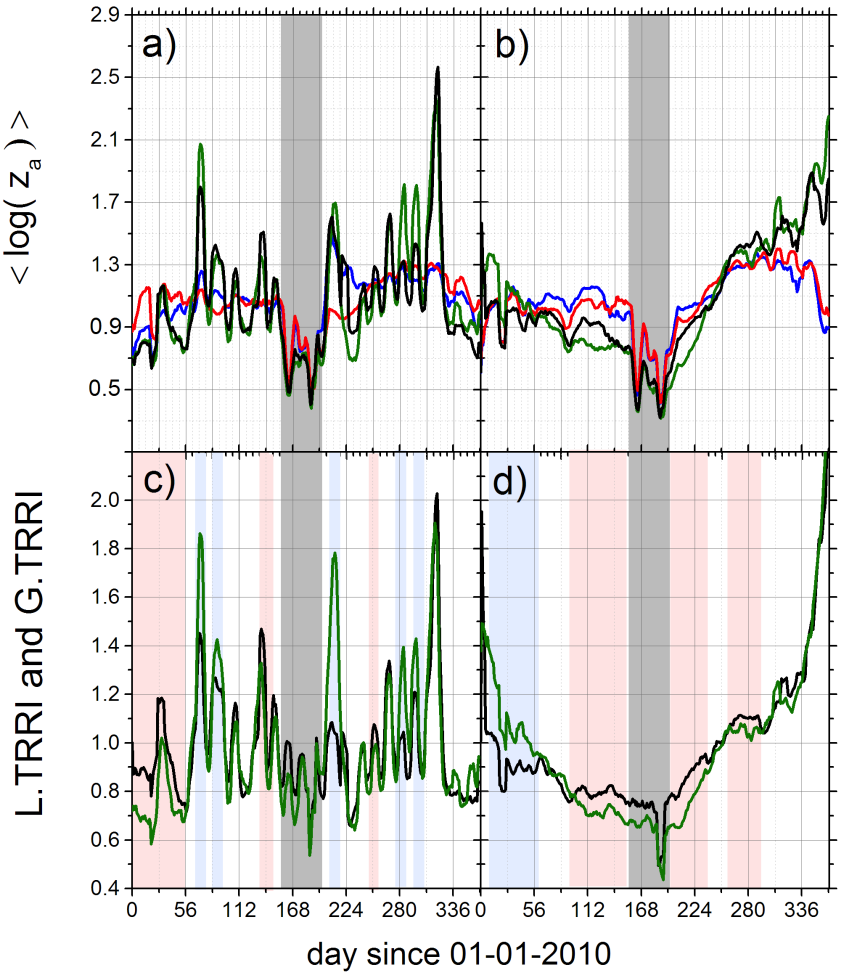

Figure 7.7.: **Detrended acces-rates in the same representation as for raw data in figure 7.3**. Detrended access-rate time series reveal seasonal patterns, and stability of Wikipedia page activity. On the left, one can clearly see strong peaks on Sundays, especially at the beginning and at the end of the *'Formula One'* season (a,c). This increased interest is triggered by races, which take place on Sundays. The right panels (b,d) show a significantly lower interest in the topic *'Influenza'* during summer time, which is plausible and intuitive. The time range during which technical problems were recognized are marked in gray. Red and blue areas indicate positive and negative *activity polarization* of the central node CN. The curves are colored as in figure 7.3.

8. The Life Cycle of Social Content Networks

> Look deep into nature, and then you will understand everything better.
>
> *(Albert Einstein)*

Different Wikipedia projects grow very differently. This is not surprising, because they are maintained by different communities. Thus, they are influenced by different economical, political, and cultural conditions. Because each Wikipedia sub project is created in a different language, one can say, that each Wikipedia project represents a different cultural context wherein different topics are important. This view is based on differences in how languages are used and how different cultural aspects are reflected by community driven content aggregation.

A second perspective exist. On a higher abstraction level, all Wikipedias can be unified, by ignoring the cultural and lingual differences. In this case one can say: Wikipedia is the global encyclopedia. It is also a crowd-based information and knowledge creation system, a growing system with inherent memory. For multiple languages, there exist several Wikipedia instances. All are interconnected clusters, representing subsystems, and may have comparable properties. If a reasonable approach for normalization of the data would exist, one could compare the project life cycle phases of each Wiki. Thus, we analyze the growth of four Wikis, the English, Swedish, Dutch, and Hebrew Wikipedia projects. We find that comparable properties exist beyond numbers of pages and links. This allows us to describe the life cycle phases based on growth rates and structural embedding ratios.

8.1. Cultural Aspects of Global Online Networks

How individual cultures influence and impact Wikipedia - the content and the communities which drive the dynamic processes such as content creation and information retrieval are in the scope of this work but beyond the scope of this chapter. Here we study structural properties of Wikipedia and follow a natural path, by organizing the data by language. This is easy, because Wikipedia projects for different languages coexist and they are interlinked already. Each language defines one *lingual dimension* for a global analysis. Further dimensions are topics based on automatically extracted topic models (see [168, 169]), and time, as used in this work.

A different approach which originates in social science uses very different dimensions. Hofstede [170] derived those dimensions from global survey data. In this way, it is possible to apply factor analysis to data to determine the predominant cultural dimensions. Hoftstede initially defined four cultural dimensions regarding fundamental anthropological problem fields. The dimensions are named: power distance (PDI), individualism (IDV), uncertainty avoidance (UAI) and masculinity (MAS). Long-term orientation (LTO) and indulgence versus restraint (IVR) were added later as additional cultural dimensions. Although this approach is data driven, it is not applicable to a global system like Wikipedia. Topic analysis algorithms do not necessarily require predefined topics rather they are able to adapt to the changing nature of data.

Wikipedia covers multiple different topics in a variety of languages. A clear segregation of cultures by topic is not possible. One reason for that is that many people use multiple languages. Even if they have a different actual intention, they may contribute to a particular Wikipedia project depending on their current working- or activity-background. Culture is one context, but obviously not the only one which influences the representation of topics within Wikipedia (see figure 11.4.2) for an illustration of the impact of the lingual context on a topic's representation in different languages. Therefore, Wikipedia seems to be a good source for advanced studies on lingual differences in knowledge formation and knowledge sharing, which is related to cultural contexts as well.

8.2. Growth of Wikipedia Projects

Wikipedia projects are more than just networks of pages. New pages are added over time. Pages provide information, they innovate and innervate new ideas, lead to questions, and as a consequence, more new pages are added and changed by different people. The editorial process can be highly controversial as Yasseri *et al.* [117] and Eckstrand *et al.* [171] show. The technical system (it consists of a server infrastructure to provide the core functionality) is embedded within user communities which consist of international editors and readers. Not all people contribute to Wikipedia, but a critical mass of users seems to be required by a Wikipedia project in order to survive. The evolution of Wikipedia project sizes was already analyzed by Ortega *et al.* [172]. They found that the contributions to Wikipedia are dominantly made by several so called "power users". Based on a calculation

of the Gini coefficients[1] for the top ten Wikipedias Ortega *et al.* [172] state that approximately 90% of all users are responsible for less than 10% of the content in Wikipedias of different languages . As in open source software projects, a small fraction of users are very active contributors. A comparable distribution of user activities in several other wikis - none of them are Wikipedia projects - was found by Stuckman and Purtilo [173]. Such a strong bias towards some very active users has to be taken into account. Furthermore, I think that analysis of the Wiki article life cycles should not only be based on the editorial activity as presented by Gorgeon and Swanson [174]. They studied the evolution of the *topic* (or *concept* as we call it) named "Web 2.0" in Wikipedia based on article size, number of editorial actions, and number of contributors. As a result, they define four phases for an article: *seeding*, *germination*, *growth*, *maturity* (for details see section 5 in [174]). The life cycle phases already take the activity and controversial character of editorial events into account. One can clearly conclude, that editorial activity does not always lead to an increase of content, because higher quality can be achieved by clear statements, which are often the result of shorter sentences. Too long articles can be seen as misleading or distracting. Too short articles are not providing background information. Several different categories or types of articles exist in Wikipedia. Ortega *et al.* [172] show, that article size distributions are bi-modal for English and Polish Wikipedia projects.

These studies ignore the network structure and embedding of articles. However, based on the node degree or on a centrality measure one can differentiate between leaf nodes, which contain definitions and well accepted facts, and more central pages, which are related to many topics, which define context as they aggregate several leave nodes. Such additional aspects show, that edit activity is not only related to a change of words or sentences but also to a structural change. Furthermore, the embedding of a page is important. In many cases it is even not possible to work with just one page, because the selected topic is represented by different linked pages within the same language.

Aggregation over all pages belonging to a topic - or even a full category - and contextual normalization within the local embedding was developed as a part of this work (see chapter 11). Such aggregated measures can contribute to advanced life cycle models.

A social network can be defined by interactions between people. Finally it can result in creation of a new resource in content networks or it can lead to a specific temporal state of minds among connected participants. One can analyze the underlying structure in both cases even if no explicit links exist. According to Borge-Holthoefer *et al.* [175] the evolution dynamics of a social community can be described by the size of the giant component, plotted as a function of time. Changes of growth rate can be interpreted as an indicator of existence of a particular social aspect, which might not yet be known or even has a clear physical representation. Also, topics in Wikipedia are formed by interconnected pages, not necessarily by categories only. By calculating the giant component of the page network we use more details - in particular the link structure - instead of just counting words.

In the next chapter, we apply a semantic analysis to the page content. Calculation of the semantic distance allows us to quantify the similarity of pages. Such a semantic similarity network can be compared to the static link network. This leads to a question: *Can semantic similarity be used as a pre-cursor for link creation?* But this work cannot answer this question yet.

8.3. Towards an Integrated Growth Model for Social Content Networks

As described in section 3.2, the *random graph model* is used to create new links between already existing network nodes with equal probability for all possible nodes. A second important model is called *preferential attachment* (see also [176]). New nodes are connected to an existing network, influenced by properties of existing nodes (e.g., with an attachment probability depending on node degree). Thus, nodes with many neighbors have a higher chance to get new nodes attached to them. In this model, new nodes can also be added without any link and even disconnected clusters can appear. Such simple models are helpful if only the final structure of the generated network should be analyzed. They cannot be used to describe the evolution of systems like Wikipedia entirely because they neglect the change of internal system states and the dynamic structure. They do not represent changes in the growth rate nor do they cover different phases in the system life cycle, which are characterized by variable growth rates and variable attachment probabilities. The goal of this section is to suggest a formal and generally applicable concept and to describe preliminary results from growth analysis, applied to data from four Wikipedia projects spanning a time range of 12 years.

To illustrate the model, I use the concept of radiation emissivity as an analogy. Although the analogy is weak it helps to understand the many facets within one coherent framework. Therefore, I compare Wikipedia with a physical body which consists of matter and has a given structure and temperature. In Wikipedia there is no such matter and also no temperature. Because content in digital documents can easily be copied one has not to care about conservation of mass (mass is seen as the equivalent to text content in this metaphor). In order to describe a flow of information we also have to track the embedding of the system. In the simplest case, it is surrounded by a field of information, which can be absorbed. This can lead to the growth of the system. In case of an equilibrium

[1] The Gini coefficient is a statistical measure to represent deviations from uniform distributions. It is the most commonly used measure of inequality in economic context, e.g., for levels of income or wealth.

- which is the ideal case - we can assume that we have a constant exchange of information between the system and the neighborhood. In case of Wikipedia, we can clearly say, the more information it contains, and the better the structure supports easy access to information, the higher the systems impact and its usefulness will be. With this in mind we can use the analogy and compare Wikipedia with a solid body, which exists in a field of radiation. The incoming energy flow leads to an increase of internal energy and to internal heating. The body emits energy according to its internal state. In an equilibrium state it emits the same amount of energy as it absorbs. A higher temperature is causing a higher radiation intensity.

In order to incorporate measured system properties into a formal description of the system's life cycle, a new integrated growth model is required. Inspired by the previously mentioned idea of radiation emissivity we use Eq. (8.1) to describe the process of network growth based on information aggregation.

$$\Delta I_{\text{system}} = I_{\text{link create}} + I_{\text{node create}} + I_{\text{node change}} + I_{\text{link change}} = \sum_{events} c_i \cdot v_{event} \approx a_{\text{edit}} \tag{8.1}$$

This model describes the growth of the system not only by counting pages and measuring text volume. Instead of volume we define the information content (comparable with a temperature) I_{system} which is changed by new links and new pages ($I_{\text{link create}}$ and $I_{\text{node create}}$). Beside adding new elements, which increase the amount of information, we can also change the internal structure of the network or the content by splitting nodes, changing text and changing links between pages ($I_{\text{node change}}$ and $I_{\text{link change}}$).

The difficulty of this idea is that it is based on a mean field approach, while the existing network growth models mentioned before are microscopic models.

ΔI_{system} is the amount of information which is absorbed by the system as a result of edit activity. This activity is not constant, instead it seems to be higher if more information is available. During the life cycle of the system, the contributions c_i of the different events i are also changing. In the beginning, we can see more creation of new elements. Later, change events dominate (see figure 8.3.a). Obviously, the creation of links and the creation of new pages are primarily structural changes. Additionally, content creation leads also to more information within Wikipedia. Because structure and context both contain information, the evolution or reorganization of the network structure leads to more information as well. If a large page is just split into smaller but interlinked pages, it is much easier to retrieve information. Relations to other nodes in the network can be found simply by traversing the links.

Context information is required in order to understand the meaning of page content and at the same time, the text and the link structure of Wikipedia pages provide such context information. This is used as implicit context for automatic information retrieval systems. A closed vocabulary (e.g., given by category pages or by an explicit ontology) are more examples for such context. Another data-driven approach for semantic context extraction was recently presented by Schwartz *et al.* [177] and is called *"The open-vocabulary approach."*

Finally, the inter-wiki link structure and the external links to referenced resources define the neighborhood and a multilingual context. A formal technical representation of semantic Wikipedia data is available as a semantic graph, as provided by the DBPedia project [158].

No matter how the semantic structure is provided or derived from data: if it exists, the Wikipedia pages can be used like a semantic network. Such implicit semantic links can also influence the growth process and thus it should not be ignored in a general growth model.

Our growth model covers the creation of new nodes as well as the creation of new links besides changes to the existing content and structure including the network topology. In the case of Wikipedia we can easily count the number of edit events. However, what goes on exactly during such edit events is not measured in our current study. Although each edit event is different and different activity leads to different results in detail, we unify this to one contribution for simplicity. Such a contribution covers one, two or all of the mentioned changes.

From the four selected Wikipedia projects we extracted all link creation events and all edit events. Each time a new link appears, also a new page can be created, if one of both pages to be linked do not already exist. All events are grouped by language and sorted by time stamp. Based on this event series the number of newly created pages n_N and the number of newly created links l_N is calculated at daily resolution.

A technical realization of a general growth model requires an integration of data analysis and simulation techniques. Assuming, that appropriate computational resources are available, one calculates the topological properties of the network as a function of time. This has to be done at a global scale for the full network and in order to allow local variations one has to track also the properties of all nodes' local neighborhood. Based on simulations it is possible to evaluate if applied parameters are consistent with available data. An important aspect is the variability of parameters in such a model, which become time-dependent values, derived from reference data sets, or simulations.

8.4. Properties of the Wikipedia Growth Process

In Wikipedia, the processes of adding new pages and adding new links between pages are coupled and cannot be separated from each other. In order to describe the growth of the four selected Wikipedia projects in more

detail we analyze the growth rates for the number of pages and the number of links separately. Because several different link types exist, we also compare the growth rates of the number of links for those types. We show the link-page ratio in figure 8.3.a. *internal links* are links within the same Wikipedia (same language) and redirects to another page of the same Wikipedia. Internal links represent semantic relations between the terms the pages are about or just relations between topics or concepts which are used within a certain page. If the meaning of a term is ambiguous, special pages help to show users all possible meanings (based on other pages). Such pages do not contribute much text, but this structural information is of a high value and increases the usability of Wikipedia. *External links* are links to another language (*inter-wiki links*) and links to pages outside the Wikipedia project (e.g., to external references). The frequency of such links represents an important quality indicator for Wikipedia articles.

8.4.1. Evolution of the Degree Distribution

All links between articles and links to external sources contribute to Wikipedia's structure. This structure and with it topological properties evolve over time. The creation of a new link is a result of an edit activity of a user. Figure 8.1 shows the temporal evolution of the internal link degree distribution for all pages of the Swedish Wikipedia. Redirects and external links are disregarded in this plot. Already since the beginning in 2002 the degree distribution can be described by a power law, with the exception of pages with a very low degree (low number of links). While pages are added over time, the distribution changes and its power-law shape becomes more obvious,

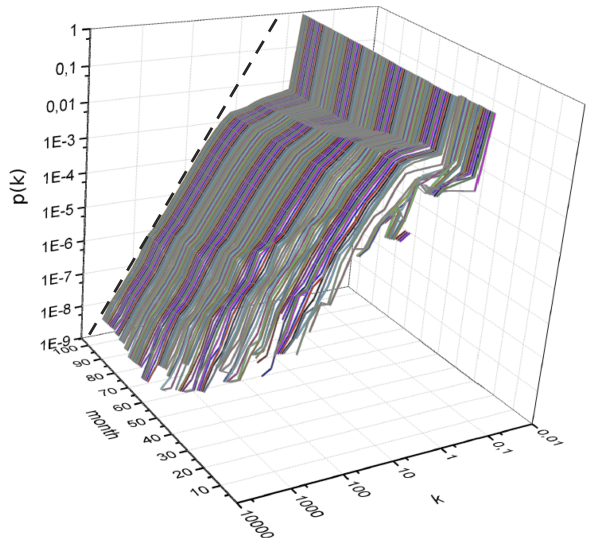

Figure 8.1.: **Evolution of Degree Distribution**. Degree distribution $p(k)$ (i.e., distribution of the number k of links per page) for internal links in the Swedish Wikipedia project. One curve is shown for each month from January 2001 till December 2009. The black dashed line illustrates a scaling exponent $\gamma \approx 1.67$. Note, that the maximum value for $p(k)$ in each curve is below one and $\int p(k) = 1$ for all shown curves which are normalized (nr of nodes with a given degree k divided by the total number of nodes availble in this particular month).

since the range of degrees becomes wider. Actually, most of the pages have much more than ten internal links and are well described by a power-law degree distribution. Only the number of pages with less than ten internal links is smaller than assumed in the Barabasi-Albert model that predicts power-law degree distributions. This also means, that the preferential-attachment model (which is also a scale-free model) overestimates the number of pages with a small number of links.

8.4.2. Growth of the Content Network and Structural Changes

Figure 8.2 (a) shows the total number of pages for four Wikipedia projects (Swedish, English, Dutch, and Hebrew). The number of pages $N_P(t)$ is growing by the number of new pages $n_P(t) = N_P(t) - N_P(t-1)$ per time interval $\Delta t = 1$ month. Figure 8.2 (b) shows the growth rate γ for an exponential growth model $N_P(t) = N_P(t-1)\exp(\gamma)$, which has been determined by $\gamma \approx n_P(t)/N_P(t)$. Note that an increased Δt has been used if $n_P(t) = 0$.

In the beginning the growth rate γ is quite large. Later, a tendency towards saturation can be identified. This shows that the character of edit events changed over time. In the early stage of a Wikipedia project most of the edit events are related to the creation of new pages, while later on the internal structure evolves. For the English Wikipedia project, one can see an intermediate regime with a constant exponential growth ($\gamma \approx 0.07$) as marked by the red line in figure 8.2.b. Such an exponential growth cannot be unambiguously identified for the Swedish

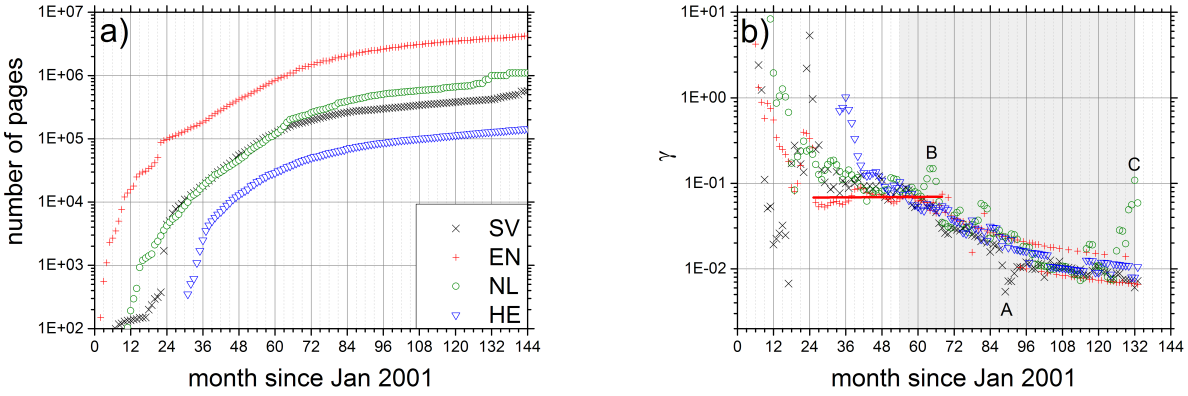

Figure 8.2.: **Comparison of growth rate for four languages**. For the Wikipedia projects in Swedish (black cross), English (red plus), Dutch (olive circle), and Hebrew (blue triangle) in (a) the number of pages and in (b) the exponential growth rate γ is shown. The grey area marks the time window were the decrease of the growth rate is comparable for all four languages. Automatic procedures like content restructuring and "bot-activity" or server failures are possible reasons for exceptions marked with A, B, and C. The red line marks a rather constant growth rate of $\gamma = 0.07$ for the English Wikipedia from 2003 to 2007.

Wikipedia. Interestingly, the page-growth rate has been drastically increasing during the last few months (in 2013, see exception C in figure 8.2.b) for the Dutch and – even more dramatically – for the Swedish Wikipedia (not shown). Actually, the Swedish and the Dutch Wikipedia started to create articles using bots.

Figure 8.3 (a) confirms that editorial activity tends to focus more on the addition of links than the creation of new pages during later states of Wikipedia evolution. It shows the ratio of the total number of pages $N_P(t)$ and the total number of links $l(t)$ as function of time. For all languages this ratio decreases during most of the time after a relatively large value (around 0.2, i.e., approximately five links per article) in the beginning. The final values are between 0.015 and 0.04, i.e. at approximately 25-60 links per article. For the Dutch and the Swedish Wikipedia the initial change (between 2001 and 2003) is quite sudden. In general, all four languages show a stronger decay of the page number to link number ratio in the beginning and a much slower decay later on. This behavior suggests that an exponential decay model may also be appropriate. However, we cannot find any regimes with unique or approximately constant decay rates for any of the considered four languages. The different decay rates of the page number to link number ratio might also be indicators for two different network growth processes.

The Swedish Wikipedia has initially ≈ 5 links per page and later the number of links per page increases to an

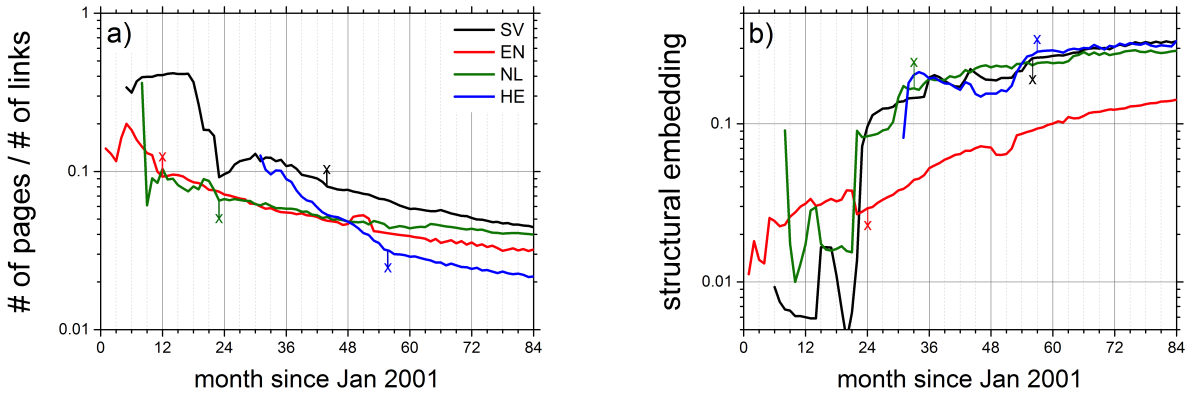

Figure 8.3.: **Contribution of content vs structural change**. (a) ratio of total number of pages $N_P(t)$ and total number of links $l_P(t) = l_{\text{int}}(t) + l_{\text{ext}}(t)$ (internal and external) as function of time between January 2001 and December 2008 for the Wikipedia projects in Swedish (black), English (red), Dutch (green), and Hebrew (blue). Note that the vertical axis has a logarithmic scale. **Change of internal structure vs embedding**. (b) ratio of number of external links $l_{\text{ext}}(t)$ to total number of links $l_P(t) = l_{\text{int}}(t) + l_{\text{ext}}(t)$ (internal and external) as function of time between January 2001 and December 2008 for the Wikipedia projects in Swedish (black), English (red), Dutch (green), and Hebrew (blue). Note that the vertical axis has a logarithmic scale.

average of ≈ 25. This is in line with the change in the degree distribution, which is shown in figure 8.1. Here one can see a continuous shift towards a dominating structural growth process, while the growth of content – measured in number of pages – becomes less important. The current ratio of page number to link number for the Swedish Wikipedia is quite similar to those for the English and the Dutch version, while the Hebrew Wikipedia has about twice as many links per article. During the quick growth of the Swedish Wikipedia article number in the last few months (see figures. 8.2 (a,b)), the article to link number ratio has slightly grown (not shown), which may indicate a slight change of the structure towards properties typical for Wikipedias at earlier stages of evolution. Although, this weak growth is still comparable with typical fluctuations of the ratio (just about twice as large), it may indicate that creating articles by bots leads to a step back in the quality of content. Next we separate the changes of internal and external link numbers. External links (to other language versions or references outside Wikipedia) are particularly important for confirmation of the article content and can thus be regarded as an important quality indicator for the articles. Figure 8.3 (b) shows the ratio of the number of external links to the number of all links (internal and external). The increasing curves show, that the ratio of external links grows for most of the time in all four Wikipedias. We note that there are two major groups of external links: just 'further reading' links (often in bad articles) and references (more likely in good articles). The habit of adding references increased in the last years, while one got more discouraged adding just simple links; they are usually also limited to 3-5 per article.

8.4.3. Phases and Phase Transitions

Finally we try to distinguish different phases - but in a much less exact way, compared to the clear definition of phase transitions in physics. If a particular property of the growth process is dominating, we consider this as a phase in the life cycle of the system. A more precise term could be *regime*. More research using more data is required before real phase transitions can be propagated.

In Fig. 8.3 (b) one can find indicators for existence of three regimes for each Wikipedia project. An initial phase, with balanced content and link contributions, followed by a second phase, with fast decreasing ratio of nr of pages and nr of links followed by a phase with a slower decreasa of page-to-link ratio. The transition time A (t_A) is determined from Figure 8.3 (a), which shows the ratio of total number of pages and total number of links. The transition time B (t_B) is based on the plot in Figure 8.3 (b), which shows the ratio of external and internal links. Table 8.1 shows the times where the qualitative behavior illustrated by figures 8.3.a and 8.3.b changes.

Language	t_A	t_B	$t_B - t_A$
SV	08/2004	08/2005	12
EN	12/2001	12/2002	12
NL	11/2002	09/2003	10
HE	08/2005	09/2005	1

Table 8.1.: **Analysis of life cycle stages.** The times when the qualitative properties of the Wikis change are the transition time t_A (determined from the ratio of total number of pages and total number of links) and time t_B (determined from the ratio of external and internal links).

For all four languages we find that t_A is before t_B, and the differences vary from 1 to 12 months depending on the language. This means, an internal structure formation starts before the external embedding is improved.

The Swedish Wikipedia has like the Dutch and the Hebrew Wikipedia a higher ratio of external links compared with the English Wikipedia. Systematically, all four Wikipedias have continued to increase this ratio (see Fig. 8.3.b). This is an indication for a very good reference quality of average articles in the Swedish, Dutch, and Hebrew Wikipedia. Note that the ratio of external links is very much lower in the English Wikipedia, just approximately half as large as in the Swedish Wikipedia (data from 2008).

Not shown in this version of figure 8.3 is a slight drop of the Swedish curve in the last months. It is probably associated with the drastic increase of the total number of articles (see Fig. 18 in [10]). However, it is too weak to be considered as an indication of a drop in article reference quality, and there was a significant larger increase during 2012 just before the slight drop. We note that bot generated articles usually have a quite high density of references, meaning just one sentence but 2-3 references to publications which, however, may not be linked using a web link.

Summary

As shown in this chapter, one can measure and study the life cycle properties of Wikipedia projects, based on access activity logs and the page edit history. Simple system properties, such as number of pages, number of links and their change rates were shown in Figures 8.2 and 8.3. Since each new page and each link creation event can be extracted from the Wikipedia edit history and via the Wikipedia API, also real time studies are possible in the future. Although Wikipedia does currently not provide aggregates of edit events on a daily or an hourly base, it

seems to be reasonable to provide such data beside the available access log data. This would enable the research community to study coupled dynamic processes of content creation and information consumption on a global scale with much less overhead, which is currently caused by expensive data extraction and pre-processing procedures.

9. Modeling Complex Systems as Networks to Connect Physics, Social Science, and Economy

> Reality is not a function of the event as event, but of the relationship of that event to past, and future, events.
>
> *(Robert Penn Warren, All the King's Men)*

Large data sets allow better statistical accuracy. Combining different types of data enables multi-faceted models. Both of this is recently facilitated by the rise of affordable data analysis engines and large-scale storage systems.

Some well-known pioneers of large-scale data analysis are companies such as Google, Yahoo!, Facebook, Apple, and Twitter. Some of their research activities are rather unknown outside the companies, but nevertheless, the scientific community benefits from recent technological improvements.

For example, Google has continuously improved its ranking algorithm since its invention. Recently, a Google research team has included a measure of a page's trustworthiness. Historically, the dominating factor in search was a page's reputation measured by the page rank algorithm. Now, they simply count the number of incorrect facts within a page, as they assume, that a page with less wrong information should be considered to be more trustworthy. [178]. This *'Knowledge-Based Trust'* approach uses a multi-layer model to represent many relevant aspects without the negative side effect of loosing granularity by early aggregation. Google also started early to provide a rich personal user adaptation. Therefore, they contextualize search results according to a user's language. The browser history and cookies are inspected, to learn more about the user. All this information is collected and merged into a final result: the scores of search items.

For advanced analysis, especially for studies of system dynamics, such tight integration of many factors into one score is not appropriate. In order to understand how a system responds to an external influence, one has to measure and analyze the system's properties as a function of time. Google offers access to some internal data. In order to create new studies, new experiments and new evaluation methods, one has to build specific multi-layer models.

Physical reality defines an important facet of human life. Science is part of our life and has the goal of providing insights into our life and explanations about our observations. Scientific models should simply help to understand what surrounds us and how we can interact with this neighborhood. The better our understanding of nature, the better is our ability to adopt to it and even to influence it - this includes also the non-physical aspects like those studied in social science and cognitive science. Describing nature is not simple, especially because many different things are highly interdependent, and over-simplification would lead to insufficient results very soon. Even if we are able to understand and influence nature, it is an ongoing controversial problem, to figure out, if individuals, or organizations should be proactive and take control. In many fields this is accepted and part of our long-term strategies. There are many others, which provide a base for scientific and political debates. Many discussions are based on scientific expertise, others are influenced by beliefs and even fears.

Independent from final decisions, whether to influence nature actively or not, modeling enables us to analyze and simulate conditions, which probably are not desired in reality. In this way, we gain insight and deeper understanding for better contributions to ongoing and upcoming discussions, or just for short-term decisions in a smaller private or business context.

System theory provides concepts to understand and study complexity. Process models and well defined procedures including automation allow us to handle complex systems, but we loose control as soon as complexity increases. This can happen if such systems interact with each other. Feedback loops are a characteristic property of complex systems. Modeling is considered to be a neutral scientific approach and a convenient way of turning hypotheses, ideas, and data into knowledge about any complex system.

No matter on what scale our research is done, we can choose strong simplifications, which usually means, we also disconnect things from their neighborhood. It is important to find approaches which allow us to do our analysis without such a harmful cut in order to minimize the impact of simplification on results. If disconnecting subsystems from each other can not be avoided, it is even more important to understand the impact of this segregation on final results. The question is now: Are the results really applicable? Or do they mean nothing, because things behave very different if they are disconnected from each other?

Active measurements can have an impact on the system, about which data is gathered. Experimental setups can also influence the final outcome. We cannot easily stop doing simplified experiments, but we have to find ways to obtain data from undisturbed interconnected large-scale systems to build more realistic models. This means, that data collection procedures should not influence the systems directly and it should also not directly depend on it. A common goal is to design experiments with as less influence on participants as possible. By using so called real world data we can already reduce the artificial impact of measurement procedures and experimental setups. Especially mobile communication devices, embedded sensors in cars, cameras, and counters in airports and train stations, but also activity logs on web sites and payment logs provide a lot of data which support advanced studies of human behavior with no direct impact on the source. Modern mobile phones have more functionality than early personal computers. They have a lot of built-in sensors, such as GPS antennas, acceleration sensors, and temperature sensors, but also microphones and cameras. All these rather cheap components can be used for experimental research [179, 180, 181].

Crowd sourcing has become very popular in the last years, not only in economy also in research. This shows that modern scientific methods are integrated directly into business processes. Furthermore, it indicates, that science is not necessarily bound to product development or specific fields of application. Scientific data analysis, nowadays also called data science, becomes more important to get information from real world processes to support operational optimization, especially in critical or disaster contexts.

Experimental design in social science is not free of risks or problems, e.g., one has to care about selection bias, especially in case of social-network studies. Using internet related systems as single source - even if mobile access is possible - the results are influenced by the limited accessibility of the Internet. Infrastructure, economic status, and political decisions in individual regions on earth are affecting Internet accessibility and can thus not be ignored, especially if global systems of our society are modeled.

A simplification is possible by focusing on only one subsystem. One has to define the system's boundaries very well. In the worst case, nothing is known about the surrounding system - which means that an unknown non-quantifiable bias might exist.

Physicists study many-body-systems which consist of a large number of interacting objects. A clear description and improved understanding of the micro-, meso-, and macroscopic properties of such systems are the objectives of a vast category of physical problems. This category is called many-body problems. Beside analytic approaches

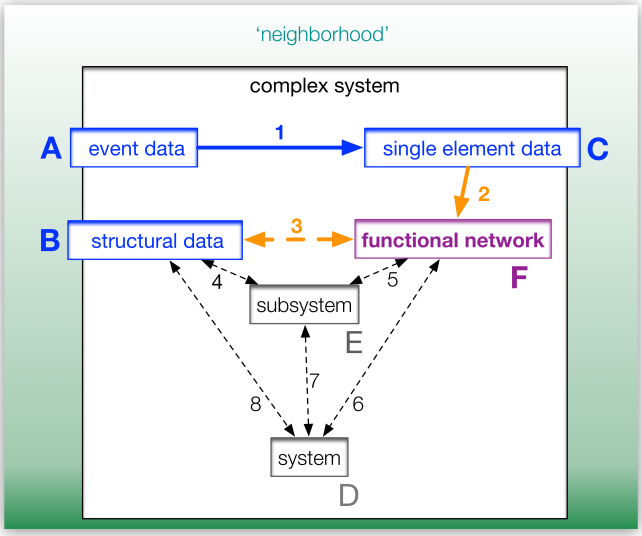

Figure 9.1.: **Accessible and hidden properties of a complex system**. To describe and study processes in complex systems, multiple data sets are required and combined (blue boxes). Because of their hidden nature - many properties are not measurable directly (purple box) - some aspects can only be calculated. For simplicity, many research projects focus on individual properties, often represented by one single network layer (E). Networks of networks promise more detailed insights into the nature of complex systems. Arrows indicate dependencies between measurable data (A,B,C) and hidden aspects (D,E,F). Aggregation of events in step 1 leads to time series data (C) which provide metadata. System elements describe specific characteristics of individual interacting parts of a complex system. For each element (C) it is possible to measure properties also directly. (B) represents the system structure - at least the obvious part. Step 2 illustrates the calculation of functional networks from individual element properties including time series. 3 represents a bidirectional dependency between structure and functionality. Static and functional structure influence subsystems (4,5) and the overall system (6,8). 7 indicates that the subsystems influence the whole system and visa versa.

in continuous space and many-body perturbation theory, various discrete numerical methods like lattice-gas and Monte-Carlo approaches have been developed and applied. Furthermore, it seems to be a reasonable approach to apply these methods to social science [182].

Although a direct transformation of the methods is not possible between different fields, according to Kulakowski [182] it is worth to try possible applications as starting points even if based on rather weak analogies, before stable scientific methods are available.

Finally, it is important to integrate multiple research disciplines. This is what *complex systems analysis* nowadays stands for. Data-driven methods, large-scale data processing, and high performance computing are the related technical aspects, which enable integration of economy, social science and physics by using networks as modeling technique. Network models aim at diverse many body systems with limited simplification and a high degree of contextual embedding.

Figure 9.1 illustrates dependencies between measurable data and hidden system properties which can be analyzed on microscopic, mesoscopic, and on macroscopic scales. Hierarchical models can also be created in this way.

As an example one can think of crowd analysis, focused on measuring the mood of a large group of people. One way to access the state of individual persons is to listen carefully to what they say. Communication between people is an important source for information. The approach depends strongly on the location of people and on their communication style. A silent crowd and a group of loud crying people walking along a street show obviously different moods. In this case, the measurable intensity of the sound that the crowd produces is a quantity, which can be related to the mood. A different approach is required if written communication is used, especially if sender and receiver are not directly connected to each other. Text analysis on exchanged messages and sentiment analysis are used to access the hidden properties of the communication which represents the mood of participating people [100, 183, 168]. Ideally, one would combine both approaches to investigate consistency and to cover multiple channels as this reflects reality in a more natural way.

Comprehensive analysis about the perception of a particular brand is another example. Also in this case one should combine data from different channels, such as Google, Twitter, Facebook, a company's own website, public communities, and also company internal communication channels. Using only public campaign data is dangerous as long as no knowledge about the system around the campaign is available.

This work introduces a method to implement reference studies based on public open data from Wikipedia. Our new approach can be generalized to arbitrary data sets. It is common to combine proprietary data with open public data. This is another advantage of large-scale data analysis, because not all research groups or companies have to own and maintain all data if they collaborate and combine their data sets.

System properties like long-term memory effects can be revealed from communication patterns, no matter if the traditional approach of analyzing the time intervals of letters sent between two persons or a modern approach is used (see Oliveira and Barabasi [183]). In the later, more and more digital messages, mobile communication devices, and multiple online systems are involved. This indicates already, that modeling techniques should be of a hybrid nature, as they combine time series and network analysis procedures.

Beside exchange of information and decision processes also dynamic motion patterns of human crowds have been in our research focus during the SOCIONICAL project. Interactions between persons and also between persons and their environment (in simple buildings or in a large stadium) were modeled using the *Social Force model* (SF, see also section 4.1)[17, 13]. Alternatively, I also developed numerical simulations for analysis of motion and information flow using the *Lattice Gas model* (LG) [184, 185, 186]. SF describes multiple aspects of interactions, which finally are superposed as forces. This superposition of attraction and repulsion forces lead to the motion of things in space.

Some qualitative properties of social networks, such as different types of relations are modeled by different interaction rules. Interaction rules are used instead of forces to model the dynamics in LG. Positions in real space and internal state properties of the elements (particles or agents) are updated depending on several conditions, and based on logical reasoning. This leads to positional updates of agents on the underlying lattice and to a specific network representation of the entire system. Topological properties of such snapshot networks, also called temporal networks, represent specific aspects, and allow studies about their evolution in time.

Although different types of interactions co-exist in both models (SF and LG) the goal is to keep the models simple by a limited number of different interaction types. Simulation and analysis of dynamic system properties is also possible in both models, but the internal structure of the system is not accessible directly, although the interactions are inherent in both models.

In integrated models, e.g., in a social network for people which participate in public events and communicate via mobile devices at the same time, the location of objects, such as persons, in real space is not the only relevant aspect. Furthermore, multiple interaction rules exist. The intrinsic structure of social systems - such as family relations or hierarchies in organizations - have an important influence. Such relations are usually entirely hidden but they must not be neglected.

Structural system properties of, e.g., social networks are analyzed with a variety of methods, but it is often hard to identify and to describe the hidden interaction roles. In general, all established network analysis algorithms require a predefined adjacency matrix as representation of the network. Because the internal structure is at least

measurable one can relate the change in structure to the variation of interaction roles.

Scalar entries of an adjacency matrix can show only one single type of interaction between two elements. This means, that multiple interaction types require multiple matrices. Such matrices are considered to be layers in a multi-layer network and each matrix describes one single facet or aspect of the system. All layers together represent the entire system including its natural embedding.

Traditionally, each layer has been studied independently. In order to describe emerging properties, the individual networks have to be combined. The result is a multi-layer network with different link types. Because different aspects can have different influence on the systems time evolution. Knowledge about the right scaling and the right weight-functions is required.

The majority of established network analysis algorithms can not handle multiple weights for one node or link, which means, they can not handle multi-layer networks directly. Some special cases are handled by bipartite or k-partite networks. Both have the limitation that no interaction between nodes of the same type are considered.

This chapter presents a generic approach to construct multiplex- or multi-layer networks from dynamic node properties. This allows calculation of dynamic correlation properties. Distance and similarity measures as structural metrics are derived from measured time series data. In this way, several computational procedures provide network layers with different meaning.

The following sections describe a formal framework for network creation procedures followed by a discussion of the application of these networks.

9.1. A Formalism for Network Reconstruction

One essential question that helps us to prepare useful networks as models of complex systems is: "How can the relevant interactions be described as network links?" Such interactions happen on multiple levels. Individual elements interact with each other (see subsystem E in figure 9.1). The collective behavior of a group influences an individual element (or person), but, all individual actions of many elements (or people) contribute to and define the collective group behavior (arrow 7 in figure 9.1). Therefore, multiple link layers are combined. A combination of layers creates in general k-partite networks with multiple types of links between nodes of different types. An integration of multiple interaction types into one unified model can lead to a hierarchy of interactions and thus a hierarchical network, or to a network of networks, where individual sub-nets are not part of a hierarchy.

Multiple layers can be defined in a hierarchy of abstraction levels. This allows microscopic and macroscopic properties to be combined in one single model. The multi-layer approach connects isolated views within one single system view and can be handled either with statistical analysis methods or numerically, or via simulations. The advantage of this approach is the combination of simulation and analysis techniques in one single theoretical and technical framework.

Link projection is required, to merge the contribution of individual link layers. Each layer represents in general one single aspect, e.g., a reaction channel or a communication channel. In our framework we introduce a *connectivity projection function* (CPF), to calculate a single link strength value for each pair of nodes from multiple link layers. Such a new link strength property includes and combines information from all available layers. Thus, a CPF reduces the number of dimensions of the system. The resulting network representation can emphasize one individual aspect without decoupling or disconnecting a component from the entire system. A comparable approach is used in the *Multiplex PageRank* algorithm to incorporate the intensity of the interaction between network layers. Halu *et al.* [82] define Additive, Multiplicative, Combined, and Neutral versions of Multiplex PageRank in order to show how each version reflects the extent to which the importance of a node in one layer affects the importance of that particular node in different layers.

Technically, nodes are represented by labeled vectors. The adjacency matrix becomes a tensor and the elements of this tensor can also be vectors instead of scalar values, e.g., if time series are taken into account. A simple summation of the link-vector components would not provide useful result. Therefore, a similarity measure (in Eq. 9.4 it is called link creation function with symbol $\mathcal{F}_{\mathrm{LC}}$) is used to calculate the strength of a link between nodes based on their microscopic properties and on the network properties of the close neighborhood. The neighborhood of a node defines a subsystem which also influences the properties of a particular node. This approach uses a direct coupling of node and system properties and leads to a closed feedback loop (in case of directed links) or closed triangles in general.

In order to model a system as a multi-layer network one has to select the appropriate type of measurable data (see A, B, and C in Fig. 9.1). Depending on the characteristics of the time series data one has to select an appropriate link creation function (see table 9.1). Finally it is important to define the direction of the links. This influences the selection of the right metric as connectivity function, because not all possible link creation functions provide information about orientation.

Our goal is to study dynamic properties of complex systems without full segregation of static and dynamical properties. There are many more questions that all influence the model definition procedures, such as:

- How fast is some internal system property changing?

- What metric allows us to describe this change as function of time?

Network reconstruction is the process of defining and quantifying relations between entities. Such links can be obvious, such as the relation between two people living in the same place, or hidden, such as a shared political opinion, which is not expressed in a direct way. Studies on opinion dynamics, crowd dynamics, election dynamics, and on the influence of religion on decision making processes are research topics with growing interest.

The article titled *'Modelling Opinion Formation with Physics Tools: Call for Closer Link with Reality'* [187] criticizes the lag of connection between research, created models, and application to real world problems, and even more importantly, the lag of analysis of real-world data. Especially the emerging amount of apparently unlimited data sources can help to change this in the future.

One contribution of this work is to formalize the procedure of reconstructing links between nodes from disjoint data sets. The formalism allows a comparison or at least a validation of comparability of obtained results. Before such a similarity relation can define a link, the similarity matrix has to be transformed into an adjacency matrix. This is done by filters, with a fixed or variable threshold, probability filters, or based on structural information (see chapter 10).

The difference between construction and re-construction is that in a constructed network all existing links really are obvious. The links exist physically. A reconstructed network is generated from data, which do only indirectly describe a relation between the linked nodes. In this way it is possible to find hidden links, which are not physically obvious. Network re-construction is a statistical method. If results are valid or not has to be validated by significance tests (see chapter 10).

Because a large variety of network reconstruction methods already exists and selected parameters for each method directly influence the outcome, it is worth to define a generic network reconstruction procedure.

Especially in case of complex dynamic systems it is important to compare network properties quantitatively. Ideally, one would define equations of motion based on the configuration properties of the system. Structural analysis of systems with obvious relations between elements or components is already well established. If links physically exist, then it is simple to measure their properties. Many obvious relations exist only in the information domain. The measurement process has to be replaced by data acquisition, and data linking procedures.

An example is the relationship between a mother and their children. In the beginning, before birth, a physical connection really exist. Later on, this connection is in general only based on the knowledge about the relation, which allows us to track the family structure. Such family networks are not static. Usually three to four generations live together at a given time. According to the book of *'Guinness World Records'* the maximum number of generations alive in a single family has been seven. Each individual birth adds one more well defined node to the network. At the same time all existing links can be interpreted in a new way. Each generation has it's own specific habits, attitudes, and beliefs. This requires additional interaction layers and additional link types for each different interaction role. Structural differences in such networks seem to be related with social structure of society on several interaction levels, from family, to local community, to a whole country including its economy.

In many cases it is not important to track all details and all members of a family nor all possible interactions between them. Network nodes can simply be connected to a common network node which acts as a stub for a particular aspect. The result is a local star structure (see Fig. 9.2). If instead of the additional node a relation between all members is generated then we would have a clique of highly connected nodes. Which approach is better? Without information about the planed analysis procedures this question cannot be answered.

We measure the change rate of a system property of an object which is influenced by the process we want to study. For network studies, this means we have to identify measures, which allow to quantify external influences. Especially if structural information is not directly accessible we investigate the structure of functional networks, created from time series obtained by objective measurement procedures on individual system elements.

Such functional networks are useful to compare different processes even if there is no direct accessible variable to measure. Our approach is based on a comparison of the impact of external influences on a system by analyzing hidden variables, which are influenced or changed by this external process.

In our examples we used Wikipedia, growing over time primarily due to added new pages. Measuring the total text volume of pages seems to be a good indicator to quantify the growth or knowledge formation process. Links are very important in Wikipedia, and links contribute also to the content. Page links form the backbone of the system and allow navigation and structural analysis. Beside these variables we can also track all pairs of pages which are edited or used in parallel, even if they are not linked to each other. This reveals overlapping interest in different topics even if no links exist between them at a given point in time. Such non-existing links cannot be analyzed directly because they do not exist.

This is why we apply computational methods in order to create functional links. It is important to note that properties of such functional networks depend strongly on the data aggregation and pre-processing procedures. An intermediate result of the network reconstruction procedure is a time-dependent adjacency matrix. This matrix can be analyzed in several ways. Depending on chosen computational methods one gets *'structural metrics'* which represent the entire system. Structural metrics quantify the impact of an external process on the system's internal properties, even if no variables are accessible directly. Figure 9.2 illustrates a process with impact on the system's structure.

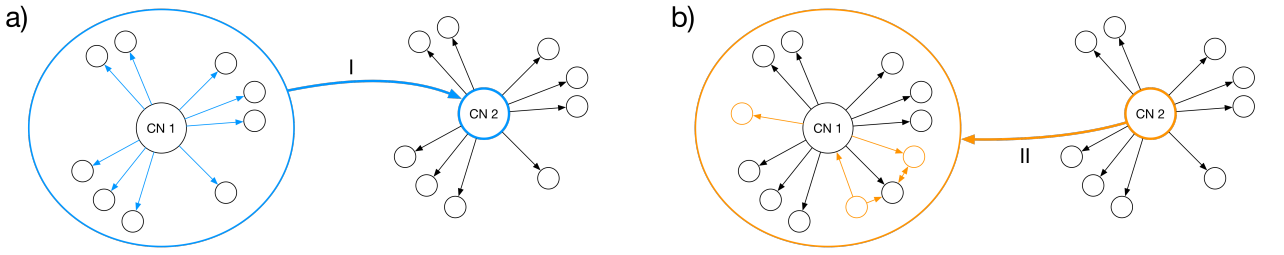

Figure 9.2.: **Dynamic processes can cause inter-dependencies between layers**. The multi-layer network approach introduces a feedback loop, although only one directed link between two example nodes CN1 and CN2 exists per layer. The functional links (I,II) represent a process on top of the static network (black arrows). Structural metrics allow time-dependent studies on large systems without a need of '*over simplification*'.

The multi-layer network approach leads to feedback loops (see figure 9.2) which are responsible for typical properties in complex systems such as non-linearity and emergent behavior.

Figure 9.3 shows two different networks reconstructed for one system. Obvious links are used to calculate an initial graph layout (see fig. 9.3.a). Clusters are colored to highlight the community structure of that network. This static facet does not explain the dynamic properties. Therefore we show the functional network from access-rate time series (see fig. 9.3.b). The size of the nodes indicate the node degrees within the networks. The node degree for each node can be different in each layer. Based on network metrics it is now possible to track the system over time. Analysis of dependencies between several aspects can be covered in future work.

The layout of the two networks in figure 9.4 uses geographical embedding. Some Wikipedia pages regarding cities provide latitude and longitude values directly, others are linked to a page for which geo-location data is available. Structural differences between the two network layers representing two different processes are visible. Detailed quantitative analysis of such systems requires additional algorithms. Our goal for this chapter is to define

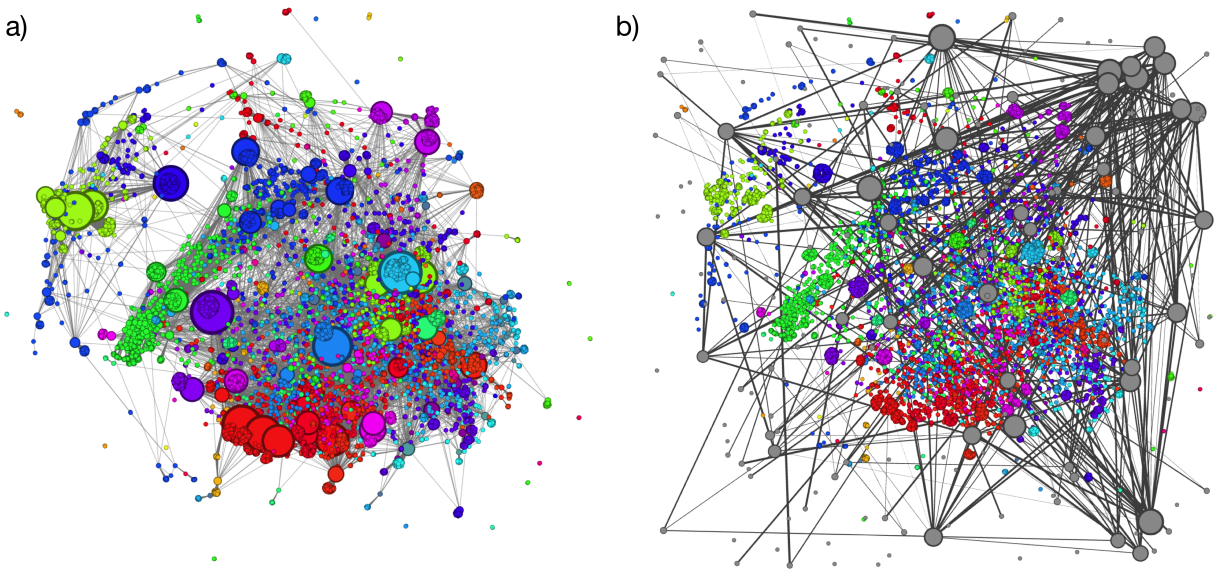

Figure 9.3.: **Comparison of the local networks** regarding topic '*Illuminati (book)*' based on (a) direct Wikipedia links between all nodes in the local and global neighborhood (CN, LN, IWL, and GN), and (b) functional links calculated from user access-rate time series. In (a) the node colors reflect the community structure of the underlying static network. The node size in both networks is proportional to the degree k. An un-directed correlation network is shown in (b). Correlation links (undelayed cross correlation) are filtered by link strength ($l_s \geq 0.75$). The layout was calculated in Gephi [18] with the *ForceAtlas2* algorithm based on the static link network, but in (b) the correlation links are plotted instead of the structural links.

Network Type	Functional Networks *for dynamic processes*		Content Networks
	time series data	*event series data*	
un-directed	Pearson cross correlation* Spearman rank correlation	Event synchronization, Q* [135]	n-gram co-occurrence cosine similarity* [188]
directed	Granger causality [189][190] partial correlation* [23]	Event synchronization, q* [135]	Semantic similarity* [191] Hyperlinks

Table 9.1.: **Link creation functions**. Different types of node interactions exist, e.g., there are functional and structural aspects. The table presents measures, which can be applied to Wikipedia time series data (generalization to other socio-technical or socio-economical systems is straight forward). The formal description in Eq. (9.4) uses the symbol $\mathcal{F}_{\mathrm{LC}}(\mathrm{G})$ for any pair- or set-function applied to time series tuples G (not G, which represents a graph as in table 9.3). The functions here allow creation of individual network layers. Marked (*) methods were implemented in the Hadoop.TS software package [11] as part of this work.

a framework for graph reconstruction - even if structural analysis algorithms are already involved at this level, network reconstruction is in the focus, not yet a detailed study of dependencies between individual properties or network levels.

The following equations describe a formal procedure to calculate structural metrics $\mathrm{S_m}$ for arbitrary complex systems based on time series data:

(1) Raw data, usually multiple sets of events $\mathrm{E}(t)$, have to be aggregated (via function $\mathcal{A}$) in order to create time series $\mathrm{X}(t)$. Measurement devices or IT systems offer direct ways to access pre-aggregated data sets with raw time series $\mathrm{X}(t)$. Measurement procedures can also be replaced by numerical simulations.

$$\mathrm{X}(t) = \mathcal{A}(\mathrm{E}(t)) \tag{9.1}$$

(2) This raw data set is processed by a time series creation function $\mathcal{C}_{\mathrm{TS}}$. A creation function can be, e.g., an (extreme) event detection function or simply a filter. Peak detection algorithms produce event series (see section 5.4.3). Results of this step are multiple time series denoted as $\mathrm{X}'(t)$.

$$\mathrm{X}' = \mathcal{C}_{\mathrm{TS}}(\mathrm{X}(t)) \tag{9.2}$$

X' is represented by a time series bucket and contains a set of time series with compatible properties. This means, they all have data for the same metric, using the same time resolution, length, start, and end time.

(3) Those time series are processed by a grouping operation $\mathcal{S}_{\mathrm{TS}}$. Groups define the context of the analysis. From this groups we create pairs and triples of time series. Depending on the chosen network creation function a set of two, three or n-dimensional vectors[1] is generated. The vectors contain individual time series as components and represent potential links.

$$\mathrm{G} = \mathcal{S}_{\mathrm{TS}}\left(\mathrm{X}'\right) \tag{9.3}$$

Creation of all pairs or triples of time series can be done as part of the analysis procedure or prior to it. For large data sets this operation can be very expensive. Especially if multiple algorithms should be compared, it is useful to store the results G and to reuse them.

(4) Now, we apply the link creation function $\mathcal{F}_{\mathrm{LC}}$ to the time series tuples G. Distance or similarity measures (as listed in table 9.1) require time series pairs, whereas triples are used for dependency networks.

$$\mathbf{A}_{\mathrm{m}} = \mathcal{F}_{\mathrm{LC}}(\mathrm{G}) \tag{9.4}$$

The result of the operation is an adjacency matrix $\mathbf{A}_{\mathrm{m}}$, which has to be cleaned by a link filter.

(5) We apply a link filter $\mathcal{F}_{\mathrm{LF}}$ in order to remove non-relevant links.

$$\mathbf{A'}_{\mathrm{m}} = \mathcal{F}_{\mathrm{LF}}(\mathbf{A}_{\mathrm{m}}) \tag{9.5}$$

After filtering an adjacency matrix $\mathbf{A'}_{\mathrm{m}}$ is available for a structural analysis per layer or by using a connectivity projection function (CPF) another adjacency matrix $\mathbf{A'}_{\mathrm{c}}$ is generated from multiple layers.

$$\mathbf{A'}_{\mathrm{c}} = \mathcal{F}_{\mathrm{CPF}}(\mathbf{A'}_{\mathrm{m_1}}, ..., \mathbf{A'}_{\mathrm{m_n}}) \tag{9.6}$$

The index m indicates a particular measure or metric and c stands for *combination* of multiple layers.

[1] The vector is a data structure, and not the mathematical vector in this case.

(6) Methods from random matrix theory [192] as well as traditional network analysis procedures can now be applied to obtain the structural metric S_m. In our case S_m was obtained by application of graph analysis algorithms $\mathcal{T}_G$ to the previously created link matrix. Usually, the topology of the network layers or the entire system is studied.

$$S_{m_{layer}} = \mathcal{T}_G\,(\mathbf{A'}_m) \qquad \text{or} \qquad S_{m_{system}} = \mathcal{T}_G\,(\mathbf{A'}_c) \tag{9.7}$$

In order to get a time-dependent structural metric $S_m(t)$ one has to use time series episodes of length l. For each time interval i defined by t_i and t_{i+l} one can obtain a value S_m, which can now be related to raw data, in order to identify the relation between function and structure.

$\mathcal{A}$ is an event aggregation operation. $\mathcal{C}$ is a time series creation method. $\mathcal{S}$ is a time series set operation. $\mathcal{F}_{LC}$ is a multivariate time series analysis procedure. $\mathcal{F}_{LF}$ is a filter function based on properties of links or based on structural properties, such as *Spanning trees* or *Planar maximal filtered graphs* (see next chapter). $\mathcal{F}_{CFP}$ is an arbitrary function, which provides a combined link strength $l_c = f(l_i)$ based on all available link types l_i. A typical example is a linear combination of individual links: $l_c = \sum_i c_i \cdot l_i$.

$\mathcal{T}$ is a graph analysis procedure which provides topological properties of the resulting network layer.

The next sections of this chapter cover the link creation phase (4) (Eq. 9.4) and the link filter phase $\mathbf{A}_m \rightarrow \mathbf{A'}_m$ (5) (Eq. 9.5). Both phases are part of $\mathcal{F}$, the link reconstruction operation.

9.2. Reconstruction of Multi-Layer Networks

A variety of network types exist. Literature such as [86, 193, 194, 39, 195] about networks and network analysis algorithms usually covers network models and specific applications but network recreation procedures and functional networks are rare. Very recently, Petter Holme published a review on new developments in temporal networks [49]

Arbitrary analysis algorithms can easily be written down, but a particular implementation can be difficult, especially if the network is too large to be stored on one single computer. Some algorithms are limited to specific network representations. A simple transformation from one into another representation is required. From a mathematical point of view this may not worth to be mentioned in some cases, but the technical and economical dimensions are important here. Even if the data is not manipulated, re-organization such as re-partitioning of a huge data set can be very expensive. This means that huge demand for large-scale compute resources with large memory, and long processing times are typical, even if the analysis procedure is not very exciting from a mathematical perspective. Recent research and engineering work (see [63] and [62]) introduced generalizations of existing data models for graph representation.

Not only technical requirements defined by analysis algorithms influence the representation of networks. Also the analysis goals have an influence. Transformations between different types are not all reversible, this means, in some cases the resulting representation is more compact and therefore more efficient, but information can be lost during the transformation procedure. To solve this problem, we handle all raw data separately. As long as

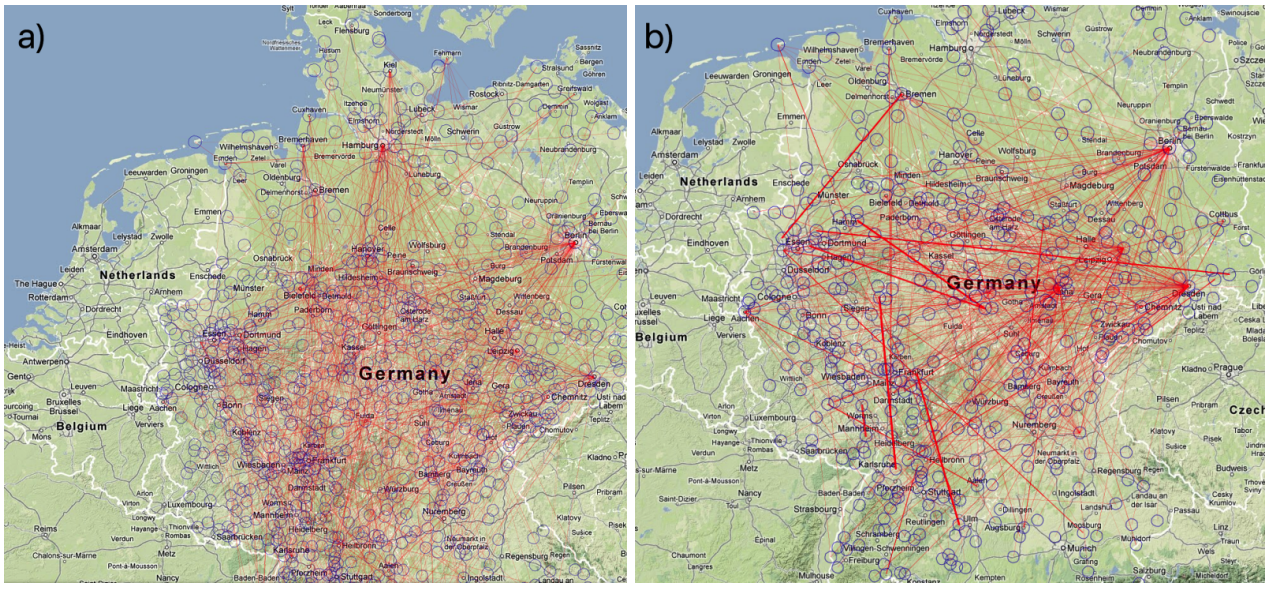

Figure 9.4.: **Comparison of two functional networks regarding German cities represented by their Wikipedia pages**. Different aspects of complex systems are highlighted by functional networks, which finally can be integrated in a multi-layer network approach. (a) shows the functional network for access activity and (b) illustrates the editorial activity.

initial data is available for each individual network node and link, it is possible to recreate all intermediate results on demand. Two access pattern have been found very useful. First, all node properties are initially stored in a key-value store. Grouping, sorting and filtering can be applied to such data in parallel. Finally, random access to individual node properties, based on the node id is essential. For known static networks, it is very efficient to store all in-going and all out-going links in an adjacency list. This means, based on a single node key, all available information can be retrieved by only three requests per node from storage. Link properties require complex keys - consisting of three components at least. A link exist between two nodes, called source and target. In general the order is relevant, but not for symmetric links. Because multiple different link types can exist between nodes, the type has to be defined by a metric name, which also will be used as a part of the key to address the link property data. For simplification we work with full link matrices. This requires memory to store $N^2 \cdot m$ values, where N is the number of nodes and m the number of metrics.

Different real world scenarios require special network representations. Very simple models require only one type of nodes and one type of links, which can even be defined by a fixed link strength. Weighted links are already more flexible. Multiple node types lead to bi-partite or k-partite networks. Multiplex networks are used to describe multiple interconnected aspects of one complex system in one model. Table 9.2 lists important network types according to typical applications.

Characteristic topological properties are often derived from network snapshots, taken at a given time. What does such a snapshot represent? Is it information, aggregated over a period of time, is the graph static, which means there is no change, or is it just static within a very short time range, which makes it quasi-static. Table 9.3 compares three important graph types with respect to time.

Table 9.1 shows useful link creation functions $\mathcal{F}_{\mathrm{LC}}$ for documents and time series, obtained from Wikipedia. The approach is also useful for other types of complex systems, in which social interactions are influencing documents, such as messages or web pages. Because the interaction of human beings also influences economy, it seems to be reasonable to use such data also to study the interaction between human communication and the economic processes. Wikipedia pages are used as stub, which represents, e.g., financial markets, companies, or products. This allows to measure user interest in specific topics. More details and preliminary results can be found in chapter 15.4.

The remaining part of this chapter shows the reconstruction of layers for multi-layer networks as individual separated steps. The approach in general is still using a simplification technique. Data is extracted for one individual aspect - other data is ignored. In this way, the layers are still disconnected from the complex nature of the entire system. The complexity is lost only temporarily. In a final step, when all aspects - where each is represented by an individual layer - are integrated in a multi-layer network, or in a network of networks, we can see each aspect embedded within the original context again.

Another approach is, e.g., comparison of different embedding scopes. By comparing the calculated structural properties for two networks with different embedding it is possible to measure the influence of contextual-variation.

Network Type	Characteristics	Applications	References
Simple Graph	Only one type of edges and vertices (respectively nodes and links) exists. Their properties are not time-dependent, just scalar values.	The majority of graph algorithms require such a representation.	[196]
k-partite Network	One type of links and k types of nodes exist.	Cluster detection and identification of central nodes are used to highlight hidden roles of, e.g., the "influencers" in social networks, or information shortcuts.	[197, 198]
Multiplex Network	Many types of links but only one type of nodes exists. For each link type a simple network (called layer) appears.	Analysis of multi-channel processes requires different link properties for each process. Parallel processes are modeled as layers.	[199, 200]
k-partite Multiplex Networks	Many types of links and multiple types of nodes co-exist.	Analysis of interactions between markets and SMA using trading volume and prices together with multiple media channels.	see chapter 15

Table 9.2.: **Some applications require specific network types**. Depending on use-cases, networks consist of homogeneous nodes and links or even of a heterogeneous mix of both.

Network Type	Characteristics	Formula
Static (StatN)	Links and nodes exist at a given point in time t_s. They do not appear or disappear during the defined time range $t = [t_s, ..., t_e]$.	$G_s(t) = G(t) = (V(t), E(t))$
Aggregated (AggN)	Links and nodes can appear during a defined range in time, the network contains all links and nodes which ever existed during the time interval $t = [t_s, ..., t_e]$	$\int_{t_s}^{t_e} G(t)\,\mathrm{d}t$
Temporal (TempN)	Links and nodes exist within a defined time range t, and if $(t_e - t_s) \to 0$ it can be seen as a static network.	$G_t(t) = G(t)\|_{t_s}$

Table 9.3.: **Classification of Network types regarding existence of nodes and links**. A graph G consists of a set of vertices V and a set of edges E. Static networks (StatN) do not change over time. For very short time ranges all networks can be seen as static snapshots or temporal networks (TempN). Aggregation networks (AggN) represent the system during a defined period in time but not at individual times.

9.2.1. Static Link Layer

The static link layer in our research context is given by the Wikipedia page links and all the inter-wiki links between pages in different languages. One can consider all existing links between arbitrary web resources as part of such a layer. In general, all hyperlinks in HTML code, which is finally interpreted by web browsers, are used to traverse the huge content graph. Static means here, that this link structure defines a skeleton on which our analysis method is built. Because these networks do change slowly we can call them quasi-static.

The network of hyperlinks in the world wide web is directed, also the Wiki page network. The network of Wikipedia pages is a very specific case. It has also links to external resources. External pages can link back to Wikipedia pages as well. In this way we can see Wikipedia networks as embedded in other global networks. The Mediawiki software allows us to collect both, in-going and out-going links for wiki pages. For other web resources, we can only find out-going links directly. In order to know all incoming links, one would have to index the entire WWW because each resource could potentially link to any given page. Indexing the entire web is not possible for individuals. For large companies it is an enormous effort to manage all that data, even for giants like Google, Microsoft, and Yahoo!, that provide the largest indexes of today's and historic web pages. They also offer convenient data analysis and search functionality. Offering content and collecting usage statistics at the same time is a very useful technique to support research. Also Wikipedia could benefit from such a tight integration of content delivery with quality and usage analysis, but currently, such a system doesn't exist.

Wiki links are not only created by humans, but also by automatic software based systems, called robots, or simply bots. Because links can be removed in case of deleting a page, we may have to reload the data for specific analysis steps or we have to load and store a snapshot at a given time.

The (quasi)-static link layer is the foundation for our analysis and can be used as a reference for studies on other media channels. Other layers can be compared with this reference layer regarding structural properties. Furthermore, it is possible to bring results into a broader context. E.g., if an unknown relation between the structure of the network and its functionality exists, the structural properties should change if usage patterns change. Different functions can be caused by different structures or visa versa. For manually selected web resources one has to expect a selection bias. With contextual reference data it is possible to identify dependencies and biases to support a reliable interpretation of results.

9.2.2. Content based Networks

Content networks can be formed by explicit links between documents. Such links can be expressed as citation (traditional document) or as Hyperlinks (electronic web documents). In this case the network also represent the static link layer.

The fact, that two documents belong to the same category can also be interpreted as a link between them where the link is not obvious. If two documents are written by the same author they can be considered to be linked as well. Finally, just the similarity of the text or parts of it can be seen as link between otherwise unrelated documents - this is essential for information retrieval methods like full text search. Instead of well defined explicit links we can extract latent links from content. Similarity of documents uses their term-vector representation. Similarity of term-vectors is measured by cosine-similarity (see Muflikhah *et al.* [188]). Advanced text and language analysis is

required for this and multilingual studies are rather complicated.

Wikipedia supports multilingual research with its internal structure. Inter-wiki links and a variety of sub-projects for all relevant languages enable our hybrid approach, which combines explicit and implicit content networks. We start with manually selected pages, than we collect more related pages with a direct link from this seed pages. Wikipedia neighborhood graphs provide a simple abstraction and a language sensitive method without a need for translation.

Similarity Networks and Distance Networks

Similarity networks consist of nodes, typically digital documents and links between similar documents. Links are generated by calculating similarity measures. If two documents are similar to each other they get linked. If similarity s has a high value, the distance d between both documents is short. Otherwise, if similarity is low, the documents are not linked. If $s \in [-1, ..., +1]$ it cannot be used as a distance. In order to have a minimal distance of $d = 0$ in case of marginally similar objects (s=+1) we apply the transformation: $d = (1 - s)/2$.

Two text documents are equal to each other, if the same words appear in the same order in both. Such a similarity measure for texts is also useful for exact identity matching. For practical reasons, texts are not analyzed word by word, rather their term vectors or n-gram vectors are used [201, 75]. A term vector is a special case of an n-gram vector with $n = 1$. Normalization techniques such as stemming and term disambiguation are applied to a document corpus during the preparation phase in order to achieve more robust results.

Calculation of cosine-similarity is often applied to compare short texts or sections in documents within a large corpus for information retrieval [188, 202]. Simply speaking, one takes the term-vector of one document or from a part of it, and looks for documents, which are similar to or nearby the given one in the term space. For practical reasons the original document is not used during analysis. Index data and search terms together with specific logical rules lead to reasonable convenient search results and drive one of the dominant access patterns in the WWW.

In network theory, the term distance is used to define how far one has to go from one node to reach another node in a network. Note, the shortest path is a very popular measure for many applications, such as efficient routing of traffic on roads, rail networks, and communication networks. This distance is related to an existing graph and simply counts the number of path segments between nodes. A link property, such as travel time or distance in real space can optionally be used as a weight. In similarity networks the, nodes are linked to each other only if they are close enough to each other in the *term space*. What close enough means depends on the application.

Event Co-Occurrence and Collaboration Networks

Since metadata are also part of many documents we are able to extract creation time, time of last change, and authors. This allows us to define event co-occurrence and collaboration networks, which are considered to be content related networks as well. Both build bridges between content and social networks. In this way, text documents can be linked indirectly to each other, e.g., by one author which contributed to multiple documents. Two authors can collaborate on one document which defines now a collaboration link between them. Many others of such relations exist. Typically, bipartite networks are used to study relations between objects of different types. One can easily eliminate one type of nodes by replacing all entities of that type by links. This procedure is called bipartite mapping (see also figure 3.3 in section 3.3). This approach allows access to a single subsystem of one node type only and thereby a comparison of subsystem properties. A possible negative impact of dominating structural metrics of one subsystem - caused by mixing two systems of very different properties - is eliminated by this transformation technique.

If one person contributed to a set of documents within a short period of time, this person might be the reason for a high correlation in edit activity (see figure 9.4.b). If more than one person work on a set of documents, we can conclude, that they share a common interest in the topic the documents belong to. We cannot clearly differentiate if this interest is an agreement or if they disagree with the content. Nor can we identify, if people are working towards the same goal or if the high activity is a result of a conflict as reported by Yasseri *et al.* [107].

Beside individual spontaneous events, also series of events can be used to construct networks. In this case we calculate the level of synchronicity for pairs of event series. Events are measured directly or as the result of a transformation of raw continuous time series by applying the link creation function (see Eq. 9.4). Using extracted features, such as strong peaks allows a variation of the intensity and it allows a massive reduction of data which has to be processed. More details about functional networks from event series are provided in section 9.2.3.b.

Semantic-similarity and Semantic-flow Networks

A third content network uses semantic-similarity instead of cosine-similarity. A comprehensive survey on text similarity measures was published by Gomaa and Fahmy [203]. They group several algorithms in three categories: (a) String-based, (b) Corpus-based, and (c) Knowledge-based similarities and demonstrate combinations of those.

Samer and Rada [204] describe *Semantic Relatedness* as *the task of finding and quantifying the strength of the semantic connections that exist between textual units.* Their approach is an unsupervised method for calculating semantic relatedness as semantic profiles for words. Those profiles are extracted *by using salient conceptual features gathered from encyclopedic knowledge* such as Wikipedia. They used two different distance metrics, cosine-similarity and SOCPMI. SOCPMI is a slightly modified version of the *Second Order Co-Occurrence Pointwise Mutual Information* introduced by Islam and Inkpen [205]. They note, that the overall performance of their approach seems to be independent from the selected distance metric. In their interpretation, a word is defined by a set of concepts which share its context and are weighted by their pointwise mutual information.

Masucci *et al.* [191] introduced a method *"to infer the directional information flow between populations whose elements are described by n-dimensional vectors of symbolic attributes."* What they call *'n-dimensional vectors of symbolic attributes'* can be seen as term- or n-gram vectors. This allows us to apply their method to text documents. They use the Jensen-Shannon divergence and the Shannon entropy, which both have a wide manifold of applications in science. Beside a genetic flow network they present a *semantic flow network*, constructed from Wikipedia pages.

Latent Semantic Analysis (LSA) is based on *Singular Value Decomposition* (SVD). The goal is to identify a lower-dimensional representation of the content. LSA provides insights into large document sets, by analyzing the relationships between the words within the documents. Therefore, a set of relevant concepts is extracted from the corpus. According to Ryza *et al.* [206] such a *concept consists of three attributes: a level of affinity for each document in the corpus, a level of affinity for each term in the corpus, and an importance score reflecting how useful the concept is in describing variance in the data set.*

Blei *et al.* [207] introduced an approach, called *Latent Dirichlet Allocation* (LDA), which is a generative probabilistic model for collections of discrete data such as text corpora. According to them, *LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.* The resulting topic model is an $N_{topic} \times N_{doc}$ matrix in which the elements describe the level of participation of a document in each topic.

Amancio *et al.* [208] proposed methods, which combine semantic and structural properties of texts. The topology allows capturing stylistic features concerning authorship and text quality, they state.

Recently, "*The Open-Vocabulary Approach"* was proposed by Schwartz *et al.* [177]. Especially in large-scale social media systems, such as Facebook, or Google+, it seems to be relevant to recognize the relation between language usage and properties of persons, which communicate with each other. They found, that a language based approach allows to distinguish people by personality, gender, and age.

As a practical example we investigate a network reconstructed from short text documents in a knowledge base. It was created using the cosine-similarity[2]. Alternatively also the Jensen-Shannon distance (JSD)[3] was evaluated for different n-gram sizes in the range from one to eight. One has to note that both measures behave complementary since one is a distance and the other is a similarity measure. Finally, both lead to comparable results. The results for $n = 3$ are shown in figure 9.5. Such a question and answer system (QnA system) contains documents of two types, questions and answers. Questions can be connected to an answer by a directed explicit link. This way, we have a bipartite network with two node types. Because some questions are not answered yet, they are isolated nodes. Valid facts can be stored as answers in such a system, even if no question is related to this fact. Also those fact-nodes wouldn't be linked to others. A second link type is based on semantic similarity, using JSD (see Eq. 2 in [191]). Figure 9.5.a shows two strongly connected clusters. It is not surprising, that the two clusters represent the two node types, questions and answers. To emphasize the inherent semantic structure, all explicit links were removed in figure 9.5.b. This highlights clusters of semantically related items, independent of an explicitly created link structure. A topic model is not required for this technique.

Explicit links and similarity links define two link layers in figures 9.5.c. This multi-layer approach contributes context information to existing items in a document collection. Here, we are able to identify related questions, because they are linked to the same or a similar answer, or because they have a high semantic similarity. Studying the evolution of such structures as a function of time, together with access-rate and edit-activity time-series (see chapters 13 and 15) can be used as an experimental setup to study self-organized knowledge-formation. This process exists because persons contribute without any obligation or task assignment. If additional rules are applied, or automatic tools such as de-duplication or disambiguation procedures operate on the data besides persons then a change in the structure can be expected as already shown in the example of the Swedish Wikipedia in figure 8.2.b (see exception C).

Information flow analysis requires multiple inputs. The content can be represented by documents or messages. Users are the source for information and also information consumers. This means that the set of users, which is for itself a network, is linked to the content by at least two different link types. In our data set we use access-rate time series to express the interest of Wikipedia users in content consumption and the edit-event time series to track

[2]The classic cosine-similarity measure is defined, e.g., in [**?**] in Eq. (6).
[3]A definition of JSD is given in Eq. 2 in [191].

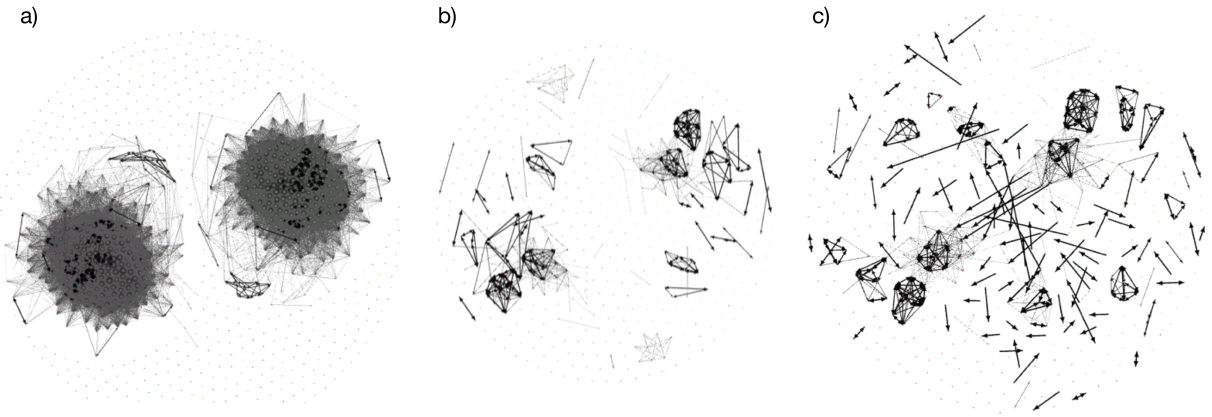

Figure 9.5.: **Semantic-similarity Network in a QnA system**. QnA systems contain items of two categories, questions and answers. Beside the explicit links between an answer given to a question (black), also implicit similarity links exist (gray). Such relations are identified by a semantic similarity measure (see Eq. 2 in [191]). Panel (a) shows all links. Links with a high semantic similarity ($s > 0.75$) excluding explicit links are visible in panel (b) and (c) shows a combination of the semantic network of undirected links (gray) and the structural links between questions and answers (black).

user contributions. Based on such data, it is possible to measure the total activity. A more detailed analysis is possible if content can be categorized. There are many different approaches for categorization and classification of Wikipedia content. Those approaches can be generalized to content from non digital sources such as books, print media, or spoken text from TV and radio stations.

Wikipedia provides an implicit content categorization. Specifically, each page can be linked to one or multiple category pages. Grouping articles by language provides an additional classification scheme. Even if no explicit content categorization is available, it is possible to derive a topic model from any text corpus.

9.2.3. Functional Networks

Functional networks represent hidden relations between network nodes, which are not directly measurable. The idea behind functional network reconstruction is, that if a property of two elements changes synchronously (or with a given time delay) one can assume a common influencing factor with impact on both.

Therefore, we apply an operation $\mathcal{F}_{\mathrm{LC}}$ to a tuple (or simply a pair) of time series to calculate a value, which represents the relation between the objects, from which the time series were obtained (see Eq. 9.4). This means, if access activity to two Wikipedia pages is correlated, we assume also a correlation in interest in both pages. Then we define a link between both pages to express common interest in both at the same time. This link is a temporary link, and depends on: (a) length of sliding window, (b) filter procedure for time series, and (c) filter procedure for link strength (see chapter 10). Some correlation methods provide only a link strength, but no orientation (see table 9.3). Cyclic patterns and strong peaks can have a dominating influence on correlation results. Therefore we study the impact of strong peaks and additional quality metrics in more detail in the next chapter.

Beside the node degree, several centrality measures and other network analysis procedures are applied (see: calculation of network profiles in section 3.5.2). In this way, functional networks provide data which describe dynamic not directly accessible aspects of complex systems. Figure 9.3 in the beginning of this chapter compares the static view of local neighborhood networks for one Wikipedia page with the corresponding functional network, based on (a) direct Wikipedia page links, and (b) functional links, calculated from user access-rate time series. Centrality and PageRank were calculated for each node. One can clearly see that the most relevant (most connected, most central) nodes are different in both representations of exactly the same Wikipedia articles.

9.2.3.a. Correlation Networks from Continuous Time Series

We consider functional networks as useful objects for studying the interaction properties between social networks of users and content networks. The approach can be generalized to arbitrary time series, such as climate data, financial data, or machine and sensor data. For each pair of nodes we calculate a link strength using one algorithm from table 9.4 to find a representation of the system as listed in table 9.3.

The value of the cross-correlation coefficient defines the strength of the functional link between the two considered nodes. Repeating the procedure for each pair of nodes yields a functional network representation of the user access-rate cross-correlations. The calculations are also performed for temporal slices of width Δt beginning at t_0, so that

time-dependent (dynamically evolving) functional networks are obtained. Specifically, we calculate for each pair (i, j) of nodes:

$$\mathrm{CC}_a^{(i,j)}(t_0,\tau) = \frac{1}{\sigma_i \sigma_j}\left[\frac{1}{\Delta t}\sum_{t=t_0}^{t_0+\Delta t-1} a_i(t)a_j(t+\tau) - \left(\frac{1}{\Delta t}\sum_{t=t_0}^{t_0+\Delta t-1} a_i(t)\right)\left(\frac{1}{\Delta t}\sum_{t=t_0}^{t_0+\Delta t-1} a_j(t+\tau)\right)\right], \tag{9.8}$$

$$\text{where} \quad \sigma_i = \sqrt{\frac{1}{\Delta t}\sum_{t=t_0}^{t_0+\Delta t-1} a_i^2(t) - \left(\frac{1}{\Delta t}\sum_{t=t_0}^{t_0+\Delta t-1} a_i(t)\right)^2} \tag{9.9}$$

and σ_j accordingly.

Similarly, editorial activity can be studied. However, since edit events are rather sparse, event synchronization coefficients should replace the Pearson cross-correlation coefficients (see next section). As weak correlations between user access-rate time series also occur randomly (because of limited statistics), we re-normalized computed link strengths (see *Palus et al.* [28]). The calculation of $\mathrm{CC}_a^{(i,j)}(t_0,\tau)$ was repeated $k = 10$ times for randomly shuffled time series $a_i(t)$ and $a_j(t)$ to determine normalization factors $\langle \mathrm{CC}_{sa}^{(i,j)}(t_0,\tau)\rangle_k$ for each pair of nodes (i, j).

Next, we calculated the **adjusted link strength** as:

$$l_{\mathrm{adj}}^{(i,j)}(t_0) = \mathrm{CC}_a^{(i,j)}(t_0,0)/\langle \mathrm{CC}_{sa}^{(i,j)}(t_0,0)\rangle_k. \tag{9.10}$$

The corresponding functional network expresses how different the links between two nodes are from random links calculated for random time series.

We tested a more robust method, called *'normalized link strength'*, which was also used by Berezin *et al.* [29] based on time-delay variance.

We calculate the **normalized link strength** as:

$$l_{\mathrm{norm}}^{(i,j)}(t_0) = \frac{max_\tau(\mathrm{CC}_a^{(i,j)}(t_0)) - \langle \mathrm{CC}_a^{(i,j)}(t_0,\tau)\rangle_\tau}{\sigma_\tau(\mathrm{CC}_a^{(i,j)}(t_0))}. \tag{9.11}$$

Instead of calculation of the time delay variance[4] ([29]), we introduce and evaluate a quality metric to measures the impact of strong narrow peaks. We replace the maximum value of the cross correlation function $\mathrm{CC}_a^{(i,j)}(t_0,\tau)$ with $\langle \mathrm{CC}_a^{(i,j)}(t_0,\tau)\rangle$ to get $\mathrm{CC}'(t_0,\tau)$ and calculate $l^{'}(t_0)$ using Eq. (9.11) from $\mathrm{CC}'(t_0,\tau)$. This allows a separation of such correlation functions with a single sharp peak (potentially caused by an artifact, or due to a very high similarity) from such functions without any peak or high fluctuations. Depending on the application, the sharp peak at a given τ can be the indicator of a strong link between nodes.

Next, we calculated the **transformed link strength** as:

$$l_{\mathrm{trans}}^{(i,j)}(t_0) = l_{\mathrm{norm}}^{(i,j)}(t_0)/l^{'}(t_0). \tag{9.12}$$

Correlation networks with link strengths calculated from Pearson correlation are un-directed networks. Even if one time series precedes another one significantly, this information is not represented in the link definition, which is just a scalar value. If the cross-correlation function with time delay $F_{\mathrm{xy}}(\tau)$ (Eq. 5.9) is used instead, one can find the characteristic delay between both processes by looking for the maximum in the correlation function. Based on this delay it is possible to extract information about the orientation or timely order, which is than translated into a direction. Berezin *et al.* [29] interpreted the variance of all τ values as a criteria to separate real links from random links.

Time delay values can be used to define a distance network, which uses a delay as link strength value to express how far away nodes are from each other or how expensive it is to traverse the link between both. This concept is also applicable to a network of states, where directed links between nodes (each representing an individual state) exist only if a state transition between them is possible. In this case the link strength represents the required activation energy to initiate the transition process. Instead of the delay value, an energy barrier is now defining the link strength. Because an energy barrier may not be measurable directly, it might be possible to derive it from measured delays in direct accessible variables.

Beside calculation of cross correlation coefficients in traditional computers, also a very efficient implementation in hardware is possible. Such a system can provide real time results, opposed to our current approach, which relies entirely on pre-aggregated time series data. David Tam [**?**] proofed, that a computational function performed by a time-delayed neural network which implements Hebbian associative learning-rules computes the equivalent of the cross-correlation function of time series. He shows the relation between the correlation coefficients and the trained connection-weights.

[4] Time delay variance (TDV) as used by Berezin *et al.* [29] is a property of the distribution of delay values τ for which a maximum correlation is determined in the cross correlation function with time delay $F_{\mathrm{xy}}(\tau)$ (Eq. 5.9).

Name	Equation	Description
cross-correlation for delay τ	Eq. (9.8)	The cross correlation coefficient for a time delay τ is used as link strength. No shuffling is included.
adjusted link strength	Eq. (9.10)	Compares the cross correlation value (no delay) with the average correlation for k shuffled samples.
normalized link strength	Eq. (9.11)	Compares the maximum value of the cross correlation function F_{CC} with its average (normalized by standard deviation). No shuffling is included.
transformed link strength	Eq. (9.12)	Compare the normalized cross correlation link with the normalized link of a transformed cross-correlation function, e.g., the transformation function F_{TRANS} replaces the maximum value with the average to emphasize the influence of a single sharp peak in the cross correlation function.

Table 9.4.: **Variations of cross-correlation functions.** For link creation and also as significance tests we use variations of cross correlation functions.

9.2.3.b. Event-Synchronization Networks

Rather than just a correlation strength Q (see Eq. 5.14a), the event synchronization method also provides information about the direction q (see Eq. 5.14b), or the order in time. In this way, it is easy to see if an external process leads to significant features in one time series early, before other time series are effected. One has to be careful here, this information does not allow a conclusion about causation or causal dependencies between the elements, from which data were obtained.

Event time series are used by Malik *et al.* [137] (see page 975 figure 3). We adopted their approach and create a functional network layer to represent user contributions to Wikipedia in form of editorial activity.

An alternative metric to measure a distance between two spike trains (which are sparse event series) is presented by Houghton and Kreuz [**?**] in their paper *'On the efficient calculation of van Rossum distances'*. Like in our case, many applications require a matrix of distances between all the spike trains in a set. Furthermore, the calculation of a multi-neuron distance between two populations of spike trains is a rather expensive approach. They present an algorithm to render these calculation less computationally expensive, making the complexity linear in the number of spikes rather than quadratic.

9.2.3.c. Dependency Networks and the Context Cohesive Force

Opgen-Rhein and Strimmer published a method to generate a causation network for high-dimensional plant gene expression data [279]. They describe partial correlation as the correlation that remains after regressing the effect of other variables away. Beside the correlation they also take the variance of the signals into account and define a link direction based on the most exogeneous variable. Such a directed link only exists, if the logarithm of the two variances is significantly different from zero, which means the variances are different and allow the definition of a direction. Finally, they create a directed acyclic graph, which is a subgraph of the undirected correlation network.

Another method of reconstruction of dependency networks uses triples of nodes to calculate a link strength between two nodes in the presence of a third. We define the *Context Cohesive Force* (CCF), which is a generalization of the *Index Cohesive Force* (ICF). ICF, introduced by Kenett *et al.* [210], was used for network reconstruction from financial time series. Here, we use it as new approach to study interlinked social communication networks and social content networks together, in the presence of other systems, in which both are embedded in.

Consider two systems A and B to be bi-directionally coupled. They interact with each other and consist of elements e_A and e_B. For individual elements we select one property, in case of stock market analysis, e.g., the log of daily returns (or log of the absolute daily price differences) and the hourly access activity in case of Wikipedia pages. As financial data are available on a daily base, we use also daily access-rate data to be consistent. Even if intra-day trading data would be available, it would not contribute much more useful information, because we only have the hourly Wikipedia access-rate time series. We have to choose an appropriate length for a sliding window in order to generate time-dependent results. In particular, time resolution and the length of interval overlap are specific properties of the analysis scope.

Our goal is to measure the influence of system B on internal correlations (intra-correlation) of system A. Therefore, the intra-correlations $\text{CC}_a^{(i,j)}(t_0)$ are calculated for all pairs of nodes (i,j) in system A and all pairs from different systems $i \in A$,$j \in B$ using Eq. 9.4. Because the cross-correlation function is symmetric, we calculate the correlation strength $\text{CC}_a^{(i,j)}(t_0)$ only for time series pairs $a_{i,j}$ with $i > j$ if $i, j \in A$.

According to [23, 210] the partial correlation $\rho(i, j|m)$ between pages i and j in the context of a mediation page

m (this can also be a category page, which represents system B or a totally different mediation time series obtained from a related system) is calculated as:

$$\rho(i,j|m) = \frac{C(i,j) - C(i,m)C(j,m)}{\sqrt{(1-C^2(i,m))(1-C^2(j,m)))}} \tag{9.13}$$

where $C(i,j)$ is $\mathrm{CC}_a^{(i,j)}(t_0)$ for simplification. Now, we can interpret $\rho(i,j|m)$ as the residual correlation between the pages i and j which does not include the correlation between both and the page m.

Previously, the *Index Cohesive Force* (ICF) for stock prices grouped by stock index was defined by Kenett *et al.* as the ratio of raw and residual correlations [23]. In order to generalize this idea, we use contextual neighborhood networks around Wikipedia pages. This allows us to apply the method to semantic concepts grouped by topics or semantic categories.

We define the *Context Cohesive Force* (CCF) as the ratio of the average pair correlation and the average partial correlations within this neighborhood during a time interval t:

$$\mathrm{CCF}(t_0) = \frac{< C(i,j)(t_0) >}{< \rho(i,j|m)(t_0) >} \tag{9.14}$$

For CCF > 1 the average internal correlations are higher than the average partial correlations. If both are equal we will find CCF $= 1$ and CCF < 1 is a result of stronger partial correlation. This is interpreted as an indicator of an external influence caused by the entity from which the time series m is obtained. This external entity represents the context. The context cohesive force quantifies the influence of the context which can be either the neighborhood in which the system is embedded in, or an ensemble of which the system is a part of.

In comparison to the observed effect an index has on stock correlations we analyze the correlations between the Wikipedia page which represents the stocks index and the page's access-rate time series. Kenett *et al.* found, that larger changes of the index results in higher stock correlations. Based on those findings we assume: If movements in stock markets cause an increase of interest in financial topics in Wikipedia over time one would measure (a) an increase of intra-wiki correlations between pages regarding a given market, and (b) an increase in partial correlations between the stock market data and the Wikipedia access-rate data for related pages over time. We discuss our preliminary results in chapter 15.

9.2.3.d. Correlation Networks from Non-stationary Time Series

An increasing demand for alternative approaches, which can handle non-stationary time series as well, can be explained easily. Since more and more data become available, but data are collected in sometimes unstable or non-stationary environments, one cannot apply Pearson correlation, because the results are not reliable under such conditions. The influence of extreme events and outliers, which are characteristic properties of real-world data sets, especially from socials media systems, has to be eliminated or addressed in a specific way. Our first approach is to focus on extreme events and outliers only. The time series are transformed into event-time time series by event detection algorithms. The event-synchronization method can then be applied.

Application of random matrix theory (RMT) is a second alternative. According to Podobnik *et al.* [**?**] RMT is used to analyze time-lag cross correlations in complex systems. They address the question whether these cross-correlations exhibit power-law scale-invariant properties. Therefore they applied time-lag RMT (TLRMT) to time series from finance, physiology, and genomics. They found long-range correlations in the finance data set by comparing the calculated eigenvalues with expected eigenvalues from random matrices. In this way, they could demonstrate different properties for return and volatility[5]. Podobnik and Stanley [255] introduced a method, called detrended cross-correlation analysis (DCCA), which is a generalization of detrended fluctuation analysis (DFA, see section 5.3.2). DCCA uses detrended covariance. Investigation of power-law cross correlations between pairs of different non-stationary time series is the purpose of this method. The DCCA coefficient was introduced by Zebende [**?**]. Kristoufek *et al.* [**?**] also conclude, that the DCCA coefficient can be used to measure correlation between non-stationary time series.

Because the DCCA coefficient can be used for non-stationary series it allows analysis of raw time series, even if they contain trends or extreme events. Most importantly, the DCCA coefficient provides information about correlations at different scales. The Pearson correlation coefficient can be calculated for time series of different length and at different times, but this should not be confused with analysis on different time scales.

The DCCA long-range cross-coefficient $\rho_{ij}(s)$ measures the correlation between two series on multiple scales. In order to analyze the scaling behavior the long-range correlation exponent λ is calculated by linear regression in the log-log representation of $\rho_{ij}(s)$. Gang-Jin *et al.* [68] applied this method to data from foreign exchange market. They created a series of Minimum Spanning trees (MST) from financial time series (log-return of daily FX rates of 44 major currencies in the period of 2007 to 2012). Instead of using a time-resolved analysis procedure, they study the properties on different time scales and identified different topological properties in functional networks, created for specific time scales.

[5] See section 5.2.5 for more details about preparation of financial time series.

9.3. From Time Series to Dynamic System Properties

Each network reconstruction step provides an adjacency matrix, one for each time interval. Macroscopic descriptions of a system are based on structural properties, such as traditional network measures, or one of the many new algorithms, which were developed recently [**?** 82, 67]. Such new algorithms allow us to handle multiple node and edge properties by using vectors. We are not longer limited to scalar values for each node and edge.

9.3.1. Time-dependent Multivariate Network Metrics

Temporal networks represent the system at a specific time, usually in the middle of the interval, defined by the length of the series.

The PageRank algorithm assigns a scalar value to each node of a static graph. Static means in this case, that during the iterative calculation of the final values by using the power iteration method (see [**?**] Eq. 5) the link structure and thus the transition probabilities are not changed. Weighted transition and random jump probabilities can take the freshness and activity of webpages ito account. Berberich *et al.* [**?**] introduce the T-Rank approach, as an extension of the widely accepted PageRank. This new approach is an extension of the Markov chain model and allows a time aware authority score for nodes to express their relevance compared to others. The problem of this method is the lag of required metadata to calculate freshness and activity.

During this work we have found that exactly this kind of information is available on Wikipedia. Initially this seemed to be really easy, but the technical requirements are huge. One has to handle the full Wikipedia network (more than 10.000.000 pages) as well as multiple slices of the access activity and edit activity. Each slice represents a time range and also the time resolution. Access-rates represent the attraction of pages and editorial activity expresses the freshness of articles. Using the method of local neighborhood graphs, which was developed in this work, it is possible to present the temporal relevance of a particular node including the structural properties.

Our network reconstruction approach is based on node similarity (especially on activity similarity) - independent of the underlying static network. The T-Rank does not take the activity similarity into account, but reflects the existing network structure. Both methods appear as complementary approaches for time-dependent analysis of complex networks.

Since a link can now be either a scalar value or a vector - in case of multiple layers each layer contributes one dimension to a link vector - one has to choose appropriate network measures. Here we call those *'multidimensional topological measures'*. Especially in case of multi-layer Networks (MN) it is essential to apply modified variants of established algorithms, such as the Multiplex PageRank introduced by Arda *et al.* [82]. Weighted network measures for linear and nonlinear correlation networks were developed and applied by Donges *et al.* [25, 26, 27], in order to take the influences of multiple sensitive measures into account. Another example for multi-layer network analysis is presented by Cui *et al.* [**?**]. They could not directly use the PageRank algorithm to predict importance of authors or papers. Because the number of references of a paper is not the same as the number of out links of a web page used in the original method, they modify the PageRank algorithm. In their case, it is not the nature of a link property, which requires such a modification, but the entire model they use to study the probability of link creation.

9.4. Discussion

Essentially one can differentiate between topological node properties (such as degree k, page rank r_p, or centrality c_N) and global properties which represent the entire system (average degree, global clustering). How those properties depend on each other and on processes on top of the networks is an open question, which is related to a huge class of unsolved problems.

A scalable network analysis framework, which integrates data from several sources in a robust and repeatable reconstruction procedure, together with efficient simulation techniques are important factors. Finally, if we want to learn more about the impact of so called *influencer nodes*, we must learn how the influence is represented and what phenomenon causes the dependencies, which can be observed but often not yet explained.

Machine learning techniques are recently very successful. They provide, e.g., probabilities for certain events and thus they can support decision processes. However, those algorithms have no explanatory power. They do not identify the driving forces behind the processes.

Analysis of the time evolution of the structure of complex systems might be an appropriate tool.

Many different network types exist. They can be analyzed by a variety of network measures. Many studies show one very specific network, which stands for a particular property or aspect of the system they study. Complementary or even overlapping alternative representations of aspects within the same system have also to be analyzed. Therefore, different network types have to be combined. In our case we use reconstruction methods, appropriate for the available data.

First, we have seen rather static content based networks. Time-dependent networks describe the evolution of the system and the moving parts. Time-resolved analysis is based on creation of snapshots, which represent the

system at a given point in time. Many of such snapshots provide the data for time-dependent structural measures.

The way how time-dependent functional networks are created is somehow related to the critizised procedure of time series clustering. Keogh and Lin [**?**] write: *"Given the recent explosion of interest in streaming data and online algorithms, clustering of time-series subsequences, extracted via a sliding window, has received much attention".* They claim, that *"clustering of time-series subsequences is meaningless"* and *"clusters extracted from these time series are forced to obey a certain constraint that is pathologically unlikely to be satisfied by any dataset, and because of this, the clusters extracted by any clustering algorithm are essentially random".* The new approach they propose is based on the concept of time-series motifs.

Since network reconstruction is not a clustering approach, rather than an individual part of a more complex procedure, it allows clustering based on the obtained network data in a second step. Our own simple experiment - it was conducted before we were aware of the work from Keogh and Lin - showed, that the clusters found in the networks were characterized by specific motifs in the time series such as peaks and peak sequences. We did not apply a sliding window technique but we were able to differentiate and to isolate individual phenomena which appeared at different points in time in long time series based on typical patterns, or motifs.

With regard to Keogh and Lin it seems to be important to study the impact of their findings on presented network construction and reconstruction procedures in the future. Of special importance is also, if different time series based methods show different properties, dependent on the sliding window technique. Regarding the link strength distributions we can already conclude that transformations like filtering, detrending, and the logarithm function change the intermediate results - the link strength distribution of temporal networks. In the future it will be important, to study the relation between those transformations and the final topological properties.

10. Identification of Significant Correlation Links

"Everything must be made as simple as possible. But not simpler."

(Albert Einstein)

In many cases, links between network nodes are well defined and can be measured or observed directly. For such obvious links it is straightforward to analyze the topological properties of the corresponding networks. However, if links are not directly observable and need to be reconstructed from dynamical signals, the underlying network structure could affect these signals and thus influence the reconstruction and partially invalidate topological properties based on it.

Correlation networks and dependency networks have been used recently to describe emergence of extreme events in the earth's climate system, such as the El Niñjo [29**?**] and the interdependence of components within the global economy [286]. Initially, such networks are complete graphs width a weighted adjacency matrix. Before traditional network analysis algorithms can be applied to such weighted networks, one has to identify the significant and therefore relevant links. An alternative is to apply weighted network measures as introduced by Wiedermann *et al.* [121].

10.1. Introduction

How stable are the results in the presence of external influences and intrinsic changes? And what is the impact of different link creation and filter methods? It is important to verify, if the applied methods have a direct influence on the selected topology measure. Furthermore, it is of a high relevance to know how stable the calculated link strength distributions are over time.

In this chapter we describe a new approach to identify relevant links, based on two quality metrics in addition to the well known normalized link strength calculation procedure. One additional link property is related to the degree of randomness and the second is related to the shape of the calculated correlation function.

Furthermore, we investigate a class of algorithms, which allows filtering of fully connected networks. The goal hereby is, to obtain the most meaningful information. A very simple approach is based on a static threshold. Creating the Minimum Spanning Tree (MST) is another widely used method, especially for networks reconstructed using distance measures. A common algorithm to calculate the MST is the Kruskal algorithm [**?**]. Many other authors applied this algorithm to extract informative sub-graphs from reconstructed complete networks. Because the MST is just a tree, it is not possible nor useful to apply algorithms like clustering or motif statistics to the resulting sub-graph. The Planar Maximally Filtered Graph (PMFG) was introduced by [69] to overcome this limitation. A PMFG is another sub-graph which retains more structural information than the MST. Both types

of sub-graphs can be seen as structural link filters and both are parameter-less. Because of this good properties they were used so often recently.

One can also invert this concept by using the largest link values, e.g., such obtained from correlation and similarity analysis. Applying Kruskal's algorithm (see [**?**] and [**?**]) leads to the Maximum Spanning Tree in this case.

The assumption that only the strongest links are relevant is not true in general. Especially in case of correlation properties it turned out that a relative measure provides a more realistic view. Beside this, also weak links can have a well pronounced sharp peak in the correlation function $F_{xy}(\tau)$. In some cases, such as in social media analysis, it is not useful to separate only the strongest links, because also weak links[1] can stand for important characteristics of the system. In this case one might merge the results from two or more filter approaches appropriately. How those links can be separated from noise, which is represented by the huge amount of weak links with no specific peak is shown in this chapter.

10.2. Critique on Existing Approaches

Defining artificial link strengths between not obviously linked elements is not new. Modifications to the similarity measure are proposed and validated, e.g., Tsonis and Swanson [**?**], constructed networks from measured surface temperature for El Niño and for La Niña years in their paper from 2008. They investigated topological properties of that correlation networks and found, that in the presence of El Niño, the network has significantly fewer links, lower clustering coefficient, and lower characteristic path length. This highlights a difference in both networks: the El Niño network is less communicative and less stable than the El Niña network. They write: *"A pair is considered as connected if the absolute value of their cross correlation* $r >= 0.5$*. This criterion is based on parametric and nonparametric significance tests. According to the t test with* $N = 60$*, a value of* $r = 0.5$ *is statistically significant above the* 99% *level."* Furthermore, in a side note they state: *"The choice of* $r = 0.5$*, while it guarantees statistical significance, is somewhat arbitrary. We find that while other values might affect the connectivity structure of the network, the effect of different correlation thresholds is negligible on the conclusions reached in this"* (in their) study [**?**].

A static link strength threshold might easily be defined for one single network based on the shape of the probability density function of the link strength distribution or based on a predefined confidence level. According to Berezin *et al.* [29] the goal of filtering the link strength data is to separate the set L_P (all links caused by real physical dependence) from L_N (just random correlations or noise). Both types of links are part of the set $L = L_N \cup L_P$. Berezin *et al.* worked with a confidence level of 98%. Depending on the distribution of the link strength values a resulting filter threshold is calculated per time interval and per region, in order to have 2% of links with the highest calculated link strength in set L_P. In order to define a reliable threshold, they plot two quantities: (a) link time delay variation $STD(T_{l,r})$ and average link strength $\overline{W}(l,r)$, and find a rather stable shape in the 2D histogram which illustrates a crossover between two regimes. They thus expand the one dimensional problem to a second dimension. This way they can take a second indicator into account. Time delay variation is used beside the average link strength, to classify links or candidates for significant links. They report that the qualitative behavior is consistent across different regions on the globe, but not constant everywhere. The threshold also varies with time. They base their conclusion and threshold selection on an increased sensitivity found around the crossover region.

In case of growing social media systems, such as Wikipedia, it is not possible to apply either of the two methods directly. A climate network consists of a constant number of nodes. All nodes have a well defined position and thus also fixed distances from each other. Available climate data time series are longer than 20.000 data points (daily values for 58 years). The number of grid points and the number of stations on which weather data is collected is constant, this allows to say, the system size is constant, although some conditions may change over time, such as the quality of devices and the density of measurement points. Such variations have to be handled during the data preparation phase. We summarize: the research focus is on climate change, but the model system has rather stable boundary conditions. This is not the case for the Wikipedia page networks.

Even if the number of relevant objects is constant, such as a group of selected cities or pages about companies and products (in the context of economical analysis, e.g., for globally interwoven financial markets), the number of pages in their neighborhood, and thus the system size are not stable. The number of users, which influence the system as well is growing, and user activity shows clear seasonal patterns. Therefore it is important to apply more robust methods. They must be robust regarding all those variable boundary conditions and normalization procedures, which stabilize the measured data before the time-dependent analysis is applied. Using time delay variance (see Berezin *et al.* [29]) is an example for this.

In this work I reconstructed correlation networks based on a new concept. Instead of only the Pearson correlation or the normalized link strength, obtained from a correlation function, I use two additional measures as quality metrics. Furthermore I developed a procedure to identify the local neighborhoods to implement a contextual

[1] Weak links are relevant if the correlation which, defines the weight or link strength is significant.

detrending within the semantic neighborhood of the content network before correlation analysis is applied.

Especially if a structural analysis is applied to the resulting time-dependent networks, a fixed filter parameter, based on an absolute link strength is not a good solution. All three methods have one negative property in common: they completely ignore the weak ties. Systematically ignoring the weak ties leads to information loss.

A fixed filter threshold filter should not be used for time-dependent studies without prior normalization of the link strength values. Normalizing the values of all time ranges to a global maximum (highest value ever seen in a particular study) or even the local maximum (which is specific for each time range) can help to solve this problem. In this case it is important to specify the normalization procedure, because it has a major impact on final results. Initially, we use the same approach as described by Berezin *et al.* [29] in section: *The network construction method.* Such a normalization of the correlation function is used to obtain a rather stable correlation measure, which is not depending on any particular time delay τ. Such transformations have an obvious impact on the link strength distribution. In our case (not shown) a long tail becomes visible and the difference between the correlation values calculated for randomized data is more significant, especially since the results from randomized data are symmetric and don't have this long tail.

In case of MST and PMFG the sub-graphs include those nodes where *"edges represent the most relevant association correlations"*. Depending on the process, which is analyzed, it is not always possible to use this rather simple assumption. According to Onnela *et al.* [278] it is important to take the different role of strong and weak links (they call them weak ties) into account. They found that the size of the giant component varies differently, depending on the type of links which are removed from the network first. If the weak links are removed first then the size of the largest component decays faster than in the case there the strong links are removed first. This work is based on the so called *"land mark paper"* by Mark Granovetter [**?**] titled *'The strength of weak ties'* published in 1973. According to Granovetter, one can find links with different roles in social networks. The weak ties connect clusters formed by highly connected nodes, while the cluster-internal links are strong ties. This is also supported by figure 10.13 later in this chapter. A wrongly chosen filter could not identify such weak ties. This has many implications. The nodes connected via strong ties are in the same cluster because of their similarity. One can expect a very specific type of interaction between them. Shared interest in the same topic leads to discussions about a particular topic, but probably less communication about "off topic" content happens within this group or cluster. New inspiration, new insights or just the missing piece of a puzzle can often not be found within a group of similar persons or in documents about the same topic. Connectivity to other clusters is important now. Such connecting links are rather weak and could easily be missed.

This can also be understood by considering social interaction as an example. A close contact to colleagues within the same working group or to members of a family can be maintained with less effort (here we don't address the quality of relations). Staying in touch with relatives in a city far away or keeping the good relation with friends from school over years seems to be impossible for many people, because of the nature of such weak ties. The nodes in a different cluster, which is formed by nodes less similar to the initial node, require more attention and more effort in order to stay connected with them. On the other hand, such weak links are *"the universal key to the stability of networks and complex systems"*; this is also the title of a book by Peter Csermely [**?**] published in 2009. This gives us a strong motivation to also identify weak but significant links.

10.3. The Percolation Threshold

Percolation on a lattice or on networks is characterized by a critical occupation probability p_c, the so called percolation threshold. On lattices one can distinguish bond percolation and site percolation (see figure 10.1.a). Site percolation means, that randomly all sites of the lattice are occupied with probability p. The occupation probability at the time when a connection between two boundaries of the underlying geometry appears, gives us the percolation threshold, which is in this case also the density of the elements on the lattice. A second model is called bond percolation. Instead of occupying the sites, we draw lines between them. The density of lines in the lattice, where a connection between the boundaries emerges, defines the percolation threshold. At the critical point, there exist a continuous path between two boundaries but also several unconnected ("finite") clusters. There is not just the one connecting ("infinite") cluster. One has to repeat simulation experiments multiple times in order to find p_c with a reasonable accuracy.

Percolation on a network must be analyzed in a different way. Because spatial embedding[2] is not available in arbitrary networks, one studies the size of the second largest cluster in the network. How is the size of this cluster N_s changing as a function of the overall network size N? In general, one observes that N_s becomes constant as N increases, but not close to the percolation threshold. As long as $p < p_c$ all clusters have a maximal size (given by the correlation length) and if $p > p_c$ the largest cluster contains most nodes. In case of $p = p_c$ the second largest cluster also grows as the overall network is growing (see also [**?**]).

[2]Spatial embedding means, that the nodes of a network have coordinates in real space, e.g., expressed as latitude, longitude, and height. More general, one can say: "A spatial embedding of a graph G (or spatial graph G) is a set of points in $\mathbb{R}^3$ (corresponding to the vertices of G) and a set of smooth arcs (corresponding to the edges of G) that join appropriate pairs of vertices and intersect only at vertices (see [**?**]).

In some cases, if spatial embedding of a network is possible, a hybrid concept as illustrated in figure 10.1.b can be applied. First, one transforms the problem to a bond-percolation problem on an unregular lattice. Depending on the orientation one is interested in, two outer boundaries are defined by two parallel lines (u,l) through the two outer most nodes (A,B). This allows analysis of percolation as a function of a chosen direction ($\overrightarrow{a}$), which is orthogonal to the previously defined boundaries.

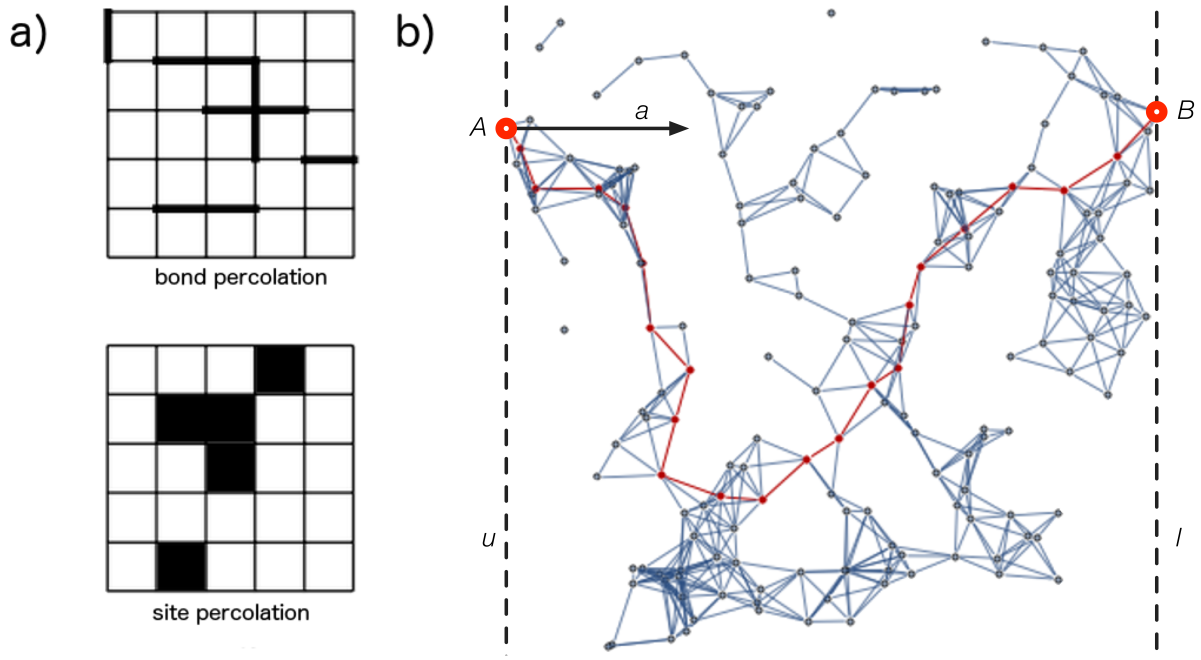

Figure 10.1.: **Percolation on lattices and networks**. Different percolation types exit on lattices and networks. (a) shows bond percolation on a lattice in the upper tile, and site percolation in the lower tile (taken from [**?**]), (b) illustrates an adoption of bond percolation to a spatially embedded network (image is based on: [**?**]). The image illustrates the percolation orientation $\overrightarrow{a}$ which is orthogonal to the boundaries, defined by two parallel lines u and l through the two outer most nodes A and B. In case of percolation a continuous path between the boundaries exist. Such a path is highlighted in red.

In general, an analytical value of p_c is not known for most geometries, especially because p_c depends on the lattice structure and the topological properties of the network. Newman and Ziff proposed a Monte Carlo based method to calculate p_c [**?**]. Karrer *et al.* [**?**] calculate p_c as the value of p where the size of the second largest cluster (component) of a network reaches its maximum.

Exact results for percolation on a 2D lattice were published by Domany *et al.* in the article *Directed Percolation in Two Dimensions: Numerical Analysis and an Exact Solution* [**?**]. More exact solutions for percolation thresholds in networks were studied by Cohen *et al.* [195] and Buldyrev *et al.* [57]. The percolation threshold can be calculated exactly for random graphs. In case of graphs with a fixed degree k (random regular graphs) the percolation threshold is $p_c = 1/k$. According to Cohen *et al.* [195] one can find $p_c = 1/ < k >$ in Erdös–Rényi (ER) networks with a Poissonian degree distribution. Also for Networks of Networks (NoN) or so called interdependent networks a critical percolation threshold could be calculated exactly (see Buldyrev *et. al.* [57]). Filippo Radicchi [**?**] published a comparison of common methods in a recent article, titled *Predicting thresholds in networks.* In his study of 109 real networks he found, that in less than 40% of the networks, advanced approaches based on the inverse of the largest eigenvalue of the networks adjacency matrix perform better than the naive approach based on the moments of the degree distribution. According to Radicchi, in general, all studied indicators behave worse as soon as the value of p_c becomes large. The percolation threshold of a network is an important property for comparison of networks from multiple domains. It should be considered in studies where many different types of networks are analyzed. The percolation threshold allows a common alignment of the networks, even if the networks have very different properties, such as density or degree distribution.

10.4. Filtering Correlation Matrices

A bi-modal or a multi-modal link strength distribution would be the ideal case. One could define a threshold between existing maxima. In reality we do not find correlation values which allow such a clear separation. Instead we find distributions with shapes where no simple model such as a Gaussian or power law can be fitted.

No matter what type of similarity measure we calculated and how the link strength distribution function looks like, it was always possible to identify significant differences between the distribution obtained from real data and those obtained from distributions calculated for randomized (shuffled time series) data. Therefore, we calculate both distributions, and compare them by a statistical test, such as the Kolmogorov-Smirnov (KS) test (see section 5.5). From this result one can conclude on a macroscopic level, e.g., if a measurable correlation between subsystems or within the entire system exist, or if the randomized data does not lead to a significantly different distribution. In this case the correlation matrix cannot be interpreted in a for network reconstruction useful way.

In case of significant correlations, one has to find out, which links are the relevant and significant links.

In general a parameter less approach (such as the KS test) is preferred. Another helpful approach is based on an individual significance test per link, e.g., by calculating the adjusted link strength using Eq. 9.10.

In the remaining part of this chapter I illustrate a major difference between threshold filters and structural filters. Both can be used to prepare a network for further topological analysis. A percolation analysis is not possible in case of MST or PMFG, as those methods already use specific topological properties to define the network.

10.5. Interpretation of Calculated Link Strengths

Initially we used only a Pearson correlation coefficient, calculated from access-rate time series pairs of Wikipedia pages to identify the hidden link structure in Wikipedia neighborhoods. We expected to find a relation between the access activity correlations and link creation events. Even if no link exists between two pages, they can have a similar activity pattern because of some real world aspects, which are not yet represented in the Wikipedia content network. The question is, which measure can be used to calculate a predictor or a precursor for link creation events from access-rate time series?

As part of this, our goal is to find out, how single peaks, long range correlation (LRC), and time series length influence the link strength of functional networks.

Also gaps and missing values influence the results clearly. One individual missing value can easily be replaced by the average value of the two closest neighbors. Such a simple replacement is not possible in case of longer periods of missing values. Missing values could be replaced by modeled data if an appropriate model exists, which can be used to simulate the process. Otherwise, we observed an artificially increased correlation which has no real meaning.

Chiu *et al.* [**?**] describe the same problems in the context of motif detection. Especailly for longer periods with low data quality or missing data they use so called *"don't care"* sections. This approach would finally also lead to a gap in the studied time series, especially in case of time-dependent analysis. Therefore, this approach seems to be useful only for data exploration, not for automatic analysis of large data sets.

10.5.1. Influence of Single Peaks

Wikipedia access-rate time series contain strong sporadic peaks as well as bursts of different shapes and periodic patterns. Keogh *et al.* [**?**] used three different types of pulses with different shape to study the influence of such disturbance on clustering properties of time series sub sequences. The shapes are in particular: funnel, bell, and cylinder (see fig. 7, 8, and 18 in [**?**]).

It is well known, that single peaks and periodic patterns in time series have a strong influence on the cross correlation coefficients. Before we apply a correlation measure to identify hidden links in large systems, it is important to understand the influence of noise, outliers, and defects in the data set. We study the impact of single sharp peaks added to white noise and simulate the cross-correlation function $\mathrm{CC_a}$ in order to get more information about the link strength l_{norm} calculated with Eq. 9.11 and show results in figure 10.2.

For our applications it is important to know, how the link strength l_{norm} calculated from correlation functions with sharp peaks behaves under certain conditions. Therefore we conduct an simulation experiment. Random time series (white noise) with one artificial peak are used for this calibration. We create time series of 28 values of a Gaussian distribution with $\langle x \rangle$=0 and a noise level defined by $\sigma \in [1, ..., 10]$ and place a peak of height h into the simulated noisy correlation function $\mathrm{CC_a}$ at position τ_p. This peak simulates a strong correlation at a given delay τ_p. The value for $\tau_p = 10$ is constant in this procedure and has no impact on the result. The peak strength h varies between 0 and 1000. This allows a variation of the signal to noise ratio. Time series without any peak can be seen as the cleaned time series, from which the strongest value was removed.

We analyze the influence of peak height h on the maximum link strength (see figure 10.2). The correlation strength increases as a function of the signal noise ratio. Figure 10.2 shows a minimal link strength of $l_{\mathrm{norm}} \approx 2.25$ and a maximum value of 5. Depending on the variance in the correlation function (without the maximum value)

one can extract the minimum peak height which would cause a certain link strength. Link strengths above 5 are not likely to be caused by a single strong peak in the correlation function. Thus we can consider higher values (if calculated via Eq. 9.11) as significant links. The curves in figure 10.2 are useful for calibration of weaker links with $l_{\text{norm}} < 5$, depending on σ of the calculated distribution of link strengths CC_a.

A Quality Measure for Pearson-Correlation Functions

The correlation function $\text{CC}_\text{a}(\text{t}_0, \tau)$ has to be calculated for an appropriate number of time delays τ. For data with recurring patterns one should choose $\tau \gg t_p$ where t_p is the length of the recurrence period of the pattern. In case of only one single sharp peak in $\text{CC}_\text{a}(\text{t}_0, \tau)$ we have a strong indicator for a strong correlation at a given delay τ. Periodic cycles or patterns would cause multiples of such peaks if the delay is chosen large enough. This allows us to define a quality criteria. First, we calculate the normalized link-strength l_{norm} using Eq. 9.11 and next we remove the maximum value from CC_a and replace it with the average value $\langle \text{CC}_\text{a} \rangle$ and repeat the link strength calculation to obtain l_{trans} using Eq. 9.12 and this finally defines $l_{\text{trans}} = QM$, our *quality measure*. This approach is not sensitive regarding the link strength and allows one to identify weak, but clear correlation peaks within noisy data.

10.5.2. Adaptive Significance Tests

In general, there exist two categories of significance tests. The first and simplest one compares the full link strength distributions p_l calculated from raw data and p_{shuffle} for which links were calculated from shuffled time series data. All relevant time series properties, such as long-term correlations, auto-correlation, and cross-correlation between two series, have been changed or removed during shuffling while all aspects, which do not influence the interpretation of results (distribution of values, maximum, minimum) have been preserved. Instead of changing the measured data it is also possible to use time series sequences from different periods to calculate the link strength distribution p_{dt}, so that no correlation between the time series has to be assumed. A systematic cross-check allows a comparison of this distribution p_{dt} with p_{shuffle}. They should not differ significantly from each other but both should differ significantly from p_l. A useful quantitative test is the Kolmogorov-Smirnov test.

This allows not yet an interpretation of individual link strengths but it helps to describe the system on an abstract level. Based on this idea it is possible to verify, if a given process can be modeled as a network, and if the results are not just random or artifacts of the measurement or analysis procedures.

A more detailed link strength significance test is based on an individual comparison for each link. The calculated link strength is compared with a number of randomized results (see l_{adjusted} calculated with Eq. 9.10). Furthermore, one can calculate multiple correlation values on a higher time resolution, e.g., instead of daily data for one month, the hourly data for each day is used. Now, one uses the median value from all days to represent the correlation during the month. This approach was also evaluated by Berit Schreck [134] in her Bachelor thesis titled: *'Rekonstruktion komplexer Netzwerke mittels Kreuzkorrelationsmethode'*.

10.5.3. Functional Links from Time Series Pairs with Long-range Correlation

How do long range correlation (LRC) and time series length influence the link strength of functional networks? In this section we investigate this problem using simulations. The shape of the probability distribution function

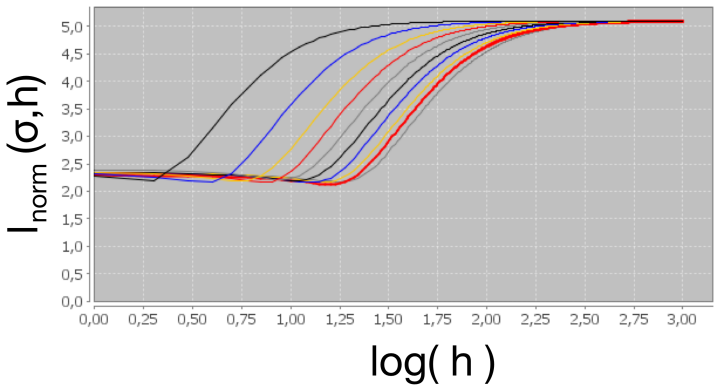

Figure 10.2.: **Link strength in the presence of sharp peaks in correlation functions**. Simulated correlation functions $\text{CC}_\text{a}(\text{t}_0, \tau)$ with $\tau = \pm 14$ were used to calculate the link strength l_{norm} for variable noise σ and peak heights h. An artificial peak of height h was added to a series of white noise with variance σ equal to 1.0 (black), 2.0 (blue), 3.0 (yellow), 9.0 (thick red line), and 10.0 (gray) (curves are ordered by sigma from left to right). For Eq. 9.11 we find link strengths l_{norm} between 2.25 and 5.

(quantified by the moments of the distribution) should change significantly compared to the shape calculated for shuffled time series if LRC exist in the data. This is our Null-hypothesis. Detection of such a significant change allows the conclusion that an effect exists in the data, different from randomness. Or in different words: the observed behavior does not happen just by chance if a significant difference for raw data and surrogate data can be identified. We use the following procedure: LRC is introduced into random time series via Fourier Filtering (see section 5.6.2). In order to study the influence of LRC we vary the parameter β (see section 5.6.2 and figure 5.11 for more details on β) which influences the strength of LRC per time series. Furthermore, we vary the length of the time series. Functional links are calculated in multiple modes as shown in table 9.4 and figure 10.3.

We calculate link strength values s for 4950 pairs from 100 individual series to get the probability distribution function $P(s)$. This distribution describes the system state during the chosen time range specified by the length of the time series. Repeating this procedure on sliding windows leads to a representation of the time evolution of the system state. The time resolution or the length of the episode depends on the actual use case (see table 10.1). Sometimes we get hourly resolution (Wikipedia access statistics) or even one data point per minute (technical log data) but in some cases we can get one value per day only (public stock market data). For our simulations with variable length in the range $l = [32 \ldots 524288]$ we found the time scales listed in table 10.1. to be relevant.

Length	Usage
32	range of one day for hourly resolution or one month with daily data
512	range of one year for daily data or one month for hourly data
8.192	range of one year with hourly data
524.288	one data point per minute for one year

Table 10.1.: **Time series length and their particular usage context.** The numbers have to be multiplied by the amount of memory one uses to represent a data point, e.g., 4 bytes for a number, in order to estimate the overall memory consumption per time series operation. This is a limiting factor even in parallel computation environments and influences system design and configurations.

Experiments were conducted with $\beta \in \{0.0, 0.2, 0.5, 0.8, 1.0, 1.2, 1.5, 1.8\}$ for 100 time series of length $l = 32$ and $l = 8192$. The four link strength calculation modes for which link strength distributions are shown in figure 10.3. are named *mode 0* (Eq. 9.11) in (a), *mode 1* (Eq. 9.8) in (b), *mode 2* (Eq. 9.10) in (c), and *mode 3* (Eq. 9.12) in (d).

Qualitatively we find larger deviations between the distributions for raw and surrogate data for larger β which means, larger LRC in both of the two time series lead to larger correlation link strengths (in mode 0, mode 1, and mode 2). Furthermore, s is dependent on the length of the time series. We found that the highest link strength values grow with increasing β and with an increased length of the series (see figure 10.5). Also, for mode 0 the link strength distribution is asymmetric and allows an easier interpretation of the change of the entire link strength distribution using its moments while mode 1 and mode 2 link strength distributions seem to be symmetric. For mode 1 we can identify a strong dependence between LRC and the width of the distribution (quantified by the

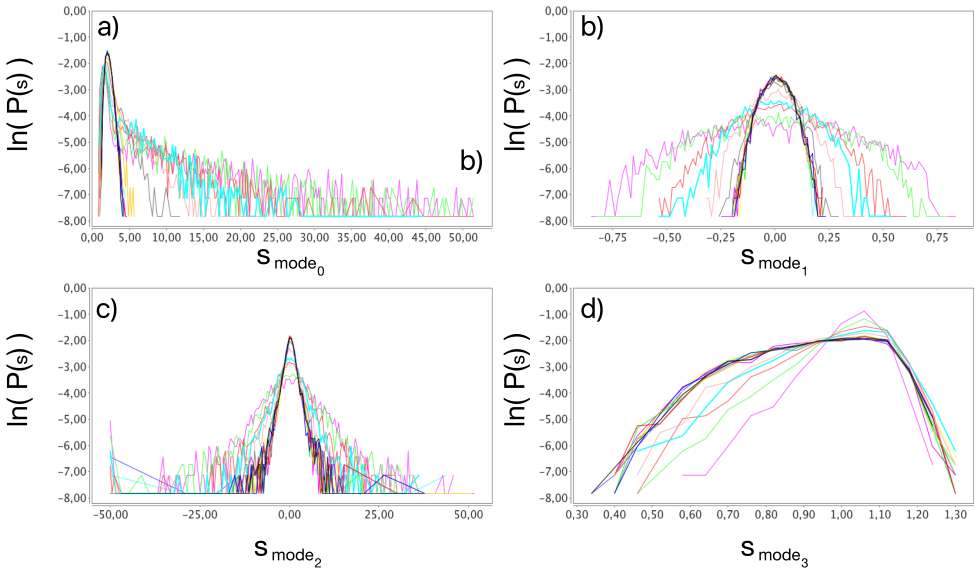

Figure 10.3.: **Link strength distributions as a function of long-term correlation.** For different link strengt calculation modes (a) Eq. (9.11), (b) Eq. (9.8), (c) Eq. (10), and (d) Eq. (9.12) for variable long-term correlation strength β. Note, that y-axis uses log scale.

variance).

We use mode 3 to test if LRC causes narrow peaks in the cross-correlation function when calculated by mode 0 (not shown). Our simulations show that there is no evidence for such an influence. Independently of the control parameter β we can see a maximum around 1. The number of strong links, which are changed by replacing the maximum value by the average of the cross-correlation function decays with increasing LRC. For comparison we added an artificial peak to the cross-correlation function. This illustrates that if a strong link for one particular delay was identified, this link is not likely to be a consequence of LRC independently of the long-term correlation strength per series.

Observation: Asymmetric deviations in link strength distributions calculated by mode 2 are not caused by LRC in the sample series. Such deviations can not be interpreted as artifacts caused by LRC. In this way the 3-rd moment of the distribution can become an indicator for tracking the activity related correlation. The average value of the distribution function (1-st moment) can only be calculated and compared with others if the PDF converges. The second moment (variance) describes the spread of the values. In many publications such values were provided to track the time evolution of the link strength. As a result of our calibration experiments we propose to use two different indicators, which can both be calculated based on obtained probability distributions from two different link strength computation modes. Regarding LRC we find stable and independent mean values for mode 1 and mode 3, independent skewness for mode 1 and an independent kurtosis for mode 1 and mode 2.

In general, the mean value of the distribution should not be used for comparison of probability distributions if they are not Gaussian distributions, or if the shape of the distributions is not stable. Finally, mode 0 and mode 2 are useful for future studies. We identify two measures to describe the link strength distribution based on our findings:

The third moment (also called skewness) has a high positive value if the tail of the distribution is longer on the right and a negative value in case of a long tail on the left. A symmetric link strength distribution function would lead to zero.

Figure 10.4 illustrates a dependency of skewness on β for mode 0 link strengths in (a) and mode 2 link strengths - also used as significance level - in (b).

Beside the multi-modal fit (a parable open to the bottom and a linear function on the right side as proposed in section 10.7 we calculate the kurtosis, which is known as the fourth central moment of a distribution. The heaviness of the tail of the distribution in comparison to a Gaussian distribution (while we assume that both have the same variance) is measured by kurtosis. This procedure is less expensive than the previously mentioned fit functions and thus it is a good candidate for screening of large data sets over long periods with high time resolution.

Because it seems to be misleading to rely only on the link strength value we investigate additional properties which are related to stationarity, or to the existence of LRC. If the time series is nonstationary ($\beta > 1.0$) we find many high correlation values which seem to be significant but are caused by inherent trends. This illustrates the importance of ensuring stationarity by application of a detrending method or a simple transformation of the raw data such as using the first differences instead of raw data.

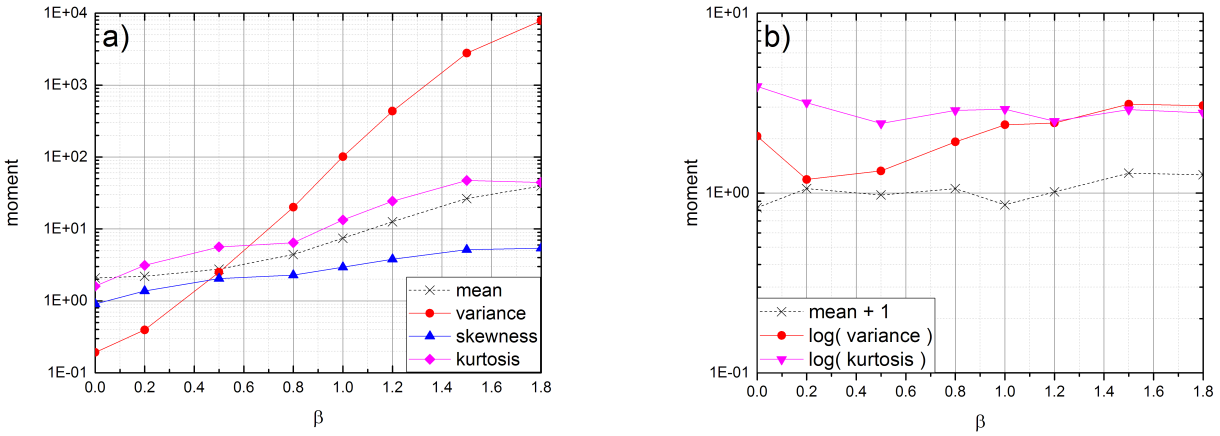

Figure 10.4.: **Quantitative comparison of link strength distributions.** (a) Link strength distributions calculated via Eq. (9.8) show a dependency between the four moments of the distribution (mean, variance, skewness, kurtosis) and the control parameter β which controls LRC. The strongest slope was found for variance. (b) Mean and kurtosis seem to be independent from β if links are calculated by Eq. (9.12). Only for variance we find a weak dependency from β. This allows the conclusion, that link strengths calculated in mode 3 is less influenced by long range correlations.

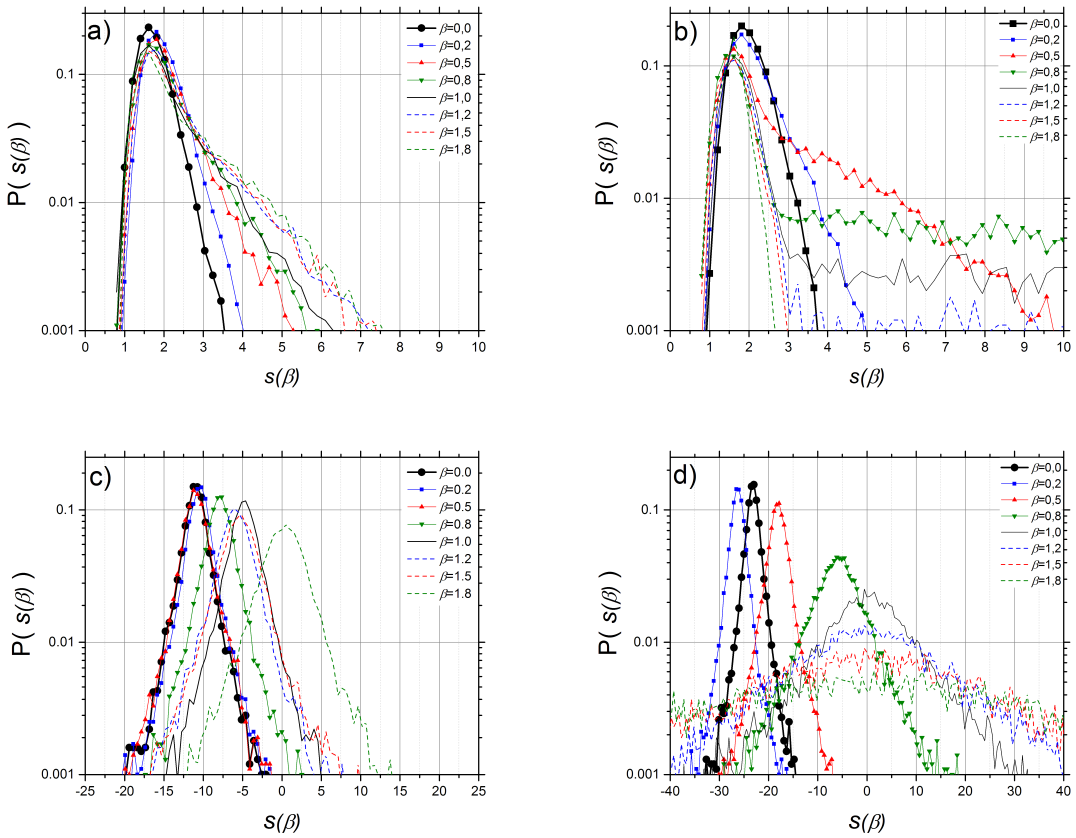

Figure 10.5.: **Link strength distributions as function of long range correlations for two link creation modes.** The influence of long-term correlations on functional link strengths is analyzed for two link strength calculation modes. (a,c) show link strength distributions for short time series (l=32) and (b,d) for longer time series (l=8192). (a,b) show link strength distributions for links calculated with Eq. 9.11 (mode 0). (a,b) In absence of long range correlations we find an asymmetric function which can be approximated by a quadric and a linear function as illustrated in figure 10.9. The asymmetry is changing as a function of the control parameter β while the overall position of the function (see left parabolic branch) stays stable. (c,d) show link strength distributions for links calculated with Eq. 9.10 (mode 2). The dependency between the shape of the distribution function and the control parameter is obvious, but the position is not stable. This means that the mean value cannot be interpreted in a reasonable way. The variance also increases with increased length l and with higher LRC generated by Fourier Filtering with $\beta = 0.0$ (black line and circle), $\beta = 0.2$ (blue line and square), $\beta = 0.5$ (red line and triangle up), $\beta = 0.8$ (green line and triangle down), $\beta = 1.0$ (black line), $\beta = 1.2$ (blue dashed line), $\beta = 1.5$ (red dashed line), and $\beta = 1.8$ (green dashed line).

10.5.4. Identify Significant Event-synchronization

The following section concludes our preliminary findings and describes practical aspects of ES implementation and usage. Furthermore, results from [164] are summarized. With our implementation of the ES algorithm in the Hadoop.TS software package [11] it is possible to calculate the synchronicity of time series pairs in large-scale data sets using Apache Hadoop or Apache Spark on clusters of distributed computers.

In order to interpret Q values as a representation of a hidden functional link in a correlation network one has to distinguish functional event-synchronization from random event-synchronization. Two computational experiments helped us to study event synchronization properties systematically. Both experiments are based on work from Malik *et al.* [136].

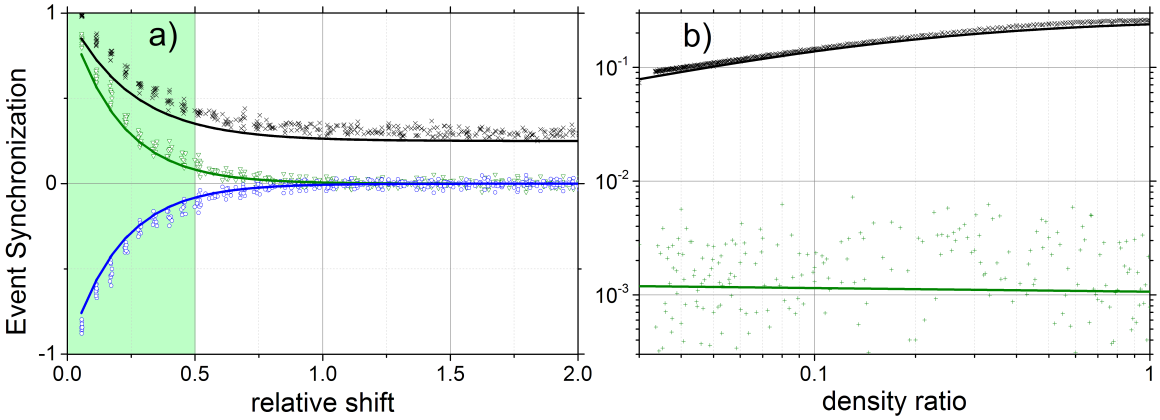

Figure 10.6.: **Event Synchronization Q as a function of relative shift (a) and density ratio (b)**. The relative shift parameter influences the event synchronization Q between identical random event time series. The relative shift is the quotient of time shift Δt and average inter-event time (reciprocal density). Q values start at 1 for identical time series and follow an exponential decay to a value around $Q_{min} \approx 0.3$. For a relative shift greater than 1.5 there is no significant decrease measurable. Depending on the density ratio the event synchronization Q for random event series can be very small as shown in (b). The green and blue lines show the average delay value q for positive and negative shifts. q is expected to be 0 for random event series. Black line shows the fit function $Q = \frac{0.7}{(5+x)^{0.6}}$. The green area marks the range in which the orientation can be determined.

As shown in figure 10.6.a, even for random event series the event synchronization strength Q_{ij} (black crosses and black curve) is not equal to zero. A minimum synchronization strength, which depends on the relative shift and on the ratio of the event densities of both series was found. In case of a pair of ETS with equal density the synchronization strength is $Q_{ij} \approx 0.28$ even for really large relative shifts. Figure 10.6 shows simulation results and the exponential fit for Q and q. The delay value disappears for higher shifts. Furthermore we found the following function Q_{rand} to approximate the dependency between random synchronization and the density ratio x:

$$Q_{\text{rand}} = \frac{0.7}{\left(5 + \max\left(x, \frac{1}{x}\right)\right)^{0.6}} \tag{10.1}$$

Experiment 1 - Transition from identical to independent event series

First, event synchronization is calculated for pairs of identical time series with random events. One of the event series is shifted against the other by Δt. Figure 10.6.a shows event synchronization strength Q (black crosses) and delay q for event series with equal density. The considered numbers of events per 8760 time steps (which corresponds to an hourly time resolution for one year) are in the range $s \in \{100, 200, 300, 400, 500\}$. q values for positive relative shifts (green triangles) and negative relative shifts (blue circles) are monotonous functions. Both series have the same density. Q reaches a maximum for $\Delta t = 0$ and a minimum of $Q_{\min} = 0.28$ which seems to be independent of relative shift and density at higher relative shifts. We subtract this empirical saturation value and calculate the parameters for an exponential fit function as shown in figure 10.6.a as a black line:

$$Q(t) - Q_{\min} = a \cdot e^{-b \cdot \Delta t / T} \tag{10.2}$$

where $a = 0.7$, $b = 4$, and $\Delta t/T$ is the relative time shift between the time series.

The exponential part of the fit-function can be explained as follows: We generate event time series of independent events. This means at each time an event can be registered with the same probability. Such a process is a Poisson process and the inter event times, which are distances between two events, follow a decreasing exponential

distribution.

$$p_{\text{poisson}}(t) = \frac{1}{T} e^{-t/T} \tag{10.3}$$

where T is the mean inter-event time and t a particular inter-event time between consequent events.

Two events $(t_l^i,\ t_m^j)$ are synchronous if δ is smaller than all four inter event times between two consequent events from within the time series for which an event was registered. Thus, four distances contribute to the value of τ according to Eq. (5.11). Because all four distances are independent from each other and from the time shift we can calculate the probability for a pair of events to be synchronized as the joint probability of having all inter event distances smaller than δ. We define $\delta = |t_l^i - t_m^j|$ and write the probability $p_{\text{sync}}(\delta)$ for finding a synchronized event pair at distance δ:

$$p_{\text{sync}}(\delta) = p_{\text{poisson}}(\delta)^4 = \left(\frac{1}{T} e^{-\delta/T}\right)^4 \tag{10.4}$$

Using this simplification we are able to explain the exponential part of the fit function Eq. (10.2), but not the vertical offset. For large time shifts the synchronization of two initially synchronized events disappears but there can be a synchronization with a following event which causes a measurable synchronization strength even if initial synchronization is destroyed.

The delay value for time shifts larger than the average inter-event time (in this case the relative shift is one) fluctuates around zero as shown in figure 10.6.a. The randomly generated event time series are supposed to be statistically independent. The green and blue curves in 10.6.a represent the exponential fit function: $q(t) = e^{-b \cdot t/T}$ with $b = 5$. The value $b = 5$ can result from overlapping the decay function for Q and the relation between q and Q. According to [164] for low shifts, all pairs of events have the same direction of delay as long as they are still synchronized, thus $|q|$ is close to its maximum: $|q| \approx Q$. Then some pairs of events start to newly synchronize. These newly synchronized pairs' delay direction is opposed to the general direction, which is given by the shifting direction between the time series, so they decrease the result instead of increasing it (as in the case of Q), leading to a fifth $e^{-t/T}$ factor and $b = 5$. Shift direction determines the sign of q. They can be distinguished as long as the relative shift is less than 0.5, indicated by the highlighted range in figure 10.6.a.

This experiment provides a first empirical calibration value, a threshold to distinguish significant (non-random) event synchronization from non-significant synchronization which is also detectable in pure random event series. The second experiment takes into account that also different densities and density ratios have an influence on Q_{rand} but hardly on q_{rand}.

Experiment 2 - Influence of density differences on ES

In the second experiment we calculated ES for two independent time series with variable densities. We expect a symmetry around the density ratio 1. Time series have a length of 17520 (which is related to two years with hourly resolution). Event numbers for series i are in the range [10...1000] for series j the number of random events was $s_i \in \{10, 260, 510, 760, 1010\}$. Figure 10.7 shows the results for 1000 computations in a log-log plot together with the fit function Eq. (10.1) for comparison.

Results from both experiments allow implementation of an automatic calibration method. We know the influence of density differences on synchronization strength. So we calculate a relative event synchronization for measured data, as the quotient of the calculated synchronization strength and the theoretically expected value which is calculated from density ratio and the empirical calibration function Eq. (10.1). We apply a threshold filter to identify relevant or significant links between randomly chosen pairs of Wikipedia pages. During network reconstruction procedure we use only those pairs for which the calibrated result $Q_{\text{cal}} = \frac{Q}{Q_{\text{rand}}} \geq t_f$ where t_f is a variable threshold parameter.

10.6. Threshold Filters

A threshold filter is based on a single property, derived from a density distribution function of link strength values.

No structural properties and no node group properties are used to separate relevant links L_N from non relevant links L_P (see section 10.1). Hence, in the following sections we evaluate two methods for dynamic threshold identification and for link classification. Therefore we calculate two additional properties beside the previously created correlation functions. By stretching the PDF function in a two dimensional plane we are able to identify and separate clusters of links with comparable properties.

Finally, our goal is a structural comparison of networks as a function of time. Therefore we use the adaptive threshold approach to define and identify relevant links. In a next step, a variation of the threshold shows if the requested network properties change significantly as a result of the variation. In this way, we can calculate time-dependent network properties and stability criteria. Results will be presented in chapter 13.

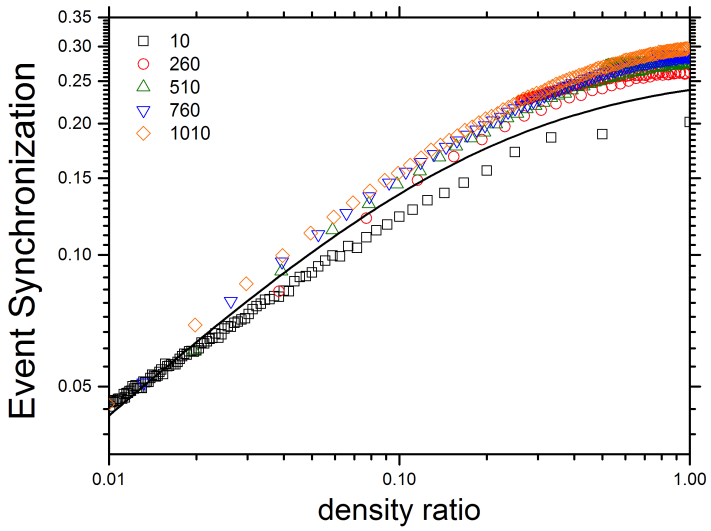

Figure 10.7.: **Event Synchronization strength as a function of density ratio**. The density ratio influences the event synchronization between random event time series. Q (open symbols) was calculated for event series of variable density. The length of the time series was 17520, which is related to two years with hourly resolution. The event numbers for series i are in the range [10...1000] and for series j the number of random events was in the range $z\epsilon\{10, 260, 510, 760, 1010\}$. A comparison of the simulation results (open symbols) with a fitted calibration function (black line) highlights a difference between both especially for higher density ratios. For small density ratios $\rho < 0.1$ the fit function can be used as a calibration function.

10.6.1. Fixed Threshold

The metric used for reconstruction of networks in figure 10.8 is semantic similarity (see section 9.2.2). In this case the weak links are not relevant. Thus, we can apply the fixed filter threshold approach. Some strongly connected clusters appear first. With increasing link strength the cluster's link density increases and at a threshold of $\approx 10\%$ also links between clusters emerge (see figure 10.8.a).

Because different link strength metrics lead to very different shapes of the PDF different thresholds are required. Also, because of variable distributions over time, a fixed threshold can not be used in general. Instead, a parameter less approach is required.

10.6.2. Dynamic and Adaptive Thresholds

Because a clear indicator for separation of significant and non significant links does not exist, and because a threshold (even if it could be found) would not be stable over time we describe an geometric approach to derive the separation threshold from the measured PDF using a simple fit procedure.

The method - illustrated in figure 10.9.a - is based on the assumption that a random link strength distribution has a Gaussian distribution and thus can be fitted by a parabola (open to the bottom) if plotted in log scale. The relevant links, caused by correlations in the time series from which links strengths were calculated follow an exponential distribution which is combined with the Gaussian distribution. The exponential part is represented by a line in the logarithmic plot. Figures 10.5.a and 10.5.b illustrate this common shape for links, calculated in mode 0 using Eq. 9.11.

The vertex position $x_{\max}$ (position of the maximum value $y(x)$) of the parabola is related to the average link strength (assuming a symmetrical distribution as in the case of correlation values for randomized data). The variance of the vertex positions obtained for multiple time intervals is interpreted as a stability criteria.

The point, where both curves (parabola and line) cross is defined as t_l and if they do not cross we set $t_l = x_{\max}$. The maximum position where the fit parabola crosses the x-axis defines $t_u = \max(x_{r1}, x_{r2})$.

To find the dynamic filter threshold for relevant links we have to identify the value $x = t_{ds}$ for which the probability of having a random link or a relevant link is 50%. If $x > t_u$, than we have a relevant link for sure, and for $x < t_l$ we consider all links as random links, regarding the property plotted on x axis.

The blue shaded area above the green curve, under the red curve and between t_l and t_u must be divided such that the resulting two blue shaded areas are of equal size.

We describe the obtained link strength distribution as bi-modal distribution, one part as a Gaussian and the other part by an exponential fit with negative exponent. In a semi logarithmic representation the fit functions become $\mathcal{P} = a \cdot x^2 + b \cdot x + c$ a parabolic and $\mathcal{L} = m \cdot x + n$ a linear function with parameters a, b, c, m, and n.

First we have to obtain the fit parameters from correlation data from each time step. In this case the fit

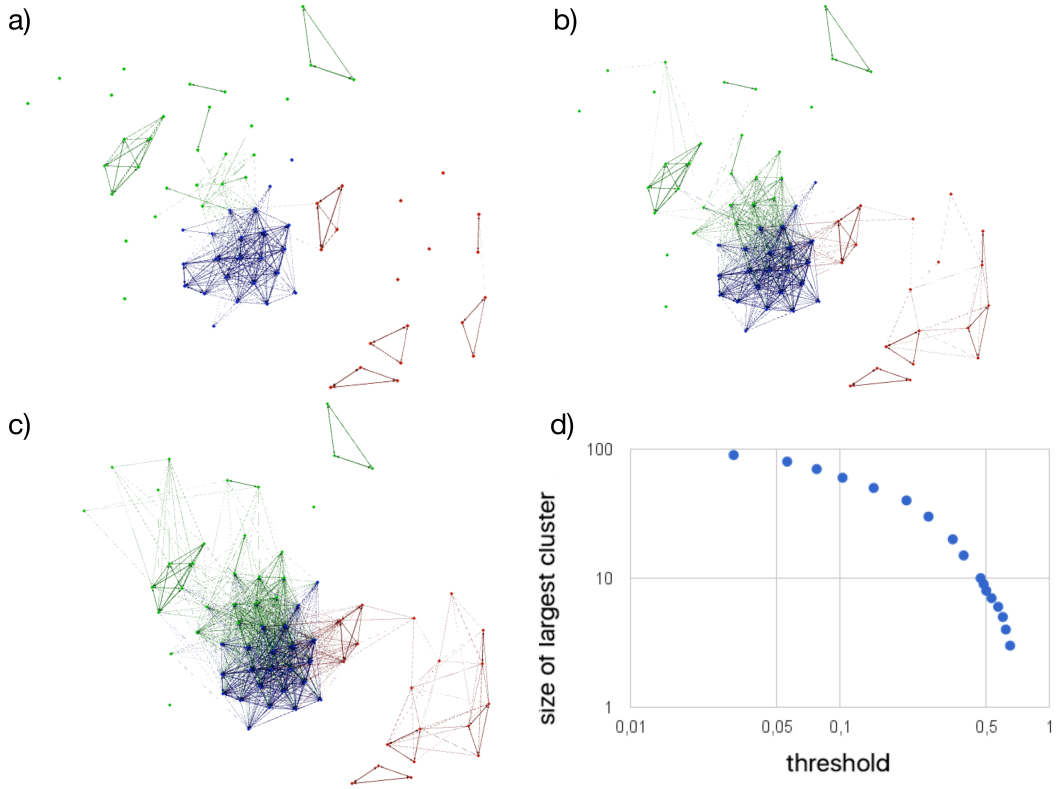

Figure 10.8.: **Emerging clusters**. Clusters of strongly connected nodes emerge as a result of decreasing filter threshold. Links between the document clusters emerge after about 10% of links were added to the network, starting from strongest to weakest. Filter thresholds correspond to graph sizes (a) 10%, (b) 20%, and (c) 30%. The Maximum Spanning Tree (see Fig. 10.11.c) contains only 1.47% of all links. **Network size as a function of filter threshold**. (d) shows how the size of the extracted (filtered) sub graph decreases monotonously with decreasing filter threshold.

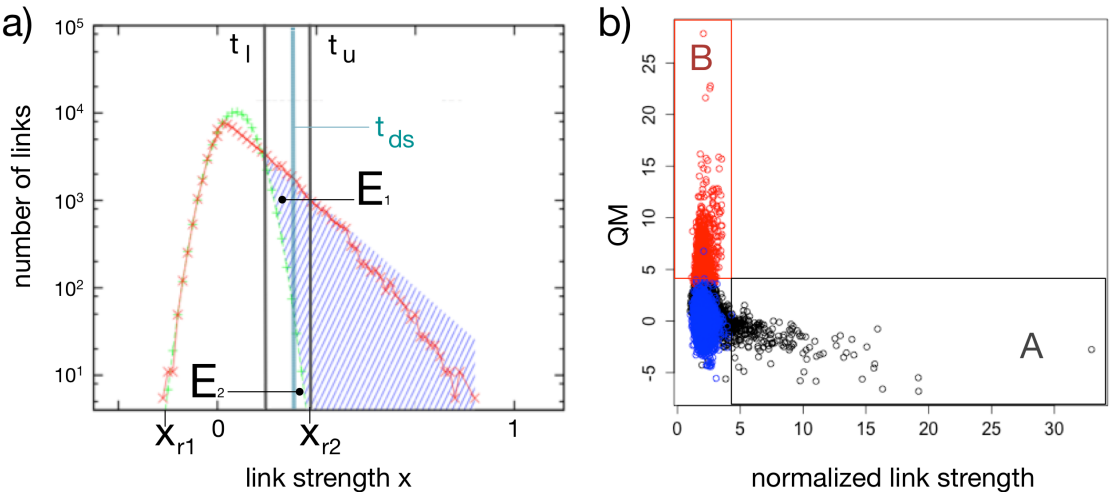

Figure 10.9.: **Adaptive filter threshold calculation**. The adaptive filter threshold respects systematic differences (as a result of extreme events) and also trends (found in open systems which are not in equilibrium). The first approach (a) defines the threshold in a way, which splits the overlapping range in two equal parts, which means we can have a significant link or an artificial links with the same chance. Grouping links using k-means clustering as shown in (b) is a new approach based on an automatic machine learning method.

parameters are time-dependent properties of the system.

$$\int_{x=t_l}^{t_{ds}} (e^{\mathcal{L}} - e^{\mathcal{P}})dx = \int_{x=t_{ds}}^{x_u} e^{\mathcal{P}}dx \tag{10.5}$$

$$\int_{x=t_l}^{t_{ds}} (e^{m\cdot x+n} - e^{a\cdot x^2+b\cdot x+c})dx = \int_{x=t_{ds}}^{x_u} e^{a\cdot x^2+b\cdot x+c}dx \tag{10.6}$$

Next, we have to find t_{ds} in Eq. 10.6 using a numerical approach.

Berit Schreck presented first results of this approach in her Bachelor thesis [134] in tab. 5 on page 22. After normalizing the fit parameters she obtained time series to characterize the time evolution of the system, based on a collective property representing the whole system. A seasonal structure seems to be visible in figure 15 on page 28 in her work, but due to the short time series we cannot conclude this for sure from her data.

Finally, we developed another approach, which uses data points in a 2D plane as shown in Fig. 10.9.b. We work with three 2D planes defined by two calcualted values. The planes a defined as follows: (name: equation for values on x-axis, equation for values on y-axis): (AC: Eq. 9.11, Eq. 9.10); (AD: Eq. 9.11, Eq. 9.12); (CD: Eq. 9.10, Eq. 9.12). Figure 10.10 (a,d) show AC-plane, (b,e) show AD-plane, and (c,f) show the CD-plane.

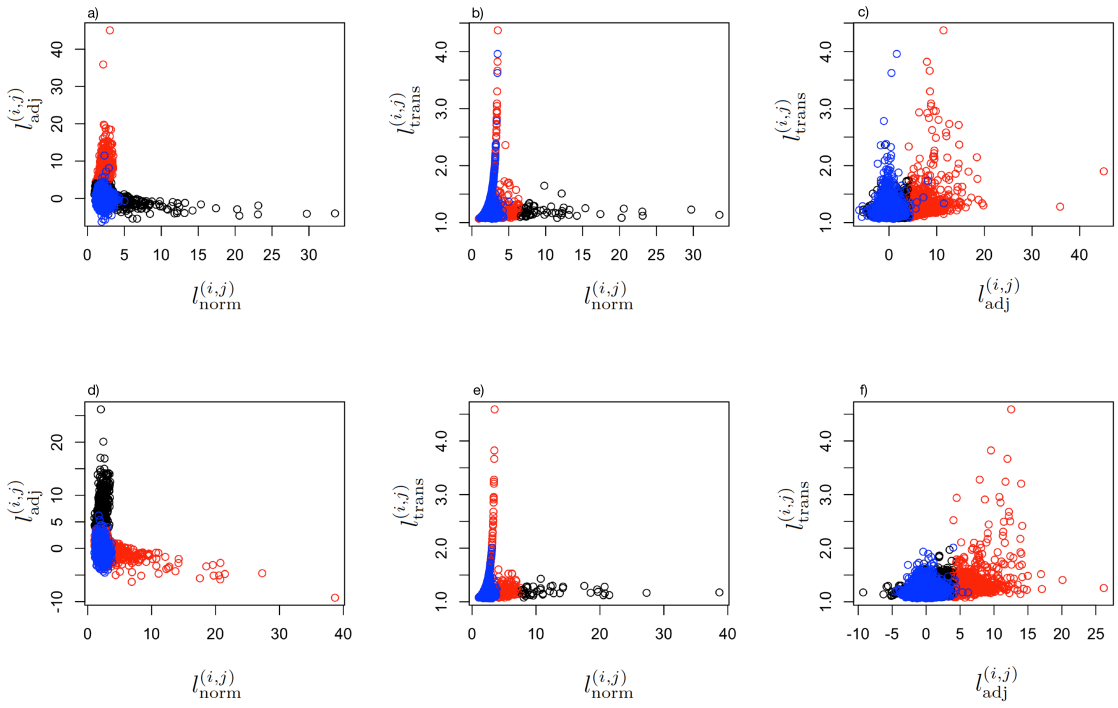

Figure 10.10.: **Comparison of link strength and quality measures**. Filter thresholds can be found in 2D histograms or in scatter plots using k-means clustering. Daily access-rate time series from the German Wikipedia page '*Amoklauf von Erfurt*' (a,b,c) and for the Wikipedia page '*Illuminati (book)*' from 1-st quarter 2009 were used to calculate three metrics $l^{(i,j)}_{\text{adj}}$ (Eq 9.10), $l^{(i,j)}_{\text{norm}}$ (Eq 9.11), $l^{(i,j)}_{\text{trans}}$ and (Eq. 9.12).

As comparison with randomized data the blue circles are plotted on top of the two clusters (red,black) obtained from k-means clustering. k-means clustering works best in AC plane (see panels a,d) and allows us to detect the threshold along the y-axis for adjusted link strength. The AD plane allows identification of a threshold along the x-axis for normalized link strength. Finally, the CD plane illustrates the effect of the quality measure. Even below the threshold, there exist sharp peaks in correlation functions which lead to rather weak links. When the logarithm is applied to raw time series, this effect is slightly weaker, which gives some more control on the filter behavior.

By applying the k-means clustering algorithm it is possible to identify two groups of significantly different links. The figure shows the group of strong links (A) in black and the weak, but significant links in red (B). The thresholds along the x-axis and the y-axis can be found as explained above in this section. By using k-means clustering with three initial cluster centers we can also estimate the boundary between weak but relevant and strong links.

10.7. Structure-Based Filters

The full graph can be used as source for information especially to apply graph analysis algorithms. For useful visualization an appropriate filter is required. A very common approach is to analyze properties of clusters as shown in figure 10.8. The clusters in figure 10.12.a are more obvious than those in 10.11.a. Thus, our mixed approach uses a link strength filter and an algorithm to find connected components. In this case the size of connected components is influenced by the filter threshold. Only if this size is a stable property one can assume to have a usable filter. Otherwise one can also test the filter threshold by measuring the change in the structural property caused by a variation of the filter threshold. Figure 10.11 illustrates the difference of the maximum spanning tree (c) and the minimum spanning tree (b). Depending on the selected metric, it is necessary to select the right spanning tree or to transform the link strength values.

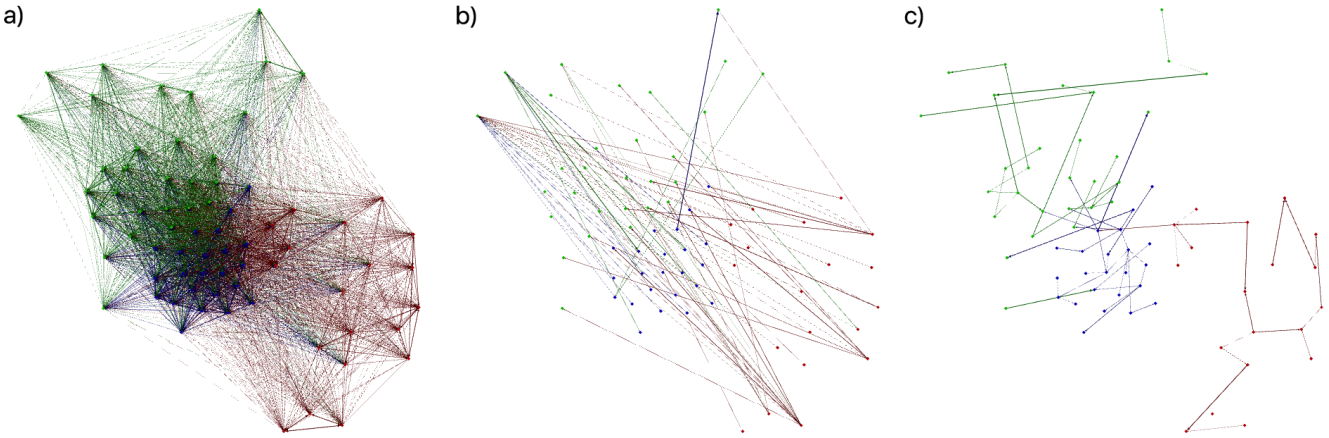

Figure 10.11.: **Comparison of structural filters**. Minimum- and Maximum-Spanning trees obtained from an unfiltered correlation network. (a) An initial layout was calculated for the full network, using a force directed layout algorithm. (b) shows the minimum spanning tree, which is not very useful in a case where the stronger links are the relevant links and the weak links are likely to be just random. Instead, the maximum spanning tree is created in (c) to extract a meaningful sub graph.

Fan *et al.* [**?**] published a paper in 2004 titled: *'Network of Econophysicists: a weighted network to investigate the developement of Econophysics'*. They use a weight to express special properties of network nodes, dependent on the node type. We can use such weights also to modify multi-layer networks which can finally be projected or transformed into one graph for which finally classical network measures are calculated. Specific clusters of interest are percolation clusters and clusters, obtained from clique percolation as shown in the right image in figure 10.12 (see also [**?**]).

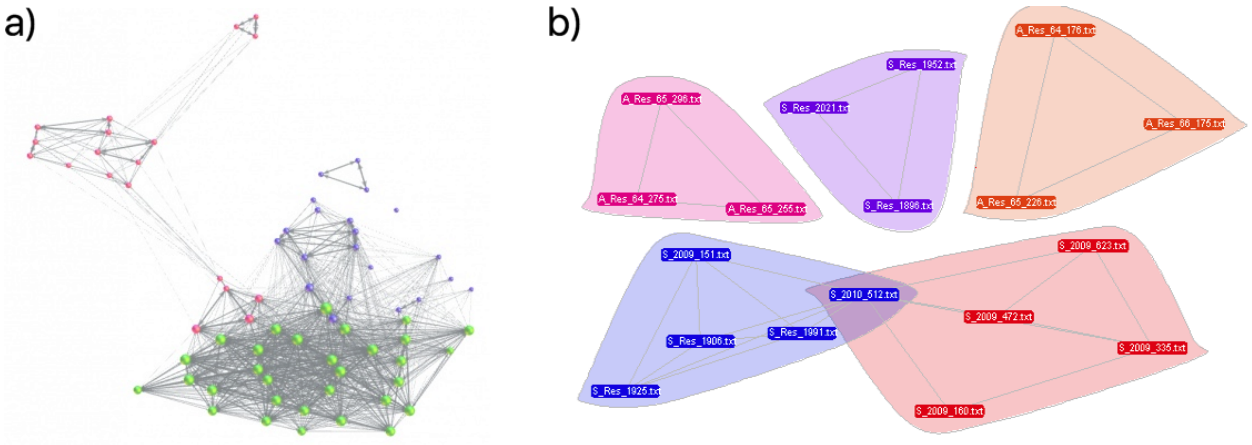

Figure 10.12.: **Threshold filter vs clique percolation**. (a) Network clusters become visible after application of a *filter*. (b) Clusters of heavily connected nodes can also be found by *clique percolation*. The software CFinder (see [**?**]) is an alternative approach to filters.

10.8. Conclusion

An appropriate selection of the right correlation measure and the right filter methods including the right parameters are important aspects in time series based network reconstruction and analysis. Therefore, calibration and a systematic analysis of external influences are required. Such influences can be, e.g, very strong peaks or patterns in raw time series. Data inspection procedures serve as a foundation for decisions regarding filter algorithm and parameter selection.

Different link creation measures emphasize different functional aspects of a system. A structural comparison of those networks allows one to identify differences between subsystems of complex systems. One can extract metrics to describe the coupling between subsystems by simple analytic functions, but all this is misleading if the selected procedures are not in line with the desired outcome, e.g, weak ties can be overseen and existing obvious structures could be over estimated.

The distribution properties, such as average link strength, or the moments of the distributions are helpful for analysis on a macroscopic scale and allow fast access to time-dependent properties on a system level, but they can not be used to characterize individual links.

The individual significance of each link should be taken into account. This can be achieved by calculating a relative link strength as the ratio between the measure on real data and randomized data. Additional quality metrics allow application of parameterless adaptive methods for automatic link classification.

If a similarity measure is used and only a high similarity between nodes is considered as a relevant information, the Maximum Spanning Tree and the Planar Maximal Filtered Graph are useful to extract informative sub-graphs. Alternative approaches are static or adaptive threshold filters, but one has to be aware of the fact, that weak links are systematically ignored by this methods.

Therefore, an alternative was introduced, e.g., one can separate groups of random and meaningful links using machine learning techniques, e.g., k-means clustering as shown in figure 10.10.

Functional networks can finally be created from multiple independent link groups resulting from multiple filters. The layers are combined in a property graph, which means, links of different categories, types, or groups are merged to one multi-layer network.

Another problem arises, if link strengths differ much over time, if they span multiple wide ranges of values, and if they are not normally distributed. In this conditions especially in case of time-dependent analysis the whole ensemble of links can not be represented by the average value.

As shown in this chapter, many different aspects influence the strength of a correlation link and the link strength distribution, especially if the system is not a closed system or if it is out of equilibrium. Finally, the right filter approach leads to a useful representation of the entire system.

11. Measuring Context Sensitive Relevance

> It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.
>
> *(Richard P. Feynman)*

This chapter introduces a new method to calculate a time-resolved context-sensitive measure for the relevance of interlinked digital resources like web pages, messages, news articles, or arbitrary documents. Such objects are usually embedded within a specific context. In Wikipedia such contexts can be represented by category pages or by a local neighborhood graph (see figure 6.2). In general, on the WWW contexts are defined by links between web pages. In the following we will define the representation index REP and the time-resolved relevance index REL, which both allow to distinguish content of a more local relevance from content of a more general or even global relevance at variable geo-spatial and temporal resolution.

The data is collected from Wikipedia in a semi-automated procedure, which allows a near real time analysis of time series data. Continuous time series, event time series, explicit link structures, and implicit semantic annotations can all be used together. Results are visualized as relevance plots. This chart type supports an advanced interpretation of the results in various interdisciplinary research contexts, especially including econo-physics and socio-physics.

11.1. Introduction

Since the numbers of hypertext pages and hyperlinks in the WWW have been continuously growing for more than 20 years, the problem of finding relevant content has become increasingly important. This led, for example, to the growth of Google Inc. with its mission statement *'to organize the world's information and make it universally accessible and useful'* [**?**]. Initially, the WWW was mainly a content network and did not reflect relations between authors. It provided structured and connected information. However, the appearance of Social Media Applications (SMAs), such as Facebook, LinkedIn, and Twitter, with friendship and follower relations between individual users has led to the creation and simultaneous evolution of novel user community networks (social networks) together with content networks. Such SMAs can be regarded as networks of networks because the underlying user and content networks are closely inter-related with each other [**?**]. The collaborative creation of linked content became very popular. Another impressive example for this is Wikipedia, a multilingual, web-based, free-content encyclopedia project supported by the Wikimedia Foundation. Because of the intertwined networks involved in the creation and presentation of information in the WWW, the identification of relevant content has become increasingly difficult. Additional problems arise if the time evolution of content relevance shall be traced and if local and global relevance of content shall be distinguished.

Ranking vs. Relevance

Typically, ranking algorithms like (Google) PageRank [**?** 75] or the HITS algorithm [**?**] are used to calculate the relevance or importance of a given page (node) in the WWW. However, *relevance* is not an exactly defined term and cannot be measured in a unique procedure. According to [**?**] relevance can be assigned to a thing or to information named **A** in the context of a given task **T**. Only if **A** increases the probability of achieving the goal **G** of task **T**, **A** is relevant to **T**. Without a task and without a related goal, relevance does not exist. The identification of relevant information thus requires a context. In this chapter, we use the term 'relevance' in the same way as 'importance' or 'meaningful within a given context'. Measuring relevance of a node can be done according to (i) its intrinsic properties (e.g. text length of a WWW page), or (ii) the relative value of an intrinsic property (e.g., text length divided by the average text length of a group of related pages), or (iii) based on structural properties of one of the networks in which the considered node is embedded. PageRank is an example of a structure based ranking. The PageRank expresses the probability to find a random surfer in a given node [**?** **?**] and thus exploits mainly the structural properties of the network. Similarly, the HITS algorithm classifies a node either as a hub node (many outgoing links) or as an authority node (many incoming links) [**?**]. Both algorithms are applied to directed graphs and require a dataset describing the full graph. This is challenging for very large systems with billions of nodes.

Relevance is also a time related issue. Both, Page-Rank and Hits do not respect time. The T-Rank was introduced by Berberich *et al.* [**?**] to overcome this limitation. They define a temporal focus of attention (tfa), which can be a time range or even a single point in time. Based on this approach, they are able to distinguish different relevant topics as a function of time (see [**?**]). By using temporal weights as annotations on links and a modified Markov chain model they are able to handle page relevance in a time-aware manner without additional slicing of Web graph data.

A New Approach

In order to achieve a meaningful interpretation of analysis results all possible influencing factors have to be taken into account - which is impossible in practical applications. In socio-physics these factors are, e.g., demographical and ethnological influences such as the embedding into cultural contexts, or political and economical influences such as availability of technical infrastructure or even access restrictions regarding resources. Such aspects might also influence the data collection procedure and cause a hidden bias. Our new approach was developed to allow identification and qualitative interpretation of such hidden properties.

Besides this, multiple research disciplines look at different parts of the data set. In order to connect and compare results of diverse research projects, scientific methods for social network analysis require robust and flexible frameworks which enable and support interdisciplinary approaches. Therefore, we aim at a comparable measurement of a node's relevance within a local graph defined by the node's local neighborhood. Especially local link structures, article length, user activity, and editorial activity are considered. The key properties of the new method are:

- The local neighborhood defined by explicit links and implicit semantic annotations is examined.
- The context can be defined by a common language or by any other set of semantic concepts. This enables a connection to cultural aspects, related to regions and languages used by specific groups of people.
- The semantic relation between pages in different languages can be used to aggregate data related with a certain topic, e.g., for studies related to news, market data, (Twitter) messages, or communication in the context of large important events or movements in societies.

11.2. Definition of Representation Indexes

Here, we define and evaluate several parameters that measure the ratio of local representation with respect to global representation for selected topics. Our first approach is based on the numbers of articles (nodes) in each group, n_{LN}, n_{IWL}, and n_{GN}. Specifically, we define the *local representation index* for node degrees by

$$\mathrm{REP}_k = \frac{n_{\mathrm{LN}} + n_{\mathrm{IWL}}}{r_{\mathrm{GN}} + n_{\mathrm{IWL}}} = \frac{k_{\mathrm{CN}}}{\langle k_{\mathrm{IWL}} \rangle} \qquad \text{with} \quad r_{\mathrm{GN}} = \frac{n_{\mathrm{GN}}}{n_{\mathrm{IWL}}}. \tag{11.1}$$

Note that the nominator is identical with the so-called degree k_{CN} of the CN, while the denominator is the average degree $\langle k_{\mathrm{IWL}} \rangle$ of all nodes in IWL, i. e., the average degree of the node regarding the considered topic in all other languages. Note, that labels SCN and IWL are synonyms. In the context of Wikipedia analysis we prefer IWL like here but SCN allows a more generic application - this is why in the published article SCN was used instead.

In our second approach, we consider text lengths v instead of node degrees k. The total text volume per page and the average text volume per group are used as indicators of how well a certain topic is represented within a certain language. We assume that a topic is better represented in the language, in which it has a more comprehensive explanation. Specifically, for the total text volume v_{CN} of each CN and the total text volume v_{IWL} of all n_{IWL} other language versions, we define

$$\mathrm{REP}_v = v_{\mathrm{CN}} \frac{n_{\mathrm{IWL}}}{v_{\mathrm{IWL}}}, \tag{11.2}$$

since $v_{\mathrm{IWL}}/n_{\mathrm{IWL}}$ is the average text volume of all IWL pages.

Thirdly, we study *time-dependent local representation indexes* $\mathrm{REP}(t)$ based on the time series of the hourly rates of user accesses $a_i(t)$ or editorial events $e_i(t)$ for each CN and each node in the corresponding IWL groups. A high number of page views (or editorial changes) can indicate an increased interest of the user community. Although page view data are anonymous, it is possible to use the relationship between users and their preferred language to measure user interest per language. Specifically, for a time slice of width Δt beginning at $t = t_0$, we define

$$\mathrm{REP}_a(t_0) = \frac{n_{\mathrm{IWL}} \sum_{t=t_0}^{t_0+\Delta t-1} a_{\mathrm{CN}}(t)}{\sum_{i \in \mathrm{IWL}} \sum_{t=t_0}^{t_0+\Delta t-1} a_{\mathrm{IWL}i}(t)}, \tag{11.3}$$

where i runs over all nodes in the IWL group corresponding to the considered CN. An analogous definition is used for the editorial events to define $\mathrm{REP}_e(t_0)$. Clearly, these indexes will be large if there is more user-access activity or more editorial activity, respectively, regarding the CN compared with the averages in other languages.

11.3. Definition of Relevance Indexes

The local representation indexes (REP_s) are related to a given (or selected) semantic concept, expressed by a Wikipedia page (the CN) in a chosen language and all IWL pages. They indicate how well a topic is *represented* in a given language, irrespective of its embedding within contexts in this language. Only the core of the neighborhood network is considered (see figure 6.2 (a,b)). However, it turns out that text lengths, user-access activities, and editorial activities are hardly comparable across the different Wikipedias, i.e. across the language versions and cultures. Therefore, more meaningful results can be obtained if we divide by average quantities determined for articles within *the same* language community. Such indexes will characterize how *relevant* a topic is within the selected language or within the global context.

Specifically, we study the ratio of the parameters $\mathrm{L.REL}_v^{\mathrm{LN}} = v_{\mathrm{CN}} n_{\mathrm{LN}}/v_{\mathrm{LN}}$, representing the relevance of the CN in the chosen language, and $\mathrm{G.REL}_v^{\mathrm{GN}} = v_{\mathrm{IWL}} r_{\mathrm{GN}}/v_{\mathrm{GN}}$, representing the average relevance of all IWL pages within their combined contexts, i.e. the relevance of the selected topic within the other languages. The corresponding *relevance index* is thus defined as

$$\mathrm{REL}_v = \frac{\mathrm{L.REL}_v^{\mathrm{LN}}}{\mathrm{G.REL}_v^{\mathrm{GN}}} = \frac{v_{\mathrm{CN}} v_{\mathrm{GN}} n_{\mathrm{LN}}}{v_{\mathrm{LN}} r_{\mathrm{GN}} v_{\mathrm{IWL}}} = \frac{\frac{v_{\mathrm{CN}}}{1} \frac{v_{\mathrm{GN}}}{n_{\mathrm{GN}}}}{\frac{v_{\mathrm{IWL}}}{n_{\mathrm{IWL}}} \frac{v_{\mathrm{LN}}}{n_{\mathrm{LN}}}} = \mathrm{REP}_v \frac{\frac{v_{\mathrm{GN}}}{n_{\mathrm{GN}}}}{\frac{v_{\mathrm{LN}}}{n_{\mathrm{LN}}}}. \tag{11.4}$$

In addition, we compare local *time-dependent relevance indexes* $\mathrm{L.TRRI}(t)$ with corresponding global time-dependent relevance indexes $\mathrm{G.TRRI}(t)$ for user-access activity ($a(t)$) and for editorial activity ($e(t)$) respectively:

$$\mathrm{L.TRRI}_a(t_0) = \frac{n_{\mathrm{LN}} \sum_{t=t_0}^{t_0+\Delta t-1} a_{\mathrm{CN}}(t)}{\sum_{i \in \mathrm{LN}} \sum_{t=t_0}^{t_0+\Delta t-1} a_{\mathrm{LN}i}(t)} \qquad \text{and} \tag{11.5}$$

$$\mathrm{G.TRRI}_a(t_0) = \frac{n_{\mathrm{GN}} \sum_{i \in \mathrm{IWL}} \sum_{t=t_0}^{t_0+\Delta t-1} a_{\mathrm{IWL}i}(t)}{n_{\mathrm{IWL}} \sum_{i \in \mathrm{GN}} \sum_{t=t_0}^{t_0+\Delta t-1} a_{\mathrm{GN}i}(t)}. \tag{11.6}$$

The width Δt of the considered time slices must be optimized so that random fluctuations are damped while the temporal changes of the relevance indexes remain visible.

11.4. Evaluation and Interpretation

11.4.1. Static Representation Index for Networks of Linked Pages

Figures 11.1, 11.2, and 11.3 compare the values of the local representation index, REP_v, and the relevance index REL_v regarding text volumes v (Eqs. (11.2) and (11.4)) for selected sample data sets in double logarithmic plots. In addition, the local representation index REP_k regarding the node degrees k (Eq. (11.1)) is represented by the areas of the circles.

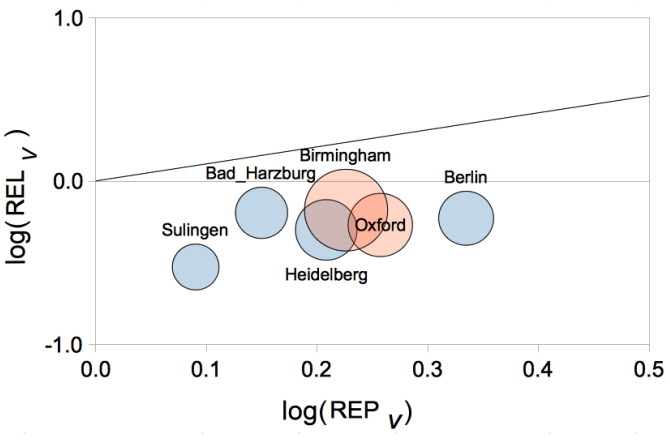

Figure 11.1.: **Comparison of three static representation measures.** X-axis shows $\log(\mathrm{REP_v})$ the related text volume based relevance measure $\log(\mathrm{REL_v})$ is plotted on y-axis. The area of the circles represents the $\mathrm{REP_k}$ value. In general, Wikipedia pages for German cities and UK cities show comparable $\mathrm{REP_v}$, $\mathrm{REL_v}$, and $\mathrm{REP_k}$ values. This indicates that pages from this specific category are equally represented in both languages.

For cities in Germany and the UK (see figure 11.1), one can see that Birmingham has clearly the largest REP_k, while all the others have a somewhat similar REP_k. In particular, REP_k for the most important city, Berlin, is smaller than the values for all other cities except the very small German cities of Sulingen and Bad Harzburg (which would not even qualify as cities by English standards). This first result for a homogeneous group of topics (cities) indicates, that the values of REP_k are *not* comparable across language versions, although they may be useful for estimating representation among articles of similar topics in the same language version.

The text-volume based representation index REP_v performs clearly better, see Fig. 11.1. The sequence obtained by ordering according to REP_v (Berlin - Oxford - Birmingham - Heidelberg - Bad Harzburg - Sulingen) reflects quite well the importance of the cities, also across languages, with Berlin and Birmingham having more than 3 and 1 million of inhabitants, respectively, and Berlin, Oxford, and Heidelberg having major universities. The

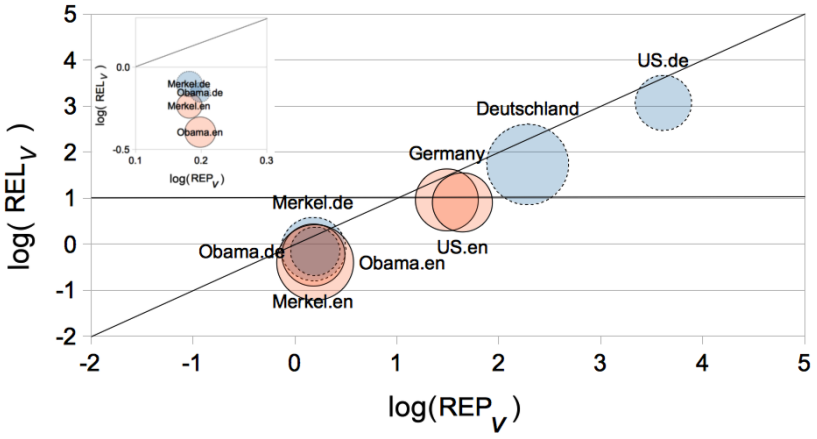

Figure 11.2.: **Comparison of three static** REP **measures.** X-axis shows the $\log(\mathrm{REP_v})$ and the related text volume based relevance measure $\log(\mathrm{REL_v})$ is plotted on y-axis. The area of the circles represents the $\mathrm{REP_k}$ value. The page for Angela Merkel (the Federal Chancellor of Germany) is equally represented in both languages. The page named Barack Obama is much better represented in English language than in the German Wikipedia sub project. This shows the influence of the selected linguistic context on the local representation index and allows a context dependent ranking for different terms within one language or a comparison between multiple languages.

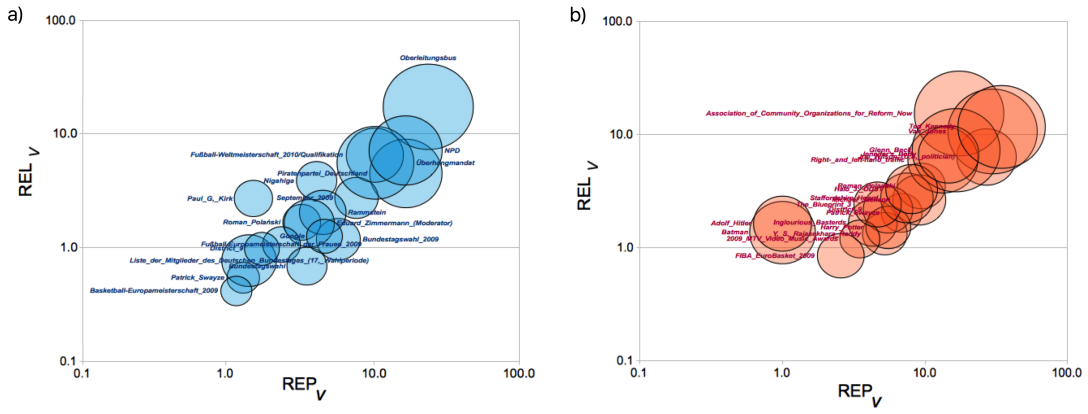

Figure 11.3.: **Relevance plot** for 20 most edited articles (09/2009) from a) German and b) English Wikipedia.

corresponding relevance index, REL_v, which was designed to make results better comparable across language versions, actually performs much worse – for example Bad Harzburg turns out to be above Berlin in this ranking.

How do the three static indexes perform when different kinds of topics are compared? To answer this question, we use figure 11.2). It shows that both text-volume based indexes, REP_v and REL_v, seem to distinguish between countries (large values) and persons (small values). In addition, both indexes are very large for the German versions of the articles regarding CNs 2.1 ('US') and 2.2 ('Germany') compared with the English versions, although one would have expected that the English versions are more important. Apparently, the articles on Germany and on the USA in the German Wikipedia are particularly long, more than 100 and 1000 times longer than those in other Wikipedias, respectively. This leads to the large values of REP_v. Simultaneously, both German articles are much longer than the articles in the corresponding local neighborhoods; this leads to the large values of REL_v. In English language, they are merely 30-50 times longer, which may be due to some material moved to sub-articles. These large differences between the German and English version of the articles regarding the countries are not justified and indicate that both text-volume based indexes cannot be used for classification on their own.

Interestingly, however, the articles regarding the two top politicians of both countries do not differ much in length between the language versions. They all thus have nearly identical local representation REP_v. However, the inset in Fig. 11.2 shows that the English versions of both articles have a slightly lower relevance REL_v – contrary to expectations. This suggests that REP_v is a better indicator for importance than REL_v, in agreement with the observations in dataset 1. All cities in dataset 1 are actually located in the same region of the REL_v−REP_v plot as the chancellors.

The local representation regarding the degree, REP_k, is not agreeing well with the expected importance of the CNs. REP_k shows that CN 2.2 ('Germany') has many more links, i.e., a much larger context, in its German version than in its English version, while the opposite holds for CN 2.3 ('Obama'). For CNs 2.1 ('US') and 2.4 ('Merkel') REP_k is similar in the German and the English versions of the article. REP_k can thus not be directly related with importance of the considered CN. Nevertheless, it is rather independent of the representation and relevance indexes regarding the text volume. The general trend is that text-volume based relevance and representation are approximately similar, $\mathrm{REL_v} \approx \mathrm{REP_v}$ with REL_v being a bit smaller in nearly all cases, see Fig. 11.2 in particular. However, if one looks at detail, REP_v seems to be much more indicative of a CN's importance than REL_v, see Figs. 11.1 and inset of Fig. 11.2 in particular. REP_k, on the other hand, is rather independent of the other two indexes, and its values are comparable only within a given kind of topics and within a language version.

The results for data sets 3 and 4 confirm this general picture, see Fig. 11.3. Furthermore, Fig. 11.3 shows that none of the three indexes, REP_k, REP_v or REL_v can distinguish between articles of rather local relevance (red circles) and articles of rather global relevance within the selected data sets. Different cultural aspects and differences between the structure and usage of languages might be considered as reasons for this.

11.4.2. Temporal Relevance Index for Time-resolved Relevance Analysis

Wikipedia is a highly dynamic system. The page networks and the related user networks grow and change their internal structure while new pages and links are added or existing pages are edited. During a single page's life cycle the article can be split into smaller, more specialized pages linked to the main page where the content originated from. Links also go to other related pages, which cover totally different topics.

Nevertheless, because the growth process of large Wikipedia projects is not very fast, after they reached a saturation (see Fig. 16.b in Schreck *et al.* [10]) the system can be handled like a static system on a weekly or even

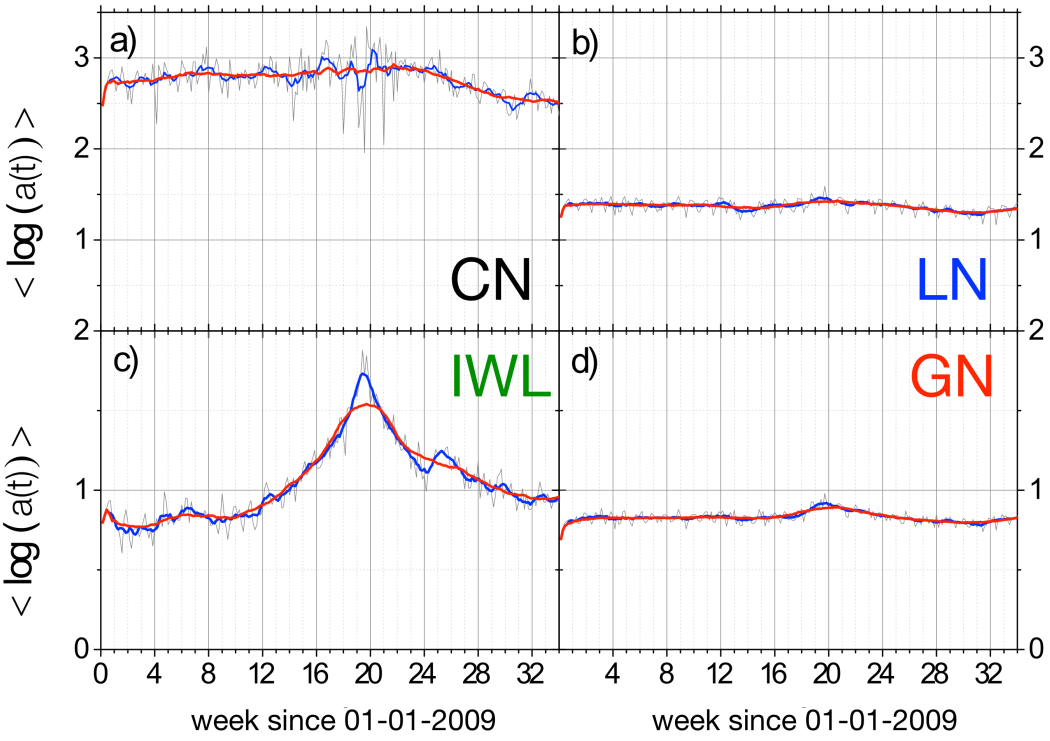

Figure 11.4.: **Raw data for time-dependent relevance measures.** The logarithm of access-rate time series is shown for the central node CN (Wikipedia page '*Illuminati (book)*' from German Wikipedia) in a). The averages of the logarithms of the access-rates for the local neighborhood LN are shown in b), for the inter-wiki linked pages IWL in c) and for the global neighborhood GN in d) in three time resolutions (daily in gray, weekly in blue, and monthly in red color) The color-coding of labels in large capital letters is related to the color of the curves in figures 11.5.a and 11.5.b.

monthly time scale. This allows a time-dependent analysis of the link structures and of the content of pages.

In contrast to this, user activity varies with daily cycles. According to [117] weekly patterns can also be found in Wikipedia edit-event time series. Daily and seasonal patterns in access time series have been reported in [10].

This study is focused on context sensitive temporal relevance. In previous studies the page groups have been selected based on the language, according to the total access activity of the pages, or depending on a classification as 'stationary' or 'non-stationary' access-rate time series. Now we select page groups dependent on their meaning. Our novel approach uses relative measures to eliminate hidden biases within the local neighborhood LN or the global neighborhood GN around a central node CN and all pages with inter-wiki links IWL to CN.

A classification of access-rate peaks in single time series was done to distinguish 'exogeneous' and 'endogeneous' bursts (see [165, 8]). Within a context network, such bursts or peaks, detected in one single access-time time series might be interpreted as a result of a sporadic information flow within the local network or as the source of such a flow. Because our goal is to study external influences and their origin in more detail, we take the context of a page represented by the neighborhood into account. We can study, if an access-rate peak has an influence on the pages in the neighborhood or if it was triggered by an increasing or dropping activity in the neighborhood. We thus want to analyze if an excitation is propagated through the network.

User interest or attention to a certain Wikipedia article can easily be measured by access-rates, which are usually calculated from aggregated raw data for a domain-specific appropriate time range. Fig. 11.4 compares daily, weekly and monthly acces-rates. During the last 3 months, the access-rate for CN decreases continuously. However, it cannot be determined from Fig.11.4.a alone, if this is an intrinsic property of the node or some external effect. Therefore, we use Figs. 11.4.b, 11.4.c, and 11.4.d to find a stable access-rate in the local and also in the global neighorhood of the selected page. The IWL pages show a much stronger access-rate peak, compared to the CN, with a maximum in week 19 in 2009. At the same time the variance of the access-rate measured for the CN is much higher. How strong the international influences are and if such even exist cannot be explained with individual time series without a context. By taking the network structure and the time series data into account, our approach uses a context sensitive comparison. This allows an explanation and even the separation of external and intrinsic features, found in the uni-variate time series.

For a direct comparison of access-rates for groups CN, IWL, LN, and GN the averages of the logarithms of access-rates are plotted into one chart (see fig.11.5). Fig.11.5a indicates an exogenous increase of the access-rate

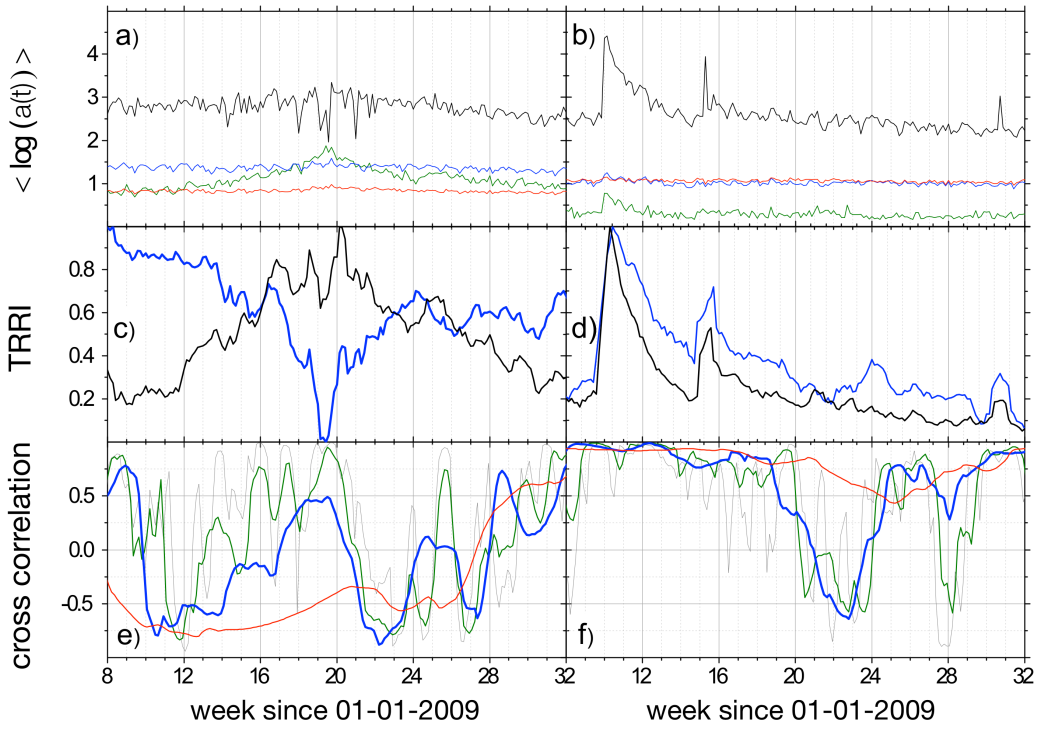

Figure 11.5.: **Comparison of local and global REL measures for two selected topics.** Selected topics are (a,c,e) '*Illuminati (Buch)*' and (b,d,f) '*Amoklauf von Erfurt*'. The first row shows daily access-rate data for CN (black), IWL (green), LN (blue), and GN (red) like those presented in figure 11.4. The second row shows L.TRRI$_a$ (blue) and G.TRRI$_a$ (black). (e,f) The cross-correlation between L.TRRI and G.TRRI for sliding windows of length 24 hours (gray), 7 days (green), 2 weeks (blue), and 3 months (red) are shown in the bottom row.

for all languages except German, while for the German page the variance of the access-rate increases. Fig.11.5b shows endogenous bursts in week 9 for the CN node and also for the IWL group but no strong change in the average access-rates for both neighborhoods. This is a clear indicator for an exogenous burst, which is not influenced by a flow of attention coming from other topics in the neighborhood (unlike the case in figure 11.5.a).

Beside this qualitative interpretation we use the relevance index, defined in equations (11.5) and (11.6), to compare the local and global relevance of a topic in a quantitative way. Fig. 11.5.c shows an increase of the global relevance, while the local relevance drops at the same time. This indicates, that in a local context, the topic does not attract that much attention compared to the global context. The decreasing value of the local relevance index also indicates an increasing interest in the local neighborhood. Fig. 11.5.d shows two peaks in the local and the global relevance index at the same time which supports the classification as an exogeneous burst according to [165]. Fig. 11.5.e and Fig. 11.5.f show the cross-correlation for the two functions from Fig. 11.5.c and Fig. 11.5.d respectively. Higher correlation values can be found in Fig. 11.5.f, while the trends of both functions are comparable. Especially the presence of bursts in both indices yields strong correlations. On the other hand, low correlation values indicate different trends in both indices. They allow to conclude, that there are also differences between the local and global relevance of the selected topic.

Language Dependent Interpretation of TRRI

Figure 11.7 shows a comparison of the global and local relevance index for three selected Wikipedia pages to illustrate the influence of the selection of language. In case of the companies Oracle and Capgemini we find a stable value while in case of the Apache Hadoop page a significant increase in March 2011 is visible. The local relevance index starts to dominate over the global relevance after it exceeds the threshold $log($ L.TRRI $) > 1.0$. after this point in time the Hadoop project attracts more user traffic than the pages in the neighborhood.

Finally, we show an example for the strong language dependency in figure 11.6.b. The page about the French company Capgemini illustrates the impact of a wrong selection of the lingual context. The relevance index for English pages is relatively low, while for international pages, especially for French pages a high TRRI was measured. One can clearly conclude, that the context has to be adjusted according to the topic, otherwise an unknown bias still exists. The best context selection is given if the difference between local and global relevance index is maximized. The representation plot is used as a tool to evaluate and validate results from TRRI plots, especially in the context

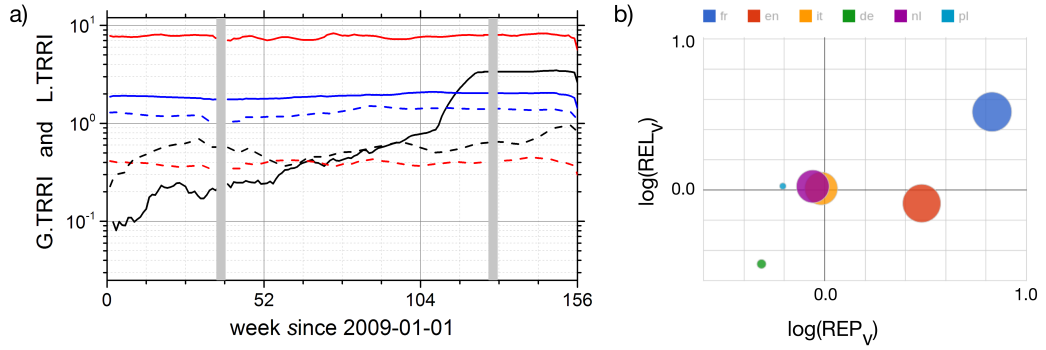

Figure 11.6.: **Influence of CN language.** (a) shows the contextual time-resolved relevance index (TRRI) for Wikipedia pages about companies Oracle (red), Capgemini (blue), and the open source software project Apache Hadoop (black) between January 2009 and December 2011 with weekly resolution. Representation and relevance index depend on the selection of CN. (b) shows the representation index for one Wikipedia page for 6 different languages with the highest representation in French language. According to (b) one can still see a high relevance, based on user access to English and French pages for the company Capgemini. In this case the local context can be defined as a hybrid context by two central nodes - French language because of the country of origin, and English language because of international IT business. If only one language is used, as in (a), one can not clearly differentiate between local (straight line) an global relevance (dashed line) as in the case of the pages for Oracle (red), and Apache Hadoop (black).

of multilingual content networks.

11.5. Conclusion

Computed representation and relevance index values can be used as ranking indicators, .e.g, within a set of given Wikipedia articles without the need of processing the whole Wikipedia text corpus. Static REP index (and in some cases also the static REL index) values can be useful weights for ranking and filtering the results from search engines. Search engines (like e.g., Apache Solr, Elasticsearch) could use those values to enrich search results consisting of several hundreds of documents in near real time if the weights are pre-calculated. Because this computation requires only partial information, not the full page graph (unlike in the case of the PageRank algorithm), it can easily be parallelized. In general, a calibration is need, to identify the sensibility of the individual measures in a particular context (see figures 11.1, 11.2, and 11.3).

Beyond a direct comparison of local and global TRRI - see 11.5(c,d) - also, the cross-correlation between $\mathrm{L.TRRI_a}$ and $\mathrm{G.TRRI_a}$ as shown in figure 11.5.(e,f) can be used as another measure to study information flow in context networks. In case of high correlation, the core and the neighborhood network behavior are comparable. Different properties or different dynamics can be assumed if the correlation is close to zero or even negative.

Therefore I suggest to use a threshold $t_s = 0.5$ to define a new property of a local context network, the *context polarization*. Context polarization is zero if the cross-correlation between $\mathrm{L.TRRI_a}$ and $\mathrm{G.TRRI_a}$ is in the range above $-t_s$ and below $+t_s$. In case of a correlations higher than t_s the core is *aligned* with the context and in case of a negative correlation with values less than $-t_s$ core and neighborhood are *complementary* to each other.

12. Structure Induced Stress: A Measure to Quantify the Impact of Functional Networks

> Headlines, in a way, are what mislead you because bad news is a headline, and gradual improvement is not.
>
> *(Bill Gates)*

According to [95] the analogy to real world systems and nature can be used in order to draw esthetical images of complex networks. If it looks like something known, often it appears also plausible. One has to be careful. Such analogies do not replace the mathematical proof nor a clear theory.

In this section I suggest a novel concept which allows a quantitative network layer comparison and a comparison of dynamic processes if they are represented as graphs.

Structure Induces Stress (SIS) is based on the idea of elements of an n-body system connected by elastic forces (springs). In a dynamic equilibrium the nodes have well defined positions. Different properties of the system components, such as node properties, edge properties, and also the link structure influence this equilibrium, and different conditions lead to different positions of the nodes.

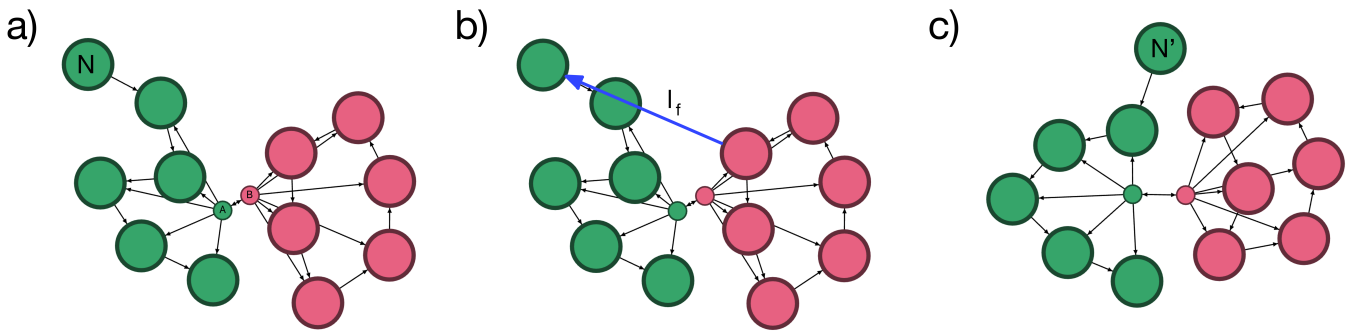

Figure 12.1.: **Structure Induced Stress (SIS)**. Functional links (see blue link l_f) can cause stress on nodes in static network layers (see black links). The result is an additional displacement of the node ($N \Rightarrow N'$), which can influence the underlying structure as a consequence. Calculated SIS vectors allow a quantitative interpretation of the impact of functional layers on the static structure, induced by dynamic process on top of the network. In this way we can also studied how different functional networks interfere with each other by comparing the impact of multiple layers or even combinations.

Also, two networks with different structure lead to different layouts. Each node's displacement is a result of the influence of the entire network. Thus, we can interpret the displacement as the result of the impact of a structural change. We do not measure the different forces, since they are inherent properties of the processes which are compared to each other. Rather, we measure the consequences of the different influences on the network in total. This allows us to express the difference in the process using forces which are not comparable in a direct way. The reasons are (a) there is no real force, and (b) the characteristic of the process which leads to a certain placement of nodes can be very different. For example, the static link network and the access activity network are defined by very different concepts.

We are interested in multi-layer networks, in which layers represent different aspects. Generally speaking, we can now measure the impact of one aspect on another one, based on a transformation of node positions within a plane. Because no direct mathematical expressions and no simple property exists to describe the impact of one network layer on another, such a transformation is helpful. This transformation is achieved by a graph layout algorithm, such as the force directed layout. This approach allows us to embed the network layers in a consistent way. Based on this we define a structural measure for process comparison. A comparable rather simple approach is called spring layout. The attracting forces between nodes are based on Hooke's law. M. Geipel [209] published a new, more generic graph layout algorithm called ARF for *'attractive and repulsive forces'*. This algorithm relies on the balancing of two antagonistic forces which is comparable with the approach, developed by Fruchterman and Reingold [95].

Even if a spatial embedding of graphs is not generally possible, the analogy of forces is often used. Examples are the *'Social Force Model'* by Helbing and Molnár and the *'Index Cohesive Force'* by Kenett *et al.* [210]. The later inspired this work and is the base for a more generic approach which is not specific to financial markets. It can be used in the context of arbitrary semantic networks and is called: *'Context Cohesive Force'* (see equation 9.14). The initial idea in this work was to define a functional network from a similarity measure. We came up with several representations of such networks, depending on the selected layout and filter algorithms. From node positions we could not learn much - but relative positions and the overall pattern - which appeared over time - we could already derive information which are also reflected by quantitative measures, such as clustering and modularity. The following question remained: *How can two networks be compared with each other?* An absolute measure would be required. Without a natural embedding into the same environment a calibration and a direct comparison is not possible.

Our novel approach uses only local properties. We quantify, to what extend the functional network "disturbs" the structural network underneath. One can also see it the over way round. Therefore we calculate the graph layout as described by Fruchterman and Reingold [95]. The Fruchterman Reingold layout algorithm is available in many software packages.

The algorithm works in two phases:

- (1) The initial layout based on structural links is calculated. Node positions are not absolute and do not contain information alone. Functional links do no contribute to this initial layout.
- (2) The second phase starts as soon as the algorithm converges and an initial layout exists.

From now on, also the functional links contribute to the placement of nodes. In case of two equal networks which perfectly overlap, we expect no differences between results from phase 1 and phase 2. If the functional network is not well aligned with the structural network, it influences the node positions already in the next layout step. We calculate the influence for one more time step and interpret the displacement of a particular node, or the average displacement as the influence of the functional network on the underlying structure.

This procedure is somehow comparable to molecular dynamics simulations. The core difference is the absence of coordinates in real space. Network nodes have by default no position - at least in most cases - even if geographical locations are known, this kind of data would not contribute new information. As a consequence we calculate the displacement of the positioned nodes on a plane. The initial positions are the results of a transformation of the raw data, e.g., a layout procedure, which uses the concepts of attracting and repelling forces, such as the one described by Fruchterman and Reingold.

Displacement vectors are defined by the initial location and end at the point where the node is moved to in phase 2, under the influence of a functional network. For all nodes - or just for groups (clusters, partitions) - one can average the displacement vector lengths. A quantitative comparison of different systems is now possible, even if they represent different real world systems for which measured data is not comparable in a direct way.

Figure 12.1 shows a simple social network with a formal organizational structure and an informal communication network on top as example. Both of the two clusters have very central nodes (A and B). The ground-state of the system is defined by the structural network shown in (a) which was used for the initial layout calculation. No *official* link exist between the nodes around B and the nodes around A. The displacement of node N is caused by an additional functional link, see blue link in (b). This link represents an *in-official* communication activity, which forms a functional network and leads to a short cut in this case. This informal communication network is not in line with the underlying organizational network and can cause stress on node N, which gives the method the name. However, based on this measure we cannot conclude if the impact is negative - which means if it disturbs the function of the overall system, or if it is helpful and improves the system functionality. Comparison with specific target variables and their time evolution are the key in this case.

Part III.

Applications and Results

As mentioned before, this work was inspired by several projects from different scientific disciplines. In this part the focus is on applications of the previously introduced measurement techniques and on the interpretation of first results. Additional studies were published in several journals (see publication list) and are partially reproduced in part IV (appendix).

13. Dynamics of the Complex System Wikipedia

> Our scientific power has outrun our spiritual power. We have guided missiles and misguided men.
>
> *(Martin Luther King, Jr.)*

Internet-based social networks often reflect extreme events in nature and society by drastic increases in user activity. We study and compare the dynamics of the two major complex processes necessary for information spread via the online encyclopedia 'Wikipedia', i. e., article editing (information upload) and article access (information viewing) based on article edit-event time series and (hourly) user access-rate time series for all articles. Daily and weekly activity patterns occur in addition to fluctuations and bursting activity. The bursts (i. e., significant increases in activity for an extended period of time) are characterized by a power-law distribution of durations of increases and decreases. For describing the recurrence and clustering of bursts we investigate the statistics of the return intervals between them. We find stretched exponential distributions of return intervals in access-rate time series, while edit-event time series yield simple exponential distributions. To characterize the fluctuation behavior we apply detrended fluctuation analysis (DFA), finding that most article access-rate time series are characterized by strong long-term correlations with fluctuation exponents $\alpha \approx 0.9$. The results indicate significant differences in the dynamics of information upload and access and help in understanding the complex process of collecting, processing, validating, and distributing information in self-organized social networks.

We analyzed Wikipedia access-rate and edit-event time series regarding the long-term autocorrelations. As will be shown below, detrended fluctuation analysis (DFA) and return interval statistics (RIS) show long-term memory effects clearly for access-rate time series but not for the edit-event time series. This allows us to conclude that the collective editorial and the information consumption processes in Wikipedia are fundamentally different. One can assume a more balanced system during an early stage, when the majority of users are also contributors. At a later stage, when more and more people get attracted by the system a different situation emerges. Many more consumers per contributor can be observed. Contributions to the system and usage are not balanced any more. At the same time the following questions arise: Are the processes of contribution and consumption of information always (also in other SMAs) asymmetric like here? How are process properties related to system size? Being able to understand the process of information contribution and consumption means also to be able to identify life cycle phases of a complex system.

This chapter covers two types of analysis: (1) univariate time series analysis followed by (2) bi-variate time series analysis which is a special case of multivariate analysis.

Univariate Analysis: Multiple approaches allow an inspection of system properties based on individual node's properties. Models for single node activity such as on a single web resource, like the Poisson process or Hawkes process do not include any of the additional structural information available in the neighborhood.

Ratkiewicz *et.al.* [211] explain the fat tails of the typical distributions of dynamic properties on a macroscopic level. They provide a quantitative, large-scale, temporal analysis of the dynamics of online content popularity in two massive model systems: the Wikipedia and an entire country's Web space. Crane and Sornette [165] provide a microscopic view into the behavior of a single resources (YouTube videos in this case). Such videos are exposed to the social community of interconnected YouTube users. This user community is also a complex system on its own like Wikipedia's user community.

The fundamental model to describe each node's access-activity is the so called Hawkes process according to Crane and Sornette [165]. The Hawkes process describes the deviation from a simple Poisson process. Mitchell

and Cates [212] used computer simulations to study the Hawkes process systematically. Their results show three classes of decay exponents as proposed by Crane *et. al.* [165]. Crane *et al.* call the identified classes universal, but according to the numerical results provided by [212] there exist different distributions. This seems to limit the universality of the Hawkes-based analysis. For Wikipedia acces-rates we could find many extreme events with different shapes but we could not find the proposed universal classes (see also section 13.1.2 and figure 13.3).

Multivariate Analysis: A model for a particular property of one specific object can be enriched by related values, such as the average group activity or topological properties of the ensemble. This is why in this work a framework for reconstruction of functional networks was created and evaluated. We adopted existing methods to investigate the interaction within the local and global neighborhood. In this way, this work provides novel tools needed to improve or modify existing analysis processes and theoretical models. Such analysis is helpful in order to understand relevance and attention to online content in a highly dynamic environment of social media applications.

Recently the phrase *attention economy* became very popular. How much attention people give to things, ideas, aspects of personal life or society is related to how much time they spend in or on the respective context. Since time is limited, higher attention in one area leads to lower attention somewhere else. Such changes in attention follow recurring patterns and contain extreme events such as exogenous and endogenous bursts. Deviations from simple random models like the Poisson process are caused by collective phenomena, also called *herding* [213, 214]. How can the level of attention and changes in attention to resources, which have usually very different properties - following power-law distributions - be measured and predicted? Our approach is based on web server logs, available as access-rate time series. These time series exhibit bursts of different kinds.

The article by Ratkiewicz *et al.* [211] describes a procedure for quantitative popularity analysis of online content, such as Wikipedia pages. They find that popularity dynamics of, e.g., Wikipedia pages are characterized by bursts. This leads to characteristic features like fat-tailed distributions of inter event times and characteristic measures of properties like traffic, in-degree, and page size. Typically, articles have a burst in the early phases of the page life-cycle, while fluctuations are observed later on. Ratkiewicz *et al.* state, that articles have more edits if they have more in-links. We have found, that a higher access-activity comes with a higher edit-activity, (see figures 3.b and 3.c in [10]).

Crane *et al.* [165] have found, that activity bursts of access-activity to videos on YouTube can be described as a Poisson process (for up to 90% of the videos). For the remaining huge amount of videos (10%) they identify power-law relaxation quotients which can be split into three categories. They state, that this is consistent with an epidemic model, in which a power-law distribution of inter-event times (or waiting times) and epidemic cascades coexist. Epidemic bursts are the cause of future actions, which could again be bursts, forming clusters of extreme events.

Yu *et al.* [215] proposed a phase representation of YouTube video popularity. They found that many videos go through multiple stages of popularity which can move towards both directions (increase and decrease). The duration can be up to several months as shown on figure 1 in [215]. Such a categorization into phases allows them to reduce the average prediction error for future view count gain significantly compared to state of the art methods. This is also an example for adding information to an existing model in order to increase accuracy based on phases as a kind of structural information about the system.

.

13.1. Bursts and Fluctuations in Wikipedia Access-Rate and Edit-Event Data

This section is a reproduction of an article which was written in the context of the SOCIONICAL project and published by Mirko Kämpf, Sebastian Tismer, Jan W. Kantelhardt, and Lev Muchnik in the Journal Physica A (see [8] or [5] in my publication list). This research project was partially supported by the European Union within the FP7 project SOCIONICAL, No. 231288.

13.1.1. Introduction

Human interaction and information spread via social networks on the internet [216, 104] is becoming of increasing importance for our contemporary technological society. A number of recent publications highlighted the significant role of scaling laws [217, 218] in the emerging network structure [219, 176, 220, 221, 222, 223, 224] and dynamical interaction patterns [218, 225, 226, 211]. Other recent work focused on modeling information flow within social on-line networks [216, 227, 228]. Broadly speaking, such networks can be categorized into communication-centered networks, where the explicit social network is used to convey direct messages between users (e. g., email, Twitter), and content-based networks (e. g., the WWW, blogs, Wikipedia), where the social network implicitly represents the overlap in interests or activities performed by the individuals. Among the specific sites, Wikipedia is particularly relevant since in contrast to many others it can be expected to be a persistent phenomenon. Recent research regarding the open on-line encyclopedia Wikipedia focused on its network structure [110, 229, 101, 230] and its

development (e. g., article life cycles) [174, 231] as well as the popularity of its articles [211]. Other recent work analyzed the structures of the social communities of Wikipedia editors by considering the collaborative usage of the network content [232]. The complexity of Wiki networks and their dynamics clearly call for methods from complexity science, including scaling analysis. Here, we study the dynamics of usage and modification of the Wikipedia content, looking in particular at the scaling behavior of fluctuations and extreme events.

Since Wikipedia is a collaborative and open project, it represent the range of the momentary interests of the population of its users. In fact, a number of researchers suggested that a significant amount of activity performed by the Wikipedia contributors is motivated by news and mass media [233, 234, 235]. It is therefore reasonable to expect that significant extended extreme events, such as bursts with large peak amplitudes in user access of specific Wikipedia articles or increased editorial activity of particular topics, may reflect extreme events in nature and society. For instance, the sudden access-rate peak shown in Fig. 13.1(e) for an article on the shooting rampage in Erfurt (Germany) is caused by another shooting rampage. In addition to such exogenously caused extremes, there are endogenous extreme events building up within the evolving network, e. g., by addition of links between articles [211], or by gradual shifts of public interests. Recently, significant effort has been made to distinguish these two classes of extreme events in observational data [236, 237, 238, 239, 165] and to develop and test corresponding models [236, 240, 241, 242, 243]. A number of empirical studies focused on Amazon book and music sales dynamics [236, 237, 238], capital fluxes [239] and videos on YouTube [165].

Here we define bursts as drastic increases in activity for an extended period of time (at least several hours) in Wikipedia access statistics. We show that their frequency scales as a power law of the duration of the burst. However, the classification of the individual bursts does not allow a clear separation of endogenously and exogenously caused extremes, since the shapes of both sides of the bursts can be similarly well fit by power laws and exponentials (see also [238]). Furthermore, we show that the shapes of the bursts are nearly independent of their peak amplitudes. These results indicate that features of both, exogenous and endogenous processes as well as article changes due to edit processes are relevant for most Wikipedia access-rate bursts.

In the last years, the scaling behavior of fluctuations [217, 244] has emerged as a useful means for classifying time series that either characterize complex systems during different regimes or stages [32, 245, 246, 247, 248] or at different locations [249], or distinguish sub-components of complex systems [248, 250]. The cited exemplary studies analyzing spectroscopy data [245], geophysical data [249, 250], medical data [32, 248], road traffic [247], and internet traffic [246] are employing the detrended fluctuation analysis (DFA) method invented by Peng *et al.* in the context of base-pair sequence analysis in the DNA [28]. The method is based on random walk theory and was generalized for higher order detrending [32], multifractal analysis [251, 29, 30], data with more than one dimension [254], and cross-correlation analysis [255]. The properties of the DFA were explored in many articles [29, 30, 31]. In addition, several comparisons of DFA with other methods for stationary and non-stationary time-series analysis have been published, see, e. g. [258]. Here we apply DFA to the Wikipedia page access-rate time series and show that they are characterized by strong long-term correlations (i. e., a slow power-law decay of the auto-correlation function). Specifically, we find a rather universal scaling behavior with fluctuation exponents $\alpha \approx 0.9$ (corresponding to correlation exponents $\gamma \approx 0.2$) rather irrespective of the stationarity, i. e., the sizes of bursts occurring in the data.

One of the natural consequences of such long-term correlations is a clustering of extreme events [22, 23, 24, 126, 26, 27], i. e., a clustering of bursts with large peak amplitudes. The distribution density of the return intervals between extreme events can be well approximated by a stretched exponential for large return intervals [23, 24] and by a power law for small return intervals [126, 26], although finite-size effects are important and the description is not exact [126, 27]. For multifractal data without linear correlations, only the power-law regime occurs [261, 262]. Most importantly, the shape of the scaled distribution is practically independent of the threshold for extreme events. In addition, the mean residual time to the next extreme event increases with the elapsed time and depends strongly on the previous return intervals – a finding that makes return interval statistics useful for risk estimation and prediction [23]. Applications of return interval statistics to paleo-climate data [23], wind speeds [263], financial market data [264, 265, 266], medical data [267], telecommunication data [268], internet traffic [269], and times of wars in China [270] have been considered; note in particular the usability of the approach for online prediction of cardiac disorders [267]. Here we apply the method to Wikipedia access-rate and edit-event time series and show that the scaling stretched exponential ansatz holds irrespective of the threshold for the peaks of bursts in the access-rate data.

However, contrary to access-rate time series, the statistics of edit-event time series are characterized by the absence of clustering on all time scales except for next neighbor clustering. This surprising observation shows that the two processes, i. e., article edits (information upload) and article access (information viewing) have a rather different nature even though they are not independent. We suggest that Wikipedia page-access dynamics is mainly driven by exogenous events such as media or calendar dates or by gradual shifts of public interests, both of which may result in system-wide synchronization. On the other hand, Wikipedia edit patterns arise, at least partially, from the direct or implicit interaction between Wikipedia contributors.

The next section describes the Wikipedia databases and the required data preprocessing for this particular study, followed by a section which is devoted to the characterization of the access-rate peaks in relation to the recently

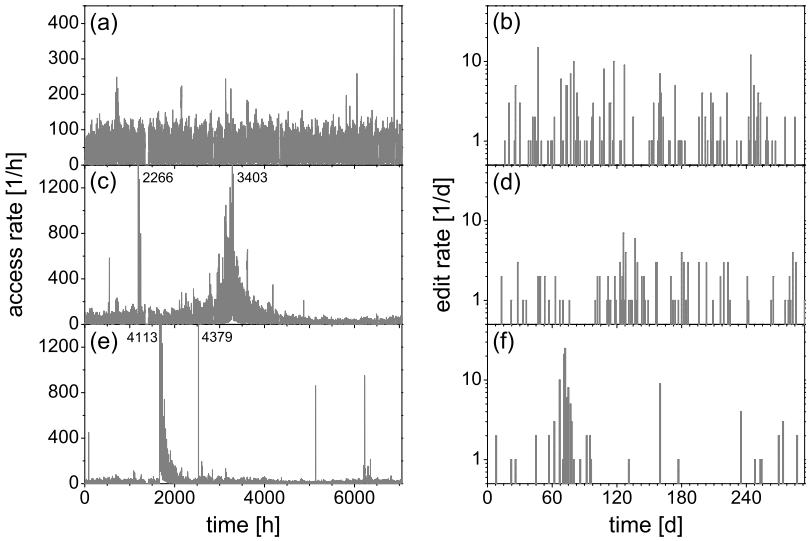

Figure 13.1.: **Examples of Wikipedia (a,c,e) access-rate and (b,d,f) edit-rate time series** for three selected articles with (a,b) rather stationary access rates (topic 'Illuminati (book)'), (c,d) an apparently endogenous burst of activity (peak on May 7, 2009, topic 'Heidelberg'), and (e,f) an exogenous burst of activity (topic 'Amoklauf Erfurt' (shooting rampage); it peaks on March 11, 2009, as another shooting rampage occurred in Winneden). The left parts show the complete hourly access-rate time series (from January 1, 2009, till October 21, 2009; i. e. for 42 weeks = 294 days = 7056 hours) with numbers in the plot giving the height of peaks truncated to show baseline fluctuations. The gap around $t = 1200$h is a systematical disruption and was found in all records. Parts (b,d,f) show daily edit-rate data for the same three representative articles, which were edited 270, 163, and 157 times in total, respectively, during the recording period.

suggested distinction between exogenous and endogenous extreme events. Than we apply, respectively, DFA and return intervals statistics to characterize the dynamics and to quantify long-term correlations of access-rate and edit-event time series.

13.1.2. Databases and Data Preprocessing

The dataset used in this work is a combination of two distinct databases. Both are open and freely available for download, however, analysis of these enormously sized data bases (see also [211, 173]) is quite challenging. Although there is some particular software for statistical analyses in wiki systems (WikiXRay [271] and WikiStatistics [272]) we have implemented our own analysis tools using C++, MySQL, MATLAB, and JAVA. First we study hourly access-rate (page-view count) data from every Wikipedia article (keyword) worldwide, i. e., including the over 260 languages. This data [273] was collected by a system monitoring and logging individual pages' access statistics. All data recordings began on January 1, 2009, and ended on October 21, 2009. For each article (with index $j = 1, \ldots, 16\,900\,000$) the number of accesses is given with an hourly time resolution, yielding access-rate time series $n^j(t), t = 0, \ldots, 7055$. The vast majority of the Wikipedia articles are relatively rarely accessed and do not experience any significant bursts of activity. In this paper we focus on Wikipedia articles j, that exhibit significant access rates of $n^j(t) \geq 256$ at least in one hour in the observed interval. This subset consists of 28 952 articles ($\approx 0.17\%$ of 16 900 000) in total. Figures 13.1(a,c,e) show three representative time series of hourly access-rate data $n^j(t)$. One can distinguish several types of variations: stationary fluctuations [Fig. 13.1(a)] with or without correlations, an apparently 'endogenous' burst [Fig. 13.1(c)] with significant precursory activity, and an 'exogenous' burst [Fig. 13.1(e)] marking the dynamic response to a major event.

In addition to random fluctuations and bursting activity, we observe daily and weakly activity patterns in the access rates for most Wikipedia articles. Periodic minima correspond to the night-time hours, although their positions depend on the time zone of the readers and thus on the language of the articles (German in the considered examples). Recurring weekly patterns present in many articles may also disrupt the following analysis. In order to minimize these cycles, we normalize the original data by dividing each observation $n^j(t)$ by the average for the corresponding time of the day $\langle n^j(t') \rangle_{\text{weeks}}$. This average is calculated for each hour t' of the $24 \times 7 = 168$ hours within a week.

$$x^j(t) = \frac{n^j(t)}{\langle n^j(t \bmod 168) \rangle_{\text{weeks}}} \qquad \text{with} \quad \langle n^j(t') \rangle_{\text{weeks}} = \frac{1}{42} \sum_{i=0}^{41} n^j(168i + t'). \tag{13.1}$$

Along with removing the recurring patterns, this transformation normalizes each data series so that $\langle x^j \rangle = 1$ for

each article j at a given time.

In addition to the access-rate data, we study the time series of the article edit events. We have collected this data from our second (complementary) database [274], which is a dump of the entire history of the Wikipedia project and contains every edit ever performed on every page in over 260 languages. It logs all edit events with time stamp, article length after the re-edition, and anonymous identification of the editor. The dump is performed periodically by the Wikimedia Foundation Inc. and is provided under the GNU Free Documentation License (GFDL). We processed this data and collected, for each article, the timeline of its access activity and its edits by individual Wikipedia contributors. Figures 13.1(b,d,f) show the edit time series of the three representative articles. In this case, the total number of article edits is displayed for each day. In the paper, we denote the number of edits per hour by $\epsilon^j(t), t = 0, \ldots, 7055$. Note that the edit-rate shown in Fig. 13.1(f) peaks on the same day as the access rate of this article (see Fig. 13.1(e)), but there is no such relation for the other two articles.

13.1.3. Characterization of Access-Rate Peaks

In this section, we study the properties of the activity bursts observed in the hourly access-rate data. We want to find out if it is possible to distinguish the classes of endogenously and exogenously caused extreme events and if corresponding models (see introduction) also apply to Wikipedia data. We define a burst at time t^* as an event for which $x^j(t^*) > 2\langle x^j(t)\rangle = 2$ (note that $\langle x^j(t)\rangle = 1$ due to the normalization in Eq. (13.1)).

Next, we determine the width of the bursts by finding the times $t_1 < t^*$ and $t_2 > t^*$ when the access rate first dropped by more than $\exp(-1)$ from its local maximal value at t^*, i. e., $x^j(t_{1,2}) < x^j(t^*) - (1/e)x^j(t^*) = (1 - 1/e)x^j(t^*)$. We do not allow bursts to overlap. In other words, for each burst interval $t_1 < t^* < t_2$, only the one position t^* with maximal access rate is accepted as burst position. We thus avoid a multiple detection of the same burst with different center positions t^*. The bursts are not necessarily symmetric, characterized by widths $t_{\text{before}} = t^* - t_1$ and $t_{\text{after}} = t_2 - t^*$. This definition is very sensitive to noise and yields 6 910 285 data points that we considered as bursts from the total number of 28 952 time series. To filter out the noise and retain only reasonable bursts, we impose a minimal duration requirement for t_{before} and t_{after} and increased thresholds for the peak amplitude in the following.

Table 13.1 reports the numbers of activity bursts with certain minimal durations before and after the maximum. Figure 13.2(a), related with the data reported in the third column of Table 13.1, shows the number of bursts of different durations before and after the peak. The functional forms can roughly be fitted by power laws

$$N(t_{\text{before}}) \propto t_{\text{before}}^{-\vartheta_{\text{before}}} \quad \text{and} \quad N(t_{\text{after}}) \propto t_{\text{after}}^{-\vartheta_{\text{after}}} \tag{13.2}$$

with effective exponents $\vartheta_{\text{before}}$ and ϑ_{after} around 2 for short bursts (< 10h) and around 3 for long bursts (> 48h). For a cumulated probability distribution ϑ would be larger by 1. Long decays occur slightly more frequently than long increases, in particular for the large peaks (see also Table 13.1). Except for this distinction, the results are virtually independent of the peak amplitude: the curves for maxima exceeding the threshold by factors of 2, 5 and 15 in Fig. 13.2(a) are parallel and the fractions for maxima exceeding the averages by factors of 2 and 10 in Table 13.1 are comparable (although large peaks 10 times above average occur much less often). The average behavior of large and intermediate bursts is therefore very similar.

burst range	factor	number	fraction	good exp.	good PL	$\Theta > 1/2$ (good)	$\Theta > 1/2$ (all)
$t_{\text{before}} \geq$3h	> 2	1292438	18.7%	10.0%	13.9%	34.8%	87.8%
$t_{\text{after}} \geq$3h	> 2	1382714	20.0%	11.1%	14.1%	34.0%	87.6%
$t_{\text{before}} \geq$12h	> 2	229097	3.3%	1.2%	1.4%	0.8%	22.9%
$t_{\text{after}} \geq$12h	> 2	267085	3.9%	1.8%	1.6%	0.8%	23.9%
$t_{\text{before}} \geq$2d	> 2	18423	0.27%	0.30%	0.16%	0.0%	0.0%
$t_{\text{after}} \geq$2d	> 2	23127	0.33%	1.27%	0.29%	0.0%	0.0%
$t_{\text{before}} \geq$3h	> 10	41115	11.1%	21.5%	20.3%	39.0%	81.9%
$t_{\text{after}} \geq$3h	> 10	63770	17.0%	24.2%	23.9%	38.5%	82.7%
$t_{\text{before}} \geq$12h	> 10	8759	2.3%	7.3%	5.4%	2.8%	23.6%
$t_{\text{after}} \geq$12h	> 10	18054	4.8%	13.2%	8.1%	3.6%	27.2%
$t_{\text{before}} \geq$2d	> 10	563	0.15%	5.5%	3.0%	0.0%	0.0%
$t_{\text{after}} \geq$2d	> 10	1765	0.47%	13.0%	1.9%	0.0%	0.0%

Table 13.1.: **Statistics of Wikipedia access-rate bursts**, reporting the number of bursts with widths of at least 3h, 12h, and 2d before and after the peak for peaks at least a factor of 2 (or 10) above the average. The fractions of good exponential and good power-law (PL) fits (correlation coefficient $r > 0.9$) for the increases and decays are also given as well as the fractions of peaks characterized by exponents $\Theta > 1/2$ among good fits and among all fits, see also Eq. (13.4) and Fig. 13.3.

Figure 13.2(b) shows that the increasing width t_{before} of many large bursts is fairly short, since large average burst heights and large standard deviations occur for $t_{\text{before}} = 1$h, 2h, and around 24h (red curves). On the other hand, the typical decay width t_{after} of large bursts range approximately from one day to one week since

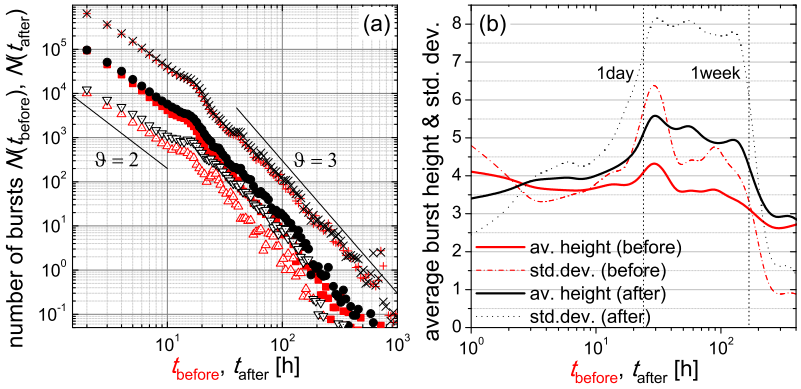

Figure 13.2.: (a) **Number of bursts versus their duration** before ($t_{\rm before} = t^* - t_1$, red symbols) and after ($t_{\rm after} = t_2 - t^*$, black symbols) the maximum. Data for maxima at least 2 times (crosses for $t_{\rm before}$ and plus signs for $t_{\rm after}$), 5 times (filled squares and circles), and 15 times (open triangles up and down) above $\langle x^j(t)\rangle$ were considered. The straight lines representing power-laws according to Eq. (13.2) with slopes $\vartheta = 2$ and 3 are shown for comparison. (b) **Average burst heights** $\langle x^j(t^*)\rangle$ (thick lines) and corresponding standard deviations (thin lines) versus their duration before (dashed and dash-dotted red lines) and after (continuous and dotted black lines) the maximum. The straight vertical lines in (b) mark the time scales of one day and one week.

increased average durations and standard deviations occur in this range (black lines). If we focus on large bursts ($x^j(t^*) > 10$) with intermediately long decays (48h $\leq t_{\rm after} \leq$ 200h), $\approx$ 11% of these bursts exhibit also long increasing widths ($t_{\rm before} \geq$ 48h) – a percentage going slightly up to 15% if weaker burst ($x^j(t^*) > 4$) are also taken into account and down below 1% if the limits for $t_{\rm after}$ are dropped. Bursts with large and small peak amplitudes thus show quite similar behavior (note that bursts with large amplitudes are much less frequent (see Table 13.1) and thus do not significantly affect the results for all bursts). Since only bursts with long widths before and after the peak ($t_{\rm before}, t_{\rm after} \geq$ 48h) can reasonably qualify as 'endogenous' events, these results also suggest that the majority of bursts are either 'exogenous' events or have independent $t_{\rm before}$ and $t_{\rm after}$ values, although there are some correlations between both width parameters (see Table 13.1).

Following previous studies of other data [236, 238] we have also characterized the increase behavior and the decay behavior of all sufficiently long bursts (at least 3h before or after the peak) by fitting two formulas: (i) an exponential law with a characteristic decay time τ,

$$x^j(t) \propto \exp[-|t - t^*|/\tau], \tag{13.3}$$

and (ii) a power law with a critical exponent Θ,

$$x^j(t) \propto |t - t^*|^{-\Theta}. \tag{13.4}$$

While Θ has been suggested as a means for distinguishing exogenous and endogenous processes [237], other authors suggested the exponential fit [238]. Figure 13.3 shows the distributions (normalized probability densities) of the fitted decay times τ [parts (a,c)] and the fitted power-law exponents Θ [parts (b,d)] for increases [fits for $t_1 < t < t^*$, black curves and squares] and decreases [fits for $t^* < t < t_2$, red dashed curves and circles]. All bursts with durations above the respective limit and with peaks (a,b) 2 and (c,d) 10 times above the average have been taken into account. One can see in Fig. 13.3 that exponential fits are more likely to be good if $\tau <$ 24h, and power-law fits are more likely to be good if $\Theta > 0.3$. However, such short decay times τ and such large exponents Θ hardly occur for long activity bursts (see insets of Fig. 13.3). For long bursts, fits of decreasing activity ($t_{\rm after}$) are usually better than fits of increasing activity ($t_{\rm before}$, see Table 13.1, fifth and sixth column), and exponential fits are better than power-law fits.

For activity bursts with $t_{\rm before} \geq$ 6h or $t_{\rm after} \geq$ 6h the distributions of fitted exponents Θ from Eq. (13.4) are nearly Gaussian with means approaching smaller Θ with increasing duration of the bursts [see Figs. 13.3(b,d) and their insets]. The results for large bursts exceeding the average at least 10 times [see Figs. 13.3(c,d)] are nearly the same except for weaker statistics and slightly larger differences between increases and decreases. This confirms that the behavior of large and intermediate bursts is rather similar in general. In addition, the distributions do not change significantly even if we consider only bursts with $t_{\rm before}$ and $t_{\rm after}$ simultaneously above certain thresholds (not shown). This observation suggests that the shapes of increases and decreases within the bursts do not allow distinguishing exogenous and endogenous processes in Wikipedia.

In their recent study of videos on Youtube [165], Crane and Sornette classified bursts with $\Theta < 1$ as 'critical' behavior and those with $\Theta \geq 1$ as 'sub-critical' behavior. The results for these classifications applied to our data

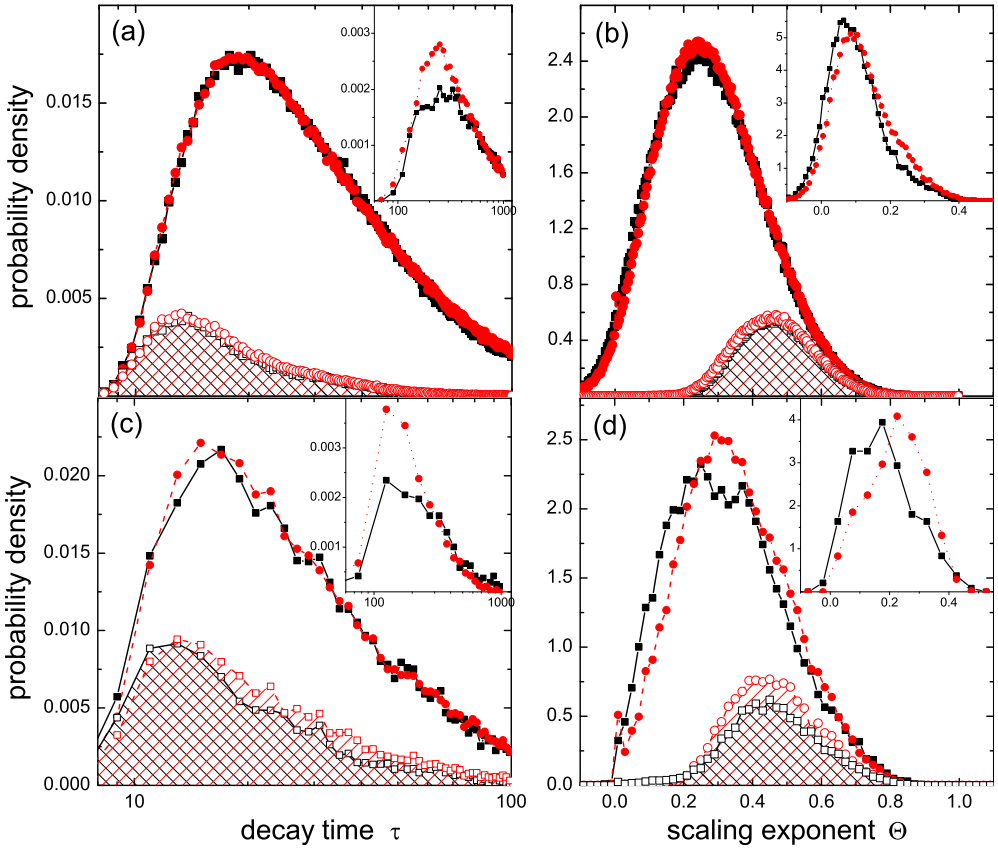

Figure 13.3.: **Distributions of fitting coefficients** characterizing increases (black squares) and decays (red circles) of access rates around bursts of activity exceeding the average accesses (a,b) 2 and (c,d) 10 times for frequently accessed Wikipedia articles from all languages (more than 255 accesses during at least one hour in the recording period). The normalized distributions of (a,c) coefficients τ of the exponential fit [see Eq. (13.3)] and (b,d) scaling exponents Θ of power-law fit [see Eq. (13.4)] are shown for bursts with $t_{\text{before}} \geq 6\text{h}$ (irrespective of t_{after}) and $t_{\text{after}} \geq 6\text{h}$ (irrespective of t_{before}), respectively. The insets show the corresponding distributions for durations exceeding two days. The filled areas under the curves with open symbols in the main panels indicate the fraction of good fits (correlation coefficient $r > 0.75$).

yield only critical peaks, except for approximately one percent of sub-critical peaks with very short durations ($t_{\text{before}}, t_{\text{after}} = 3\text{h}$). The last two columns of Table 13.1 show that even values of $\Theta > 1/2$ completely disappear with increasing peak durations.

In conclusion, it must be said that most bursts are characterized neither by an exponential increase or decay nor by a power-law behavior; the characterization being worse for the increases. The scaling exponent Θ in the power-law fit Eq. (13.4) becomes very small for long bursts, disallowing a classification into exogenous and endogenous processes or critical and sub-critical processes. Apparently, random fluctuations of different amplitudes play a major role in the dynamics of the access-rate time series. In the next Section, we therefore characterize these fluctuations to see if they can reveal some typical and more universal features.

13.1.4. Long-term Correlation Properties

In this section we study the correlation properties of our Wikipedia access-rate data to find out the degree of persistence governing the fluctuations of the corresponding time series. Quantitatively, correlations between normalized access rates $x^j(t)$ separated by s hours are defined by the (auto-)correlation function,

$$C(s) \equiv \frac{\langle \Delta x^j(t) \Delta x^j(t+s) \rangle}{\langle \Delta x^j(t)^2 \rangle} = \frac{1}{\langle \Delta x^j(t)^2 \rangle (L-s)} \sum_{i=0}^{L-s-1} \Delta x^j(i) \Delta x^j(i+s), \tag{13.5}$$

where $L = 7055$ is the series length and $\Delta x^j(t) \equiv x^j(t) - 1$. If the $\Delta x^j(t)$ are uncorrelated, $C(s)$ is fluctuating around zero for s positive. If correlations exist up to a certain number of hours $s_\times$, the correlation function will

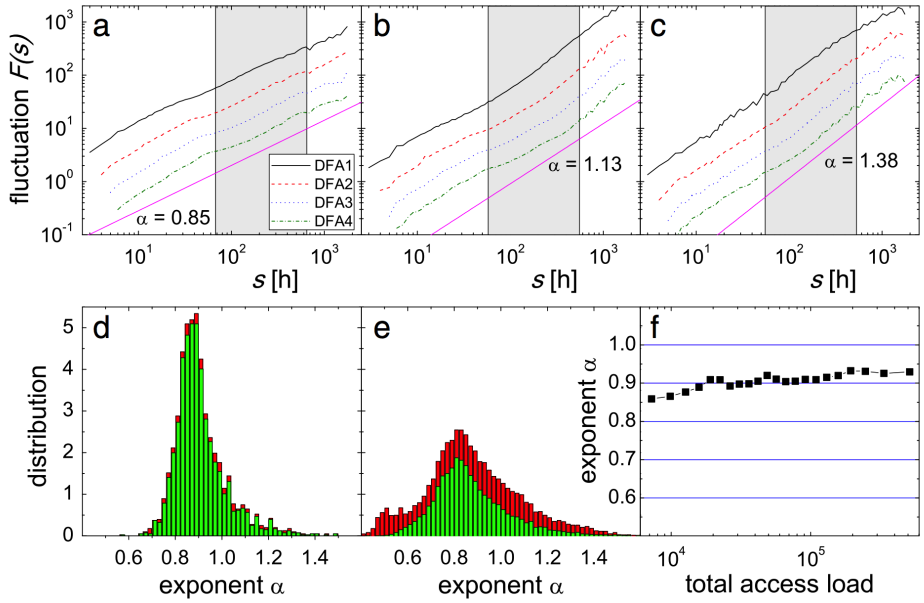

Figure 13.4.: (a,b,c) DFA of access-rate time series of representative Wikipedia articles (the same as for Fig. 13.1). The results for different detrending orders show very similar behavior: DFA1 (black solid line), DFA2 (red dashed line), DFA3 (blue dotted line), and DFA4 (green dash-dotted line); data for DFA2 to DFA4 shifted by multiple factors of 2 for clarity. The straight lines below the data have the indicated slopes and are shown for comparison. On small time scales ($s < 48$h) the effective scaling exponents α are sometimes a bit smaller due to incompletely removed daily rhythms (periodic trends with 24h period, see Fig. 13.4(b) in particular). Parts (d,e) show distributions of the scaling exponents α determined according to Eq. (5.4) for the regime 55h $< s <$ 550h, separately for (d) stationary and (e) non-stationary access-rate time series using DFA1. The green fractions of the bars indicate the proportions of time series for which very good fits of Eq. (5.4) with correlation coefficients $r > 0.98$ were obtained. (f) Dependences of the average DFA1 scaling exponents α on the total access volume that occurred within the considered 42 weeks.

be positive only up to $s_\times$. For the relevant case of long-term correlations, $C(s)$ decays as a power law,

$$C(s) \sim s^{-\gamma}, \quad 0 < \gamma < 1. \tag{13.6}$$

A direct calculation of $C(s)$ is hindered by non-stationarities in the data.

To overcome this obstacle, we apply the detrended fluctuation analysis (DFA) method [28] which has become a widely-used technique for the detection of long-term correlations in noisy, non-stationary time series [29, 30, 31]. We split the subset of the chosen 28 952 Wikipedia articles (with $n^j(t) \geq 256$ for at least one hour) into a group of 2 012 stationary articles with $x^j(t) < 10$ and a complementary group of 26 940 bursting (non-stationary) time series.

We plot $F(s)$ (see Eq. (5.4)) as a function of s on double logarithmic scales and calculate α by a linear fit in the regime 55h $< s <$ 550h. These results are illustrated in Figs. 13.4(a,b,c) for the three representative Wikipedia articles considered already in Fig. 13.1 and for several different detrending orders.

Figures 13.4(d,e) show the distributions of scaling exponents for DFA1 of all (d) stationary and (e) bursting Wikipedia access-rate time series. The distributions for DFA1 (and also those for DFA2-DFA4, not shown) are all peaking in the range of $\alpha = 0.8, \ldots, 0.9$. Comparing the red and green histograms one can see that most fits of Eq. (5.4) are excellent, indicating that the fluctuations of the access-rate time series are following power-laws very well. The access-rate time series are clearly long-term correlated with all effective scaling exponents α being larger than 0.6 for the stationary data recordings [see Fig. 13.3(d)]. The average is $\alpha = 0.91 \pm 0.11$ indicating fairly strong long-term correlations. These findings hold also for the non-stationary access-rate time series [see Fig. 13.4(e)] although the distribution of scaling exponents α is significantly wider in this case with a similar average, $\alpha = 0.88 \pm 0.22$. In particular, the distribution for the bursting time series shows an additional peak at $\alpha \approx 1/2$ exposing existence of a few articles with dominating uncorrelated access behavior among the strongly bursting articles. Figure 13.4(f) shows that the scaling exponent hardly depends on the access traffic, i. e. the data for very frequently and not so frequently accessed articles scales very similarly. There is a slight trend of increasing long-term correlations for increasing access load.

The finding of a rather universal scaling behavior with $0.8 < \alpha < 0.96$ for most articles (65.0 percent of

all 'stationary' and 33.8 percent of all 'non-stationary' articles) quite irrespective of, e. g., total access volume [see Figs. 13.4(d,e,f)], shows that the fluctuations are governed by long-term correlated dynamics similar for most articles. This observation is surprising, since users usually access Wikipedia articles by their own decision and independently of each other. They usually do not follow links (i. e., surf) by more than one or two steps either. Therefore, a naive view would suggest uncorrelated fluctuations of Wikipedia article access frequencies. The observed strongly persistent fluctuations of the article access dynamics must therefore represent the gradual emergence and shift of topics of general interest in society, which may be partly (but not fully) related to exogenous events.

Taking the findings of the previous section into account, we conclude that long-term correlated random fluctuations represent a more characteristic description for the data than individual peak characterizations. We suggest that this description might also be suitable for the process of emergence and shift of topics of general interest in society. In the next section we study if long-term correlated random noise is a good description also for the occurrence of extreme events (peaks and bursts) or just characterizes small and intermediate fluctuations.

13.1.5. Return Interval Statistics

A particularly illuminating way for identifying long-term memory effects in dynamic systems is based on the analysis of return intervals between extreme events that exceed a given threshold [22, 23, 24, 126, 261, 262, 26, 27]. Applications to paleo-climate data [23], wind speeds [263], financial market data [265], medical data [267], telecommunication data [268], internet traffic [269], and times of wars in China [270] have been considered recently. The observed scaling of the effect with the threshold is particularly useful for understanding and anticipating the universality of small and large extremes. Recent results of the analysis of telecommunication data [127] show a bimodal distribution of human interaction events. The inter-event time distributions are not completely Poisson nor power-law, but a bimodal combination of them. Large extreme events have tremendous impact on the system, and may in fact be the main objective of many studies (i. e., prediction of large earthquakes, stock exchange crashes, or extreme weather conditions). Yet they occur rarely and the observational data needed to derive the conclusions directly is scarce. Therefore, scaling approaches are used to relate large extreme events with intermediate events. To describe the recurrence of bursts exceeding a certain threshold q, i. e., $x^j(t) > q$, we investigate the statistics of the sequence of return intervals r between consecutive events.

Therefore, according to Eq. (5.4) and the findings reported in Figs. 13.4(d,e,f) we expect to find a stretched exponential distribution of return intervals characterized by $\gamma = 2 - 2\alpha \approx 0.2$. Figure 13.5 shows the results for thresholds ranging from a moderate $q = 2$ (bursts with maximum amplitude at least twice the average) to very high ($q = 12.5$), and for (a) all English as well as (b) all German Wikipedia articles. In both cases we find that the expected functional form is indeed a good model of the distribution. The values of γ used for the continuous fitted curves in Fig. 13.5, $\gamma = 0.06$ and 0.18, correspond to $\alpha = 0.97$ and 0.91 and are therefore consistent with the characteristics of the long-term correlations reported in Section 4 (where articles from all languages were analyzed together). Rescaling with R_q causes data points for different values of q to come close to the single line.

However, the data collapse is not perfect for two reasons. Firstly, for very short time scales r, i. e., for the lowest r in each data set, we observe a strongly increased probability of return intervals. This is due to additional very short-term correlations such as those we see for the edit return intervals in Fig. 13.6. The presence of these additional short return intervals affects the normalization of the distributions, leading to deviations also for larger r values. Further reasons for the non-perfect data collapse are more speculative: there may be multifractal scaling behavior involved, which would mean that the curves for small and large q scale somewhat differently [261, 262], and finite size effects may be relevant [126, 27]. Beyond these minor deviations, we find that the clustering of extremes for all values of q is basically caused by the same intrinsic dynamics driving the time series' fluctuations. This shows that large extreme events are generated by the same long-term correlated random processes that also drive the fluctuations on small and intermediate scales. Bursts with large peak amplitudes in access rate do thus *not* represent outliers or special events, and they are *not* driven by a dynamics different from the one driving normal fluctuations. This finding supports our conclusions from Sections 3 and 4 that large extremes cannot be uniquely characterized as exogenous or endogenous and that scaling long-term correlated random noise is a good model for Wikipedia access-rate time series.

Figure 13.6 shows the corresponding return interval distribution statistics for article edit events. Since Wikipedia page edits are typically quite infrequent events (compared with article accesses), we have considered every edit as an extreme event and grouped the articles by the gross number of edits as indicated in the legend. The first points of all curves in Fig. 13.6(a,b) deviate from an exponential (straight line in the semi-log plot), and the slope of the exponential fit in the gray region deviates from -1, which would correspond to Eq. (5.5). This is an indication for the presence of local clustering of neighboring edit events. To show random behavior of edit events on different days, return intervals r smaller than an offset r_{offset} are removed in the plots in Figs. 13.6(b,c). Evidently, the return interval statistics of article edits in Figs. 13.6(c) are characterized by Eq. (5.5) (with slope -1) for all subgroups of the data, indicating the absence of particular clustering. The results are in line with the findings reported in [127].

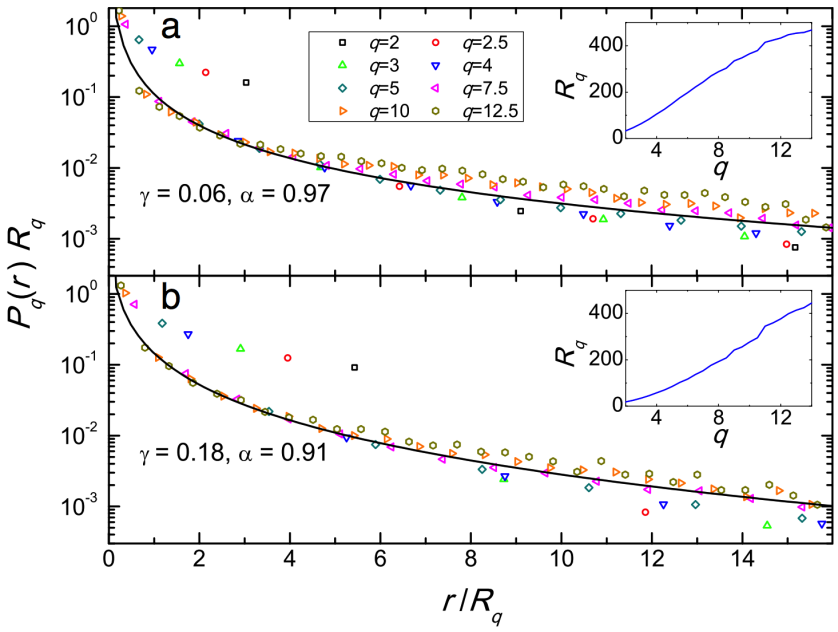

Figure 13.5.: Normalized distributions $P_q(r)R_q$ of return intervals r between bursts exceeding the different thresholds q given in the legend for Wikipedia access-rate data with hourly resolution for (a) English data and (b) German data. The dependence of the mean return interval $R_q = \langle r \rangle$ on the threshold q is shown in the insets. Due to the rescaling with R_q in the main panels, the curves for different q collapse except for the first point, which is higher and indicates strongly increased clustering of bursts on very short time scales. The continuous curves show the stretched exponential decay according to Eq. (5.6) with (a) $\gamma = 0.06$ and $b_\gamma = 7.4 \times 10^8$ and (b) $\gamma = 0.18$ and $b_\gamma = 259$. a_γ was chosen slightly larger than the theoretical values (see [24, 126]) due to discreteness effects and increased clustering of bursts on very short time scales.

Local clustering of edit events partially arises from edits that are immediately re-edited or removed by another editor (e. g., in so-called edit wars [107]) or intermediate saves performed by the same author as larger parts of a text are rewritten. The corresponding high number of quickly following edits leads to a high weight of short return intervals in comparison to all return intervals. This explains the deviations of the plots in Fig. 13.6(a) from the simple exponential Eq. (5.5) as shown by disregarding short return intervals below $r_{\text{offset}} = 24\text{h}$ in Fig. 13.6(c); note that disregarding only $r < r_{\text{offset}} = 2\text{h}$ is not sufficient [Fig. 13.6(b)].

Except for the observed short-term effects, the simple exponential decay of $P_q(r)$ indicating the absence of clustering of edit events is clearly dominating. In other words, no long-term correlations are present in the edit-event time series. This is surprising, since editorial activity could be expected to be also driven by the gradual emergence and shift of topics of general interest in society just like access activity. However, editorial activity is in fact organized in a much more complex fashion. There is a social network of editors controlling the process via discussion pages and to-do lists. In addition, some editors feel responsible for several articles and continuously check (and sometimes correct) all changes. Furthermore, exogenous events, which are often uncorrelated, might have a stronger effect on edits than on access behavior. There are thus many additional reasons for article re-editions besides changes in general public interest. We hypothesize that editorial activity is caused by several superimposed reasons and therefore appears rather random from the statistical point of view.

Editions of Wikipedia articles can thus not be the reason for the long-term correlated fluctuations and the long-term correlated burst occurrences in the access-rate time series. It is rather fascinating to see that the partly orchestrated editorial work appears uncorrelatedly random, while the access activity of many independent users leads to stable and rather universal long-term correlations. However, the two processes cannot be completely unrelated either. One can see already with the naked eye that the access-rate and edit-event time series of the article shown in Figs. 13.1(e,f) are cross-correlated (although no cross-correlations are visibly in the other two examples). A more detailed analysis of cross-correlation properties will be done in a future work.

13.1.6. Conclusion and Outlook

In this work we have characterized and compared time series representing two separate complex processes that govern the spread of information in a complex on-line system – the encyclopedia Wikipedia. The two considered

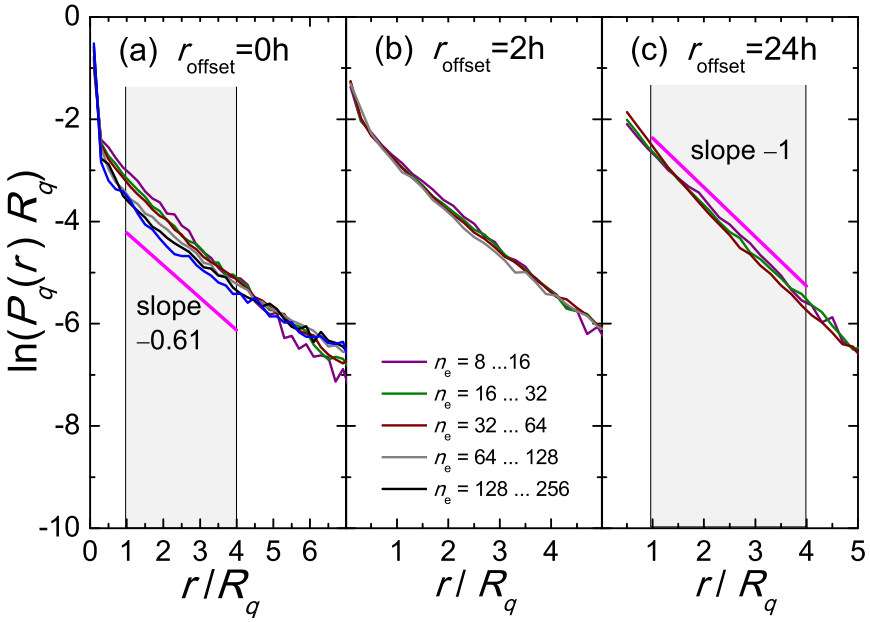

Figure 13.6.: Normalized distributions $P_q(r)R_q$ of return intervals r between article edit events for English Wikipedia articles grouped according to their total numbers of edits n_e during the observational period (see legend, $>$ 1000 articles in each group). In parts (b) and (c) return intervals r below offset times of r_{offset} = 2h and 24h, respectively, have been excluded to eliminate effects of short return intervals (e. g., edits immediately reverted by another editor in so-called edit wars [107]). The straight lines in (a) and (c) are exponential fits similar to Eq. (5.5), but with the slope given in the plots. The expected slope -1 is obtained in (c), when return intervals below 24h are excluded.

processes, i. e. editorial activity (information upload) and article access (information viewing), are performed by nearly distinct populations in different ways. The processes appear to be closely related only superficially. Instead, by characterizing the details of the dynamics, we have systematically found that access-rate time series can fairly well be represented by long-term correlated random fluctuations. The rather large fluctuation exponent $\alpha \approx 0.9$ found by DFA in Section 4 is quite universal for most Wikipedia articles, independent of the occurrence of large peaks (bursts) and valid for more than one decade of time scales (from 1-2 days up to at least one month). The return interval statistics have confirmed that long-term correlations can also explain the recurrences and clustering of extreme events and strong bursts. There are thus no different driving mechanisms for small, intermediate and large fluctuations. We think that the observed strongly persistent fluctuations of the article access dynamics represent the gradual emergence and shift of topics of general interest in society, which may be partly (but definitely not fully) related to exogenous events. Extreme events (bursts) in Wikipedia access-rate time series can thus not easily be classified as caused by exogenous or endogenous events in agreement with our results on the form of the bursts. However, the frequency of extreme events is very well described by a scaling rule (power law). We conclude that the access activity follows scaling laws (characteristic of emergent phenomena in complex systems) in several aspects, even though the behaviors of all individual users are invisible to the other users.

On the other hand, the statistical properties of edit time series appear to be much closer to white noise. We have clearly found a simple exponential decay of the probability of return intervals $r > r_{\text{offset}}$ = 24h between edit events, which holds irrespective of the total number of edits for the considered articles. This shows that long-term correlations are not important in edit time series. We think that this behavior is caused by the complex orchestration of the editing process of Wikipedia articles. The corresponding feedback loops and synchronization seem to provide a different mechanism that destroys or at least interferes with long-term correlations.

Our next step is studying and comparing the characteristics of processes that govern information spread in communication-centered networks, where an explicit social network is used to convey direct messages between users (e. g., email, Twitter). In these systems information editing and viewing are not separated, and there is no orchestration of information upload. We expect that the emergence and the decay of topics of general interest should thus also be characterized by long-term scaling correlations. We think that it might be interesting to study relations between the emergence of new topics in, e. g., Twitter messages and corresponding activity in the Wikipedia system.

Regarding Wikipedia we suggest that further research should study the interplay of the network structures with the dynamics of article re-editing and access. Since Wikipedia articles are organized in topical groups and often cross-linked, one can study and compare how these structures affect the emergence of bursts in both, access and editorial activity. In addition, the editorial processes can be studied in more detail, since (anonymous) information on the actions of many editors and their interplay is also available. Besides this, an alternative network

representation could be created by studying the strength of temporal cross-correlations between the access-rate or edit-rate time series of *different* articles. This dynamical network structure could be compared with the link-based static network structure. Further work is also needed to clarify differences between language communities. If the dynamics of article access indeed reflect the gradual emergence and shift of topics of general interest in society, one can study differences of these processes between language groups and cultures by separate analysis of data according to the languages of the articles.

13.2. Dynamics of Correlation Properties in Multi-Layer Networks

In this section we investigate the dynamics of a content access process as a function of time on a mesoscopic level using correlation networks. Functional networks are created as described in section 9.2 to extract dynamic properties of this process. We identify several types of changes in the link strength distributions which could be indicators for specific transitions within the system.

Initial Experiments: Link properties were obtained from time series episodes of length 120 days. Daily binning (aggregation of all events during a day) was applied to the raw sequences with hourly resolution. To connect this part of the study to previously presented results, we used the same Wikipedia pages as in the last section (see figure 13.1). Figure 13.7 shows raw results (link strength distributions for various neighborhoods and three time frames) of the analysis procedure applied to data for the year 2009 for page *'Illuminati_(Buch)'*. The analysis of the two other pages, *'Heidelberg'*, and *'Amoklauf Erfurt'* and for the year 2008 provided comparable results but are not shown here. Results of this initial experiment allow to identify for what kind of changes (see labels in table 13.2., and figures 13.7., and 13.8.) one should look for in a more systematic approach on more samples (which means more central nodes from multiple different topics).

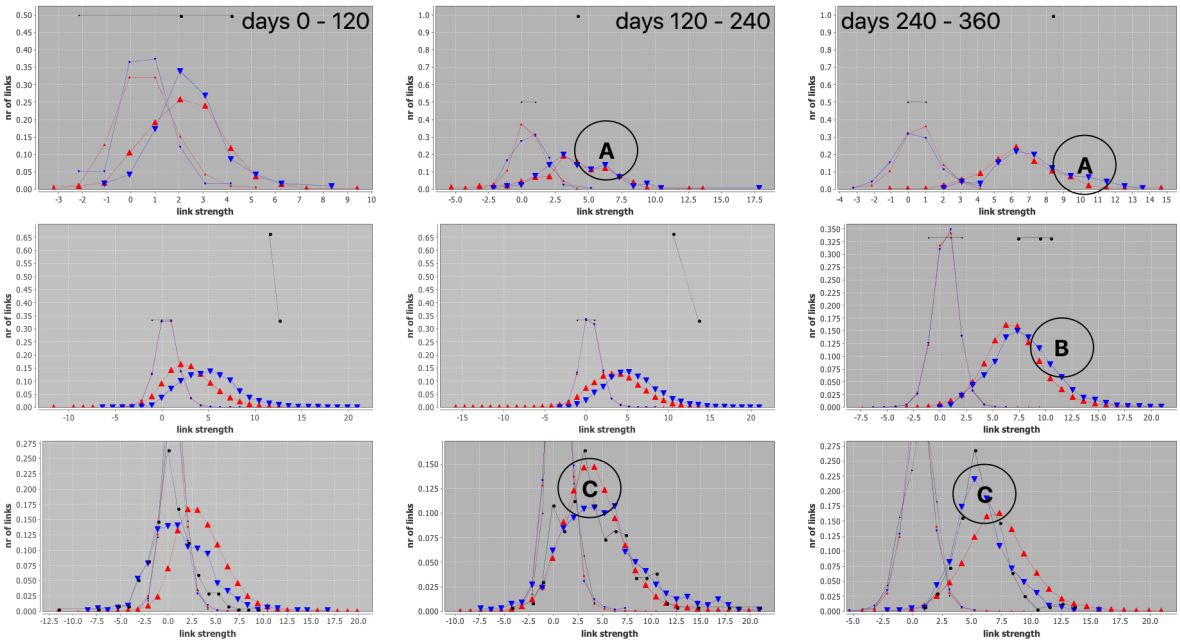

Figure 13.7.: **Link strength distributions for time dependent correlation networks**. For the Wikipedia page *Illuminati_(Buch)* the neighborhood correlations were calculated (Eq. 9.11) for non- overlapping episodes of daily access rates of length 120 days starting January 1-st 2009. Results for three non-overlapping time episodes are shown in columns. Curves with large symbols represent results from measured data and small symbols for randomized (shuffled) time series data. Top row shows *correlation for closest neighbors*, which are correlation between IWL (black curve), correlation between CN and the local neighborhood (blue curve), and correlation between CN and the global neighborhood (red curve). Middle row shows group internal correlations (*intra correlations*) for IWL (black), LN (blue), and GN (red). The *inter group correlations* are shown in the bottom row for IWL and LN (black), IWL and GN (blue), and for LN and GN (red). **A** marks overlapping bimodal distibutions in the close neighborhood during episode 2 and 3. **B** marks a small - but stable over time - difference between two distributions (red and blue) while the difference to distributions for shuffled data is huge in both cases. The blue curve has slightly higher values than the red curve in all episodes. Finally, **C** marks a change in the differences of inter group correlation. During episode 3 the red curve shows higher values than the blue which has higher values in episodes 1 and 2. Symbols and group labels are as in figure 13.9.

In order to prepare such a study, we provide two visual representations of results as shown in the next section.

13.3. Interpretation of Correlation Link Strength Distributions

Calculated links for reconstructed functional networks have to be filtered in order to extract highly relevant links which carry information. Which links are and are not relevant depends on the selected topic, the context, and the selected algorithm for link strength calculation. Furthermore, it can be influenced by the selected filter method (see sections 10.6 and 10.7). A fixed threshold, adaptive thresholds, or graph based filtering like the planar maximum filtered graph method are common methods for this step.

The distribution of the cross-correlation link strengths for pages with a wiki link to the core (a,d), for all possible page pairs within the group IWL, LN, and GN (b,e) and for all possible pairs of pages from two different groups IWL, LN, and GN in (c,f) are shown in Fig. 13.8. The definition of the individual groups is based on figure 6.2 in chapter 6.

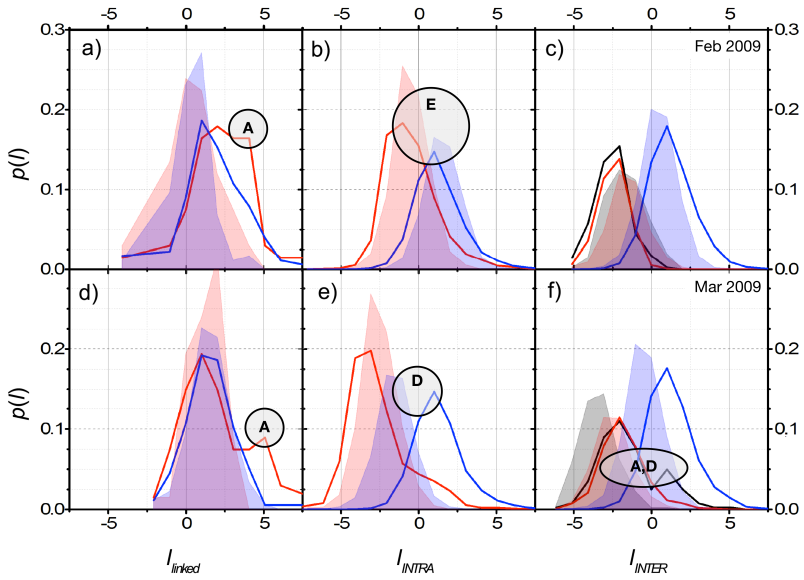

Figure 13.8.: **Time-resolved link strength distribution for functional networks calculated for the local network of a Wikipedia page.** The functional network around page '*Illuminati_(book)*' was calculated for two non-overlapping time frames of length 28 days. The top row shows the link strength distributions for February 2009 and the bottom row for March 2009. Colored lines show link strength distributions for measured data at a daily resolution and colored areas indicate the results for surrogate data based on random shuffling of raw data series. The blue curve represents the correlation between CN and the local neighborhood while the correlation between CN and the global neighborhood is shown in red in panels a) and d). Panel b) and e) show the group internal (intra) correlations for LN (blue), and GN (red). The inter group correlations are shown in panels c) and f), for IWL and LN (black), IWL and GN (blue), and for LN and GN (red).

Label	Description
A	bimodal distribution with at least two peaks
B	stable difference between distributions from raw data
C	changing differences between distributions from raw data
D	large difference between distributions from raw data and shuffled data
E	small difference between distributions from raw data and shuffled data

Table 13.2.: **Qualitative analysis of link strength distributions.** Visual inspection of link strength distributions provides features, which can be used in advanced machine learning algorithms (examples of supervised learning).

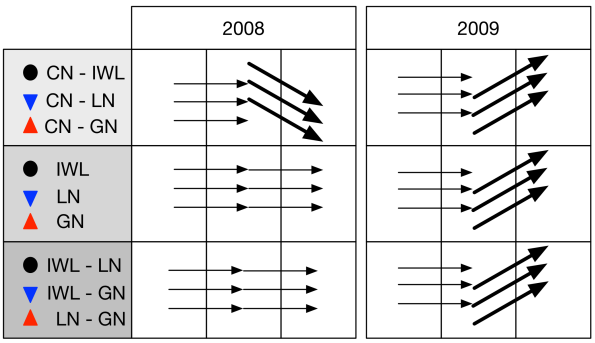

Figure 13.9.: **Drift of average correlation strength** for individual regions in neighborhood networks around the Wikipedia page *Illuminati_(Buch)* for years 2008 and 2009. Symbols and group combinations as in Fig. 13.7.

Figure 13.9 shows a scheme to illustrate changes in correlation networks based on quantitative results. In this case, the position of the maximum in the link strength distribution was used. For each group we found different shapes, but per group the shapes did not change fundamentally over time. Other metrics can be variance, skewness, and kurtosis of the distribution. If values for two consecutive episodes differ more than a defined range the figure shows an arrow up or down, otherwise a horizontal arrow indicates no relevant change between two time intervals. Usage of a fixed threshold value is often appropriate, especially if the link strength was calculated as the normalized correlation coefficient (see Eq. 9.11). For network reconstruction the link strengths ($l_s(i)$ values) are used to create an adjacency matrix $\mathbf{A}$. $\mathbf{A}_{\mathrm{CN}}(i)$ represents a local functional correlation network around a central node CN during a given time interval i. The example in Fig.9.3 shows a comparison of two functional networks - access activity (a), edit activity (b) - and the underlying static network (formed by page links) for the Wikipedia page '*Illuminati_(book)*' in English language with $\mathrm{CC}_a^{(i,j)}(t) > 0.75$. Since not much is known about the individual meaning of different link strength distributions based on different metrics it is important to compare results obtained from real world data sets with simulation results using well known time series models.

One can assume to find stronger connected (or less dense) networks depending on the level of public interest in a given topic. Correlation networks with larger connected components are the results of an access process with stronger correlated actions. Such correlation within the user community can be generated by a stronger media presence of the topics or caused by natural phenomena like earthquakes or floodings.

Uncorrelated usage of the pages, which means that there is no common interest during a given period in time in the topic, would lead to a link strength distribution like shown for the surrogate data (small symbols in 13.7, and colored area in 13.8). This is described by pattern (E in table 13.2 and in figure 13.8) where correlation links are not notably different from random links.

Exogeneous bursts in the access-rates seem to be caused by a common interest of many users within a short time interval. This means, access activity is highly correlated and leads to high correlation link strengths. Fig. 13.8 and the two sided Kolmogorov-Smirnov test [141] allow the conclusion, that exogeneous bursts have a significant influence on the link strength distribution of a functional network. During the episode before the exogeneous burst (see figure 13.1.c) significant differences between the measured link strength distribution and the surrogate data were determined for groups CN-IWL and CN-LN. For groups IWL, LN, and IWL-GN we could find a very high significance with p values less than 10^{-30} (not shown). In the presence of the burst (during the second episode) the significance of differences of link strength distributions changes. The correlation between pages within the core network decreases. At the same time the correlation between pages in the neighborhood increases. This indicates that during increasing interest in the topic - indicated by increased access-rates - people access pages within the neighborhood more often. This is interpreted as an increase of the relevance of the topic.

For such networks we calculate topological properties per time interval, and for each metric (average clustering, local clustering, average degree, size of largest component) we show time-dependent results in figure 13.10.

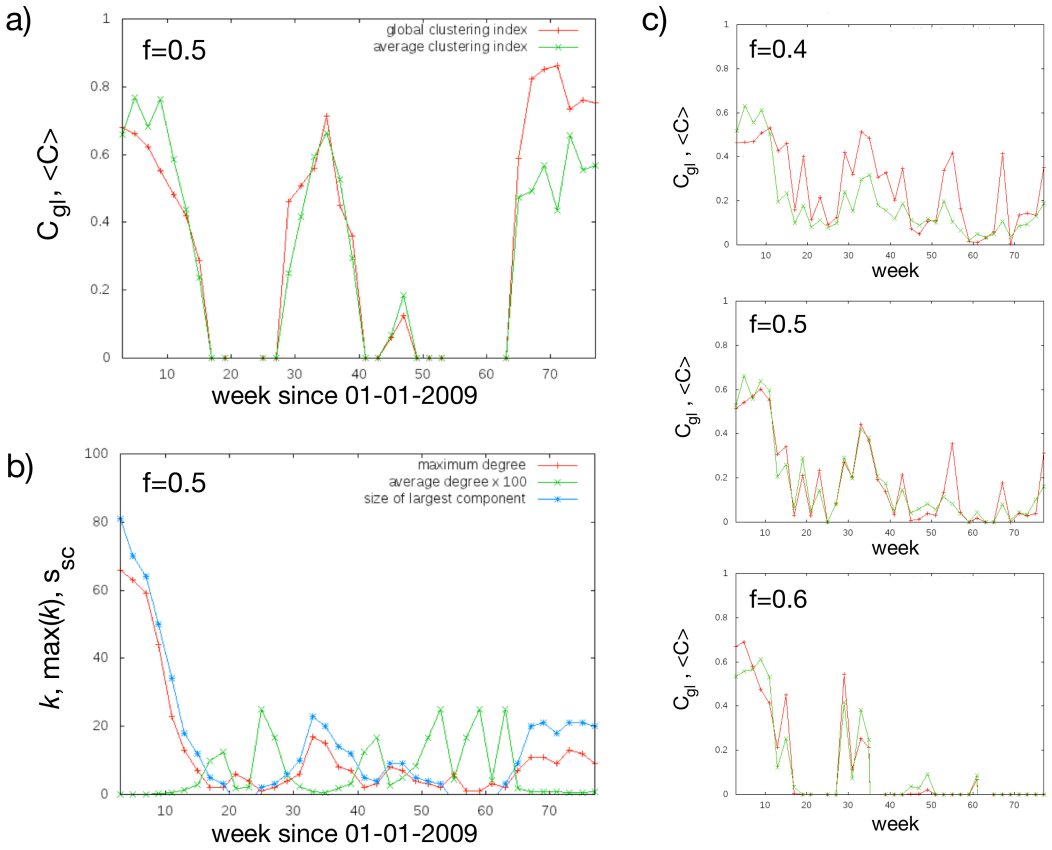

Figure 13.10.: **Change of topological properties over time.** (a) shows that global clustering index (red) and average local clustering index (green) are not equal but both show the same pattern over time. (b) shows a comparable curve for maximum degree (red) and size of largest cluster (blue) but a very different curve for the average degree (green). The sequence of three images in (c) shows the impact of the threshold filter (only links with link strength $l_s > f$ are used) on the global and local clustering index. The peaks around 55 and 68 disappear as a consequence of the filter. Relevance of a peak in such a structural property can be evaluated in this way based on the ratio of the result for two different filter thresholds. Only if a peak does not disappear by changing the filter threshold it should be interpreted as relevant. (a) and (b) are for the Wikipedia page '*Heidelberg*' and (c) for the page '*German cities*'.

13.4. Information Flow in Correlation Networks

One of our initial research question was: Can we measure a significantly higher correlation link strength for directly linked pages?

In chapter 6 we introduced the concept of local neighborhood networks. Here we use it in the special case of linked Wikipedia articles of multiple languages. One has to distinguish between simple page links, inter-wiki links, and external links (which are omitted for simplicity). Figure 13.11 shows for two of those local neighborhood networks a comparison of the average link strength (orange and black) as a function of time in the presence of the raw activity (blue, log(access rate) on right axis).

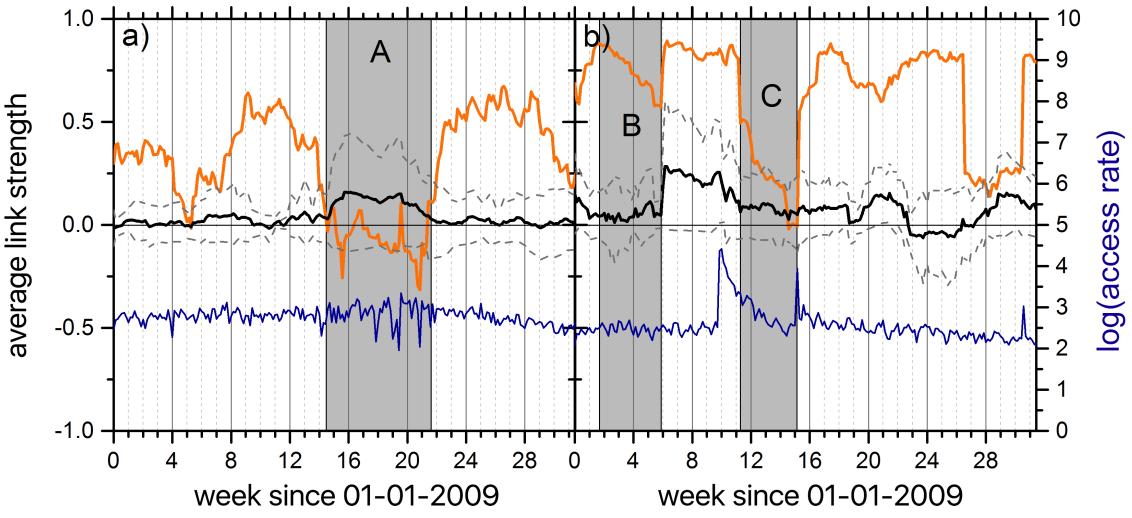

Figure 13.11.: **Time-resolved average link strength for functional networks** calculated for the local neighborhood network (orange, left y-axis) and the global neighborhood network (black, left y-axis) for functional networks (a) around '*Amoklauf von Erfurt*'. and (b) around '*Illuminati_(book)*'. The blue line show the logarithm of the daily access-rates (right y-axis) for page CN (in both cases the page in German language). The average link strength within the local neighborhood (orange) is compared with the average link strength in the global neighborhood (black). The dashed gray lines show the variance $\pm\sigma$ around the average link strength for the global neighborhood. During a period of increasing interest in the topic in English language (see burst in figure 13.1.c) the access-rate correlations drop to an average value below zero and the access time series correlations within the global neighborhood are increased significantly (see region A). The experiment shows fluctuations in link strength overlapped by sporadic drops (region C) or jumps (region B and C) in both groups (orange and black).

Fig.13.11 shows the time-resolved average link strength for functional networks calculated for the local neighborhood network (orange) and the global neighborhood network (black) for functional networks around page '*Illuminati_(book)*' (a) and around page '*Amoklauf von Erfurt*' (b). Region A indicates the time of increasing attraction within the global context. At the same time the average correlation between the access-rate time series increases, and the correlation between time series measured in the local context decreases. Correlation values around zero, which means time series are uncorrelated (see region B in Fig. 13.11) are normal for the global context of a page with high local relevance, but in the presence of the burst, also the correlation increases. Region C shows decreasing correlation in local and global context after an exogeneous burst. However, within the local context, the time series have a stronger correlation. This can also be seen in fig. 13.8. (a,d) and fig. 13.8.(b,e).

14. Conclusion and Outlook

> Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.
>
> *(Marie Curie)*

Section 1.1 lists multiple problem fields which were addressed by this work. During the project SOCIONICAL we initially were interested in properties of non-stationary processes in socio-technical systems. Especially in the example of Wikipedia, we study the creative and intellectual processes in which Wikipedia pages are created and edited by groups of individuals. The information access process by millions of anonymous readers is the second process in our focus. As a third potentially related process we consider changes in stock prices and trading volume in financial markets. All those processes coexist and might be related to each other and thus, one can expect that they also influence each other. We want to understand how those processes are coupled and if the coupling changes over time. Which properties do affect the structure of the underlying interconnected network? What is the right time scale that allows us to find significant correlations? In this work we did not yet look into causation between the considered processes and participating components. The contributions to the field of complex systems research can be organized in three groups:

I. Reconstruction of Functional Networks Based on Cross-correlation and Event-synchronization.

Contribution 1: Investigation of network reconstruction methods based on measured data, modeled data, and randomized real data (see chapter 9 and 10). According to properties of the data set an appropriate correlation analysis has to be selected. Event-synchronization is suitable for event time series, while cross-correlation analysis is better in case of continuous signals. Our approach uses normalization based on structural context (chapter 6).
Contribution 2: Investigation of the impact of multiple filters on the resulting link set (see chapter 10). If the link strength distribution is a bi-modal distribution (from two Gaussian distributions), identification of two sub-groups is possible, e.g., based on Gaussian kernel estimation. We found many cases in which a clear separation was not possible by Gaussian kernel estimation. Instead, we developed an approach to identify a dynamic cut off based on the assumption of two overlapping distributions, a Gaussian, and a Poisson distribution. A second method combines multiple link layers and thus uses a two dimensional distribution. Since individual metrics expose specific properties, we are able to identify groups of nodes (clusters) which show significant differences compared to random or randomized time series.

The underlying structure of a complex system might be hidden but accessible through a dynamic correlation network. Such functional networks are used as a tool to measure different types of hidden interactions between network nodes.

II. Visualization and Characterization of Time-dependent Functional Networks to Describe Dynamic Processes in Complex Systems.

Contribution 3: We developed a generic framework using time series analysis methods to reconstruct multiple representations of a complex system (chapter 9). Such overlapping facets allow analysis without disconnection or segregation of the components. Analysis of resulting time-dependent networks leads again to time series data which can finally be related to the initially measured raw data. This allows to study the coupling mechanisms of components in integrated systems on multiple scales.
Contribution 4: We applied sliding window techniques to identify time-dependent structural properties. Coupling between structure and function can be identified this way (see section 13.2).

III. Aggregation of Information with 'Per Element' Granularity Towards System Properties Represented by Group Averages.

Contribution 5: We revised the phase model for traffic data based on correlation properties for pairs of measured time series episodes and based on measured fluctuation properties (see chapter 18).

Contribution 6: We developed an approach to study information flows within the neighborhoods of Wikipedia articles based on access-rate statistics (see chapter 11, 13, and 15).

The definition of the right sub data set (sample) is complicated in a multilingual global social network. For example, lists of companies in a stock market (the components of an index) are well defined, but companies offer information in public web pages, but in different language. Such pages vary in content and of cause in user attraction. Also the Wikipedia pages about those companies are often not comparable to each other. Most importantly because the pages are written in different languages and speakers of different languages are not equally interested in financial or business topics.

The representation index allows us to describe the data set in more detail. Interpretation of final results can be better embedded into a quantitative context which shows, if the findings are in line with the initial hypothesis or if significant looking results could be artifacts caused by so far unknown hidden or artificial influences.

Furthermore, in many cases multiple pages exist for one company already in one language. In our approach the level of representation of topics, or companies, can be measured using a definition page and the neighborhood graph using the representation index (see P1 in section 1.1). The available amount of information in Wikipedia differs dramatically. Especially if data is aggregated from all available languages it is important to normalize this data. A context sensitive normalization of time series data was developed in this work (P2). In order to address hidden bias, we developed the representation index of Wikipedia pages (P3). This enables us to study them in context networks. A generalization allows a context sensitive analysis based on structural information extracted from arbitrary content networks, not only Wikipedia (P4).

To model complex systems means to describe the properties of the system in such a way which allows a quantification of several of its (sometimes hidden) aspects, or in simple words: we want to measure properties of a complex system, even if a direct measurement is not possible. Such indirect methods require the representation of complex systems within a scalable computational environment because one has to compute non measurable properties or one wants to simulate the systems behavior to study the time evolution based on extracted models. For several years the network approach has been evolved and more and more research was done with a strong focus on complex networks. Coupled networks (NoN) show special properties which are not observed in isolated networks. Especially such effects are results of interactions as a consequence of feedback loops and allow to call such systems *complex system*.

Instead of simplification achieved by chopping a system into pieces or slices new integrative approaches are necessary. Therefore one needs appropriate representations of complex systems which can be used in HPC[1] and HPA[2] environments. Beside these technical requirements, which are related to IT engineering and IT operations, also new analysis methods have to be implemented on top of recent large-scale IT infrastructures. Optimal data representation and access patterns to stored data[3] are crucial aspects in this context.

Research questions drive new technology which is used to answer certain questions. With advanced technology one can answer some questions in more detail or even faster and more efficiently than before. This shows, how theoretical concepts, computational experiments, IT technology, and data management are related to each other. If one wants to understand processes behind large social networks, one has to work with data, derived from such networks. Preparation of raw data with repeatable and reliable procedures require special technology and procedures. Nowadays, network algorithms can be applied to very large networks. But, so far there is no general approach for working with large time-dependent multi-layer networks available. Time-dependent properties of the many constituents of a system are measured, stored, and analyzed in scalable distributed storage systems using cloud technology. But on top of this, we need a system which is able to connect different views as multi-layer networks.

For example, different types of interactions between subsystems (groups of particles or groups of people) can be, e.g., the gravitation and electromagnetic forces or social forces in case of a social context. All forces together have an influence on the trajectories of the elements. The gravitational and electromagnetic forces in case of charged particles lead to a situation in which one knows much about the way of interaction and a common law - the equation of motion - can be written down and solved. Although the presence of the gravitational field is unlimited one skips the interaction between particles which are far away from each other for computational efficiency. In social networks we have different properties. What does it mean, two persons are far away, or people attract each other in a social network? In terms of their location (spatial embedding) it might be thousands of miles

[1] HPC: High performance computing refers to highly optimized computation on large-scale grid systems with focus on expensive operations rather than huge data sets.

[2] HPA: High performance analysis refers to rather simple analysis tasks on so called Big Data platforms. Simplified computational models are used to achieve efficient data handling on large-scale.

[3] The data can be collected data from SMAs or simulation results, or ideally both since both types of data have to be analyzed in the same way. The combination of simulation and real world data analysis by applying extracted model parameters is the essential aspect of Big Data based science.

but in terms of communication, it is just one call or a digital message between them. In this situation one has not any longer a scalar value which determines the distance, but rather multiple distances (one per layer) become the elements of a distance vector which have to be combined in the right way. If we assume superposition, a linear combination seems to be the valid mathematical operation to combine the values. But even in this case, the weights are not known yet. Having multiple so different interacting layers may cause a situation in which super position is not correct.

The distance of nodes within a functional network strongly depends on the selected properties and the method to reconstruct the connecting links. And as such a functional network is usually not stationary as seen in chapters 8 and 13. It is very important to incorporate such information in new analysis algorithms.

Multiplex or multi-layer algorithms can help to connect existing models which express well studied phenomena and properties on different scales. A scalable framework for large models, which cover multiple effects in one embedded environment is the result of our effort to address (P4).

Integrated and interdisciplinary research has to be done carefully. Not all possible integrations are useful but some might be. For example, in the context of socio-physics one can understand the individual behavior of each person. In the case of a group of persons, group dynamics can affect the average behavior of the group in a different way which can not be determined by just looking on the measurable averages. Such emergent properties are critical since they might be not known and not obvious. In which case do we have to consider the properties and the influence of single elements on the group and in which case are the group properties more relevant? This questions can not be answered without a given context and without impact analysis - but therefore we need appropriate models, contextualization methods, and the right technical framework to apply to the right data.

So far, we can find many rather specific models which are able to explain a certain aspect without looking into the environment which means, one has either a limited size of the system or one disconnects the components of the system. Without the original embedding into the natural environment many systems change their behavior.

Socio-technical systems are related to human communication and motion (not necessarily by vehicles, but also by walking in complex geometries like large buildings or even cities). One can work with the trajectory for a person which is walking in a city or a stadium. For a group of persons one has to track each person's trajectory together with structural information, which describes, if the persons within the model are related to each other and if they form a group, which might be the source for *group dynamics* (see also [193]). In case of thousands of people the data volume increases and leads to high computational cost. Furthermore, it becomes important to have multiple models for different detail levels and different scales which are connected to each other. Hence one has to work in an integrated environment which shows HPC and HPA characteristics. With large-scale multilevel models it will be possible to study so far unknown effects and phenomena in dynamic complex systems.

Based on the proposed generalized analysis framework we started the development of a distributed software system called HDGS (abbreviation for *Hadoop Distributed Graph Space*). The core of the system is formed by a metadata management system and a generic processing engine for time series buckets and time-dependent multi-layer networks. Data for analysis is stored in Hadoop clusters in different ways and depending on required access patterns (defined by algorithms) we use (a) key-value stores, (b) object stores, and (c) indexed data sets[4] to optimally keep data according to heir dominant access pattern defined by analysis algorithms.

Initially, the core functionality of our tool set was limited to a specific problem and a certain type of data - Wikipedia in our case. It was not yet possible to incorporate Twitter messages, Facebook posts, or company internal email communication. Our new software allows integration of multiple different data types from multiple data sources by combining a Big Data platform and the linked data approach.

Furthermore, in the Hadoop.TS software package we adopt the concept of an oscilloscope to visualize time series data sets stored in a Big Data environment. In this way social media applications and business data can be combined in order to measure properties and to perform dependency analysis of multiple properties from multiple different complex systems.

By working in this technical environment, researchers from physics, social science, economy, and computational science are getting closer in order to collaborate in the new fields called socio-physics and econo-physics. The large number of joined interdisciplinary research projects and the absence of a public software platform which supports such interlinked approaches is the motivation for my future work on HDGS.

[4]Indexed data sets allow full text analysis and linguistic analysis of node properties.

References

[1] Elad Segev. Mapping the international: Global and local salience and news-links between countries in popular news sites worldwide. *International Journal of Internet Science*, 5(1):48–71, 2010.

[2] Zhu Wang. Learning, diffusion and industry life cycle. Technical report, Federal Reserve Bank of Kansas City, 2006.

[3] Rosario N. Mantegna and H. Eugene Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge University Press, 1st edition, 2000.

[4] Mirko Kämpf, Eric Tessenow, Dror Y. Kenett, and Jan W. Kantelhardt. The detection of emerging trends using wikipedia traffic data and context networks. *PLoS ONE*, 10(12):e0141892, December 2015.

[5] Homepage: Wikimedia foundation. retrieved March 22, 2016 from: `http://wikimediafoundation.org/wiki/Home`.

[6] Cathy O'Neil and Rachel Schutt. *Doing Data Science*. O'Reilly Media, 1st edition, 2013.

[7] Phillip I. Good and James W. Hardin. *Common Errors in Statistics (and How to Avoid Them)*. John Wiley & Sons, Inc., fourth edition, 2012.

[8] Mirko Kämpf, Sebastian Tismer, Jan W. Kantelhardt, and Lev Muchnik. Fluctuations in wikipedia access-rate and edit-event data. *Physica A: Statistical Mechanics and its Applications*, 391(23):6101–6111, 2012.

[9] Jan W. Kantelhardt, Matthew Fullerton, Mirko Kämpf, Cristina Beltran-Ruiz, and Fritz Busch. Phases of scaling and cross-correlation behavior in traffic. *Physica A: Statistical Mechanics and its Applications*, 392(22):5742–5756, 2013.

[10] Berit Schreck, Mirko Kämpf, Jan W. Kantelhardt, and Holger Motzkau. Comparing the usage of global and local wikipedias with focus on swedish wikipedia. *ArXiv e-prints (1308.1776)*, August 2013.

[11] Mirko Kämpf and Jan W. Kantelhardt. Hadoop.ts: Large-scale time-series processing. *International Journal of Computer Applications*, 74(17):1–8, July 2013. Full text available.

[12] Mirko Kämpf, Jan W. Kantelhardt, and Lev Muchnik. From time series to co-evolving functional networks: dynamics of the complex system 'wikipedia'. In *Proc. Europ. Conf. Complex Syst. (ECCS)*, 2012.

[13] P. Gawroński, K. Kułakowski, M. Kämpf, and J. W. Kantelhardt. Evacuation in the social force model is not stationary. *Journal Acta Physica Polonica*, 121(2-B):77–81, 2011.

[14] M. E. J. Newman. Complex systems: A survey. *American Journal of Physics*, 79:800–810, 2011.

[15] Herbert Pietschmann. *Die Atomisierung der Gesellschaft*. Ibera-Verlag, 2009. `http://vorarlberg.orf.at/radio/stories/2509849/`.

[16] Complex systems modeling: Using metaphors from nature in simulation and scientific models, August 2003. retrieved March 22, 2016 from: `http://informatics.indiana.edu/rocha/complex/csm.html`.

[17] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Phys. Rev. E*, 51:4282–4286, May 1995.

[18] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 2009.

[19] Uri Wilensky. Netlogo. http://ccl.northwestern.edu/netlogo/, Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 1999.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. ISBN 3-900051-07-0.

[21] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[22] Amir Bashan, Ronny P. Bartsch, Jan W. Kantelhardt, Shlomo Havlin, and Plamen Ch. Ivanov. Network physiology reveals relations between network topology and physiological function. *Nature Communications*, 3:702, 02 2012.

[23] Dror Y. Kenett, Michele Tumminello, Asaf Madi, Gitit Gur-Gershgoren, Rosario N. Mantegna, and Eshel Ben-Jacob. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS ONE*, 5(12):e15032, 12 2010.

[24] Reginald D. Smith. The spread of the credit crisis: view from a stock correlation network. *Journal of the Korean Physical Society*, 54(6):2460–2463, 2009.

[25] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *EPL*, 87, 2009.

[26] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, 2009.

[27] J. F. Donges, H. C. H. Schultz, N. Marwan, Y. Zou, and J. Kurths. Investigating the topology of interacting networks. *The European Physical Journal B*, 84(4):635–651, 2011.

[28] M. Paluš, D. Hartman, J. Hlinka, and M. Vejmelka. Discerning connectivity from dynamics in climate networks. *Nonlinear Processes in Geophysics*, 18(5):751–763, 2011.

[29] Y. Berezin, A. Gozolchiani, O. Guez, and S. Havlin. Stability of climate networks with time. *Scientific Reports*, 2:666 EP –, 09 2012.

[30] Joel N. Tenenbaum, Shlomo Havlin, and H. Eugene Stanley. Earthquake networks based on similar activity patterns. *Phys. Rev. E*, 86:046107, Oct 2012.

[31] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.

[32] S. Havlin, D. Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J. W. Kantelhardt, J. Kertész, S. Kirkpatrick, J. Kurths, J. Portugali, and S. Solomon. Challenges in network science: Applications to infrastructures, climate, social systems and economics. *The European Physical Journal Special Topics*, 214(1):273–293, 2012.

[33] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 06 1998.

[34] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–1123, 01 2008.

[35] Martin Rosvall and Carl T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007.

[36] Tiago P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X*, 4:011047, Mar 2014.

[37] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[38] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[39] Sergei N. Dorogovtsev and José F. F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, 1st edition, 2003.

[40] Wolfram documentation center: Pseudodiameter. retrieved March 22, 2016 from: `http://reference.wolfram.com/language/GraphUtilities/ref/PseudoDiameter.html`.

[41] U. Kang, Charalampos Tsourakakis, Ana Paula Appel, Christos Faloutsos, and Jure Leskovec. Hadi: Fast diameter estimation and mining in massive graphs with hadoop. Technical report, School of Computer Science, Carnegie Mellon University Pittsburgh, December 2008.

[42] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 09 1999.

[43] Network science book. retrieved April 7th, 2016 from: `http://barabasilab.neu.edu/networksciencebook/downlPDF.html`.

[44] Stanley Milgram. The small world problem. *Psychology Today*, 67(1):61–67, 1967.

[45] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 33–42, New York, NY, USA, 2012. ACM.

[46] Benjamin Bercovitz. Viewing implicit social networks as bipartite graphs. 2009. CS322, Stanford University.

[47] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific Reports*, 2:541 EP –, 08 2012.

[48] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. Activity driven modeling of time varying networks. *Scientific Reports*, 2:469 EP –, 06 2012.

[49] Petter Holme. Modern temporal network theory: a colloquium. *Eur. Phys. J. B*, 88(9):234, 2015.

[50] F. Liljeros, J. Giesecke, and P. Holme. The contact network of inpatients in a regional healthcare system. a longitudinal case study. *Mathematical Population Studies*, 14(4):269–284, 2007.

[51] Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5):387–408, 2012.

[52] Nicola Perra, Andrea Baronchelli, Delia Mocanu, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Random walks and search in time-varying networks. *Phys. Rev. Lett.*, 109:238701, Dec 2012.

[53] Plamen Ch. Ivanov. Ideas about dynamics networks. personal communication, 09 2003.

[54] Jianxi Gao, Sergey V. Buldyrev, Shlomo Havlin, and H. Eugene Stanley. Robustness of a network of networks. *Phys. Rev. Lett.*, 107:195701, Nov 2011.

[55] Jianxi Gao, Sergey V. Buldyrev, H. Eugene Stanley, and Shlomo Havlin. Networks formed from interdependent networks. *Nat Phys*, 8(1):40–48, 01 2012.

[56] Roni Parshani, Sergey V. Buldyrev, and Shlomo Havlin. Critical effect of dependency groups on the function of networks. *Proceedings of the National Academy of Sciences*, 108(3):1007–1010, 2011.

[57] Sergey V. Buldyrev, Roni Parshani, Gerald Paul, H. Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 04 2010.

[58] Evangelia Mitleton-Kelly, editor. *Co-evolution of Intelligent Socio-technical Systems: Modelling and Applications in Large Scale Emergency and Transport Domains (Understanding Complex Systems)*, volume 1. Springer, 1st edition, 2013.

[59] Lingling Gao. Power-law decay of the view times of scientific courses on youtube. *Physica A: Statistical Mechanics and its Applications*, 391(22):5697 – 5703, 2012.

[60] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008.

[61] Jure Leskovec and Rok Sosič. SNAP: A general purpose network analysis and graph mining library in C++. `http://snap.stanford.edu/snap`, June 2014.

[62] Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, and Ion Stoica. Graphx: Graph processing in a distributed dataflow framework. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 599–613, Broomfield, CO, October 2014. USENIX Association.

[63] Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, Grzegorz Czajkowski, and Google Inc. Pregel: A system for large-scale graph processing. In *In SIGMOD*, pages 135–146, 2010.

[64] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239, 1978.

[65] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64:016131, Jun 2001.

[66] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[67] J. Heitzig, J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes. *The European Physical Journal B - Condensed Matter and Complex Systems*, 85(1):1–22, January 2012.

[68] Gang-Jin Wang, Chi Xie, Yi-Jun Chen, and Shou Chen. Statistical properties of the foreign exchange network at different time scales: Evidence from detrended cross-correlation coefficient and minimum spanning tree. *Entropy*, 15(5):1643, 2013.

[69] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10421–10426, 2005.

[70] T. Aste, T. Di Matteo, and S. T. Hyde. Complex networks on hyperbolic surfaces. *Physica A: Statistical Mechanics and its Applications*, 346(1–2):20 – 26, 2005. Statphys - Kolkata V: Proceedings of the International Conference on Statistical Physics: Complex Networks: Structure, Function and Processes.

[71] Guido Previde Massara, T. Di Matteo, and Tomaso Aste. Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, 2016.

[72] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.*, 42(1):181–213, 2015.

[73] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[74] 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, 2008.

[75] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April 1998.

[76] Robert Tarjan. Depth first search and linear graph algorithms. *SIAM Journal on Computing*, 1972.

[77] Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theor. Comput. Sci.*, 407(1-3):458–473, November 2008.

[78] Steve Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2010.

[79] Yehuda Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 426–434, New York, NY, USA, 2008. ACM.

[80] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész. Dynamic asset trees and black monday. *Physica A: Statistical Mechanics and its Applications*, 324(1):247–252, 2003.

[81] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E*, 68(5):56110, 2003.

[82] Arda Halu, Raúl J. Mondragón, Pietro Panzarasa, and Ginestra Bianconi. Multiplex pagerank. *PLoS ONE*, 8(10):e78293, 10 2013.

[83] Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R. Sugimoto. Do altmetrics work? twitter and ten other social web services. *PLoS ONE*, 8(5):1–7, 05 2013.

[84] George A. Lozano, Vincent Larivière, and Yves Gingras. The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*, 63(11):2140–2145, 2012.

[85] Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. The production of information in the attention economy. *Scientific Reports*, 5:9452 EP –, 05 2015.

[86] Ernesto Estrada. *The Structure of Complex Networks*. Oxford University Press, 2012.

[87] Roger Guimerà and Marta Sales-Pardo. Form follows function: the architecture of complex networks. *Molecular Systems Biology*, 2:42–42, 2006.

[88] Romina Cachia, Ramón Compañó, and Olivier Da Costa. Grasping the potential of online social networks for foresight. *Technological Forecasting and Social Change*, 74(8):1179 – 1203, 2007.

[89] Kun Zhao, Márton Karsai, and Ginestra Bianconi. Entropy of dynamical social networks. *PLOS ONE*, 6(12):1–7, 12 2011.

[90] Kartik Anand and Ginestra Bianconi. Entropy measures for networks: Toward an information theory of complex topologies. *Phys. Rev. E*, 80:045102, Oct 2009.

[91] Krzysztof Kulakowski. A note on temperature without energy - a social example. *n.a.*, November 2015. `https://arxiv.org/abs/0807.0711v2`.

[92] L. M. Floría, C. Gracia-Lázaro, J. Gómez-Gardeñes, and Y. Moreno. Social network reciprocity as a phase transition in evolutionary cooperation. *Phys. Rev. E*, 79:026106, Feb 2009.

[93] Dirk Helbing, Illes Farkas, and Tamas Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490, 09 2000.

[94] Michael J. Bannister, David Eppstein, Michael T. Goodrich, and Lowell Trott. Force-directed graph drawing using social gravity and scaling. In Walter Didimo and Maurizio Patrignani, editors, *Graph Drawing*, volume 7704 of *Lecture Notes in Computer Science*, pages 414–425. Springer Berlin Heidelberg, 2013.

[95] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21(11):1129–1164, 1991.

[96] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, Sep 2003.

[97] Tatsuro Kawamoto and Naomichi Hatano. An explosive diffusion on a social network, November 2012. http://arxiv.org/abs/1211.2555.

[98] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[99] Peter Sheridan Dodds, Kameron Decker 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):1–1, 12 2011.

[100] Emily M. Cody, Andrew J. Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M. Danforth. Climate change sentiment on twitter: An unsolicited public opinion poll. *PLoS ONE*, 10(8):1–18, 08 2015.

[101] Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors: Research articles. *Complexity*, 12(3):30–40, January 2007.

[102] Kevin W Boyack, Richard Klavans, and Katy Börner. Mapping the backbone of science. *Scientometrics*, 64:351–374, 2005.

[103] A. J. Reinoso, F. Ortega, J. M. Gonzalez-Barahona, and G. Robles. A quantitative approach to the use of the Wikipedia. In *IEEE Symposium on Computers and Communications (ISCC)*, pages 56–61. IEEE, July 2009.

[104] Won Kim, Ok-Ran Jeong, and Sang-Won Lee. On social web sites. *Information systems*, 35(2):215–236, 2010.

[105] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[106] Andreas Kaltenbrunner and David Laniado. There is no deadline: Time evolution of wikipedia discussions. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 6:1–6:10, New York, NY, USA, 2012. ACM.

[107] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in wikipedia. *PLoS ONE*, 7(6):e38869, 06 2012.

[108] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 575–582, New York, NY, USA, 2004. ACM.

[109] Jaap Kamps and Marijn Koolen. Is Wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 232–241, New York, NY, USA, 2009. ACM.

[110] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Phys. Rev. E*, 74:036116, Sep 2006.

[111] Bernd Ulmann. *Analog Computing*. Oldenbourg Wissenschaftsverlag, 2013.

[112] Aleks Aris, Ben Shneiderman, Catherine Plaisant, Galit Shmueli, and Wolfgang Jank. Representing unevenly-spaced time series data for visualization and interactive exploration. In MariaFrancesca Costabile and Fabio Paternò, editors, *Human-Computer Interaction - INTERACT 2005*, volume 3585 of *Lecture Notes in Computer Science*, pages 835–846. Springer Berlin Heidelberg, 2005.

[113] Alexander Mörtl, Tamara Lorenz, and Sandra Hirche. Rhythm patterns interaction - synchronization behavior for human-robot joint action. *PLoS ONE*, 9(4):1–17, 04 2014.

[114] Chris Chatfield. *The analysis of time series: an introduction*. CRC Press, Florida, US, 6th edition, 2004.

[115] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. In *In an Edited Volume, Data mining in Time Series Databases. Published by World Scientific*, pages 1–22. Publishing Company, 1993.

[116] Stephen Beveridge and Charles Nelson. A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary Economics*, 7(2):151–174, 1981.

[117] Taha Yasseri, Robert Sumi, and János Kertész. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PLoS ONE*, 7(1):e30091, 01 2012.

[118] K. J. Friston, O. Josephs, E. Zarahn, A. P. Holmes, S. Rouquette, and J.-B. Poline. To smooth or not to smooth?: Bias and efficiency in fmri time-series analysis. *NeuroImage*, 12(2):196 – 208, 2000.

[119] Yannick Malevergne, V. Pisarenko, and D. Sornette. Empirical distributions of stock returns: between the stretched exponential and the power law? *Quantitative Finance*, 5(4):379–401, 2005.

[120] S. Drozdz, J. Kwapien, F. Grümmer, F. Ruf, and J. Speth. Are the contemporary financial fluctuations sooner converging to normal? *Physica Polonica*, B34(8):4293–4306, 2003.

[121] M. Wiedermann, J. F. Donges, J. Heitzig, and J. Kurths. Node-weighted interacting network measures improve the representation of real-world complex systems. *EPL (Europhysics Letters)*, 102(2):28007, 2013.

[122] Jan W. Kantelhardt. *Fractal and Multifractal Time Series*, pages 463–487. Springer New York, New York, NY, 2011.

[123] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of dna nucleotides. *Phys Rev. E*, 49:1685, 1994.

[124] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A*, 295:441, 2001.

[125] A. Bunde, S. Havlin, J. W. Kantelhardt, T. Penzel, J. H. Peter, and K. Voigt. Correlated and uncorrelated regions in heart-rate fluctuations during sleep. *Phys Rev. Lett.*, 85:3736, 2000.

[126] Jan F. Eichner, Jan W. Kantelhardt, Armin Bunde, and Shlomo Havlin. Statistics of return interval in long-term correlated records. *Physical Review E*, 75, 2007.

[127] Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. Evidence for a bimodal distribution in human ciommunication. *PNAS*, 107:18803–18808, 2010.

[128] A. Bunde, J. F. Eichner, S. Havlin, and J. W. Kantelhardt. The effect of long-term correlations on the return periods of rare events. *Physica A*, 330:1–7, 2003.

[129] A. Bunde, J. F. Eichner, J. W. Kantelhardt, and S. Havlin. Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys Rev. Lett.*, 94:048701, 2005.

[130] Eduardo G. Altmann and Holger Kantz. Recurrence time analysis, long-term correlations, and extreme events. *Phys. Rev. E*, 71:056106, May 2005.

[131] Ernesto Pereda, Rodrigo Quian Quiroga, and Joydeep Bhattacharya. Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, 77(1-2):1–37, 2005.

[132] Radebach Alexander. Evolving climate networks: Investigating the evolution of correlation structure of the earth's climate system. Master's thesis, Humboldt-Universität zu Berlin, July 2010.

[133] Hauke Jan and Kossowski Tomasz. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, 30(2):87–93, June 2011.

[134] Berit Schreck. Rekonstruktion komplexer netzwerke mittels kreuzkorrelationsmethode. *Martin-Luther-Universität Halle-Wittenberg*, 2012. unpublished Bachelor thesis.

[135] R. Quian Quiroga, T. Kreuz, and P. Grassberger. Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Phys. Rev. E*, 66:041904, Oct 2002.

[136] N. Malik, N. Marwan, and J. Kurths. Spatial structures and directionalities in monsoonal precipitation over south asia. *Nonlinear Processes in Geophysics*, 17:371–381, September 2010.

[137] Nishant Malik, Bodo Bookhagen, Norbert Marwan, and Jürgen Kurths. Analysis of spatial and temporal extreme monsoonal rainfall over south asia using complex networks. *Climate Dynamics*, 39(3-4):971–987, 2012.

[138] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[139] Joseph Troy Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1(11), 2014.

[140] A. D. Wyner. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38(1):51 – 59, 1978.

[141] Ian T. Young. Proof without prejudice: Use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. *J. Histochem. Cytochem*, pages 935–941, July 1977.

[142] Andy Field. *Discovering statistics using SPSS*. Number 978-1-84787-906-6. SAGE Publications, Los Angeles, 3rd edition, 2009.

[143] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, November 2009.

[144] E. S. Pearson and H. O. Hartley, editors. *Biometrika Tables for Statisticians Biometrika Tables for Statisticians*, volume 1. 3rd edition, Jan 1966.

[145] Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[146] Theiler J., Eubank S., Longtin A., Galdrikian B., and Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1):77–94, 1992.

[147] B. Podobnik, F. D. Fu, E. H. Stanley, and Ch. P. Ivanov. Power-law autocorrelated stochastic processes with long-range cross-correlations. *The European Physical Journal B*, 56(1):47–52, 2007.

[148] Commons-math: The apache commons mathematics library. retrieven 2012, July from: `http://commons.apache.org/math`.

[149] Makoto Matsumoto and Takuji Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30, January 1998.

[150] Dutang Christophe and Savicky Petr. *randtoolbox: Generating and Testing Random Numbers*, 2015. R package version 1.17.

[151] Hernán Makse, Shlomo Havlin, H. Eugene Stanley, and Moshe Schwartz. Novel method for generating long-range correlations. *Chaos, Solitons and Fractals*, 6:295 – 303, 1995. Complex Systems in Computational Physics.

[152] Thomas Schreiber and Andreas Schmitz. Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.*, 77:635–638, Jul 1996.

[153] Fan Chung, Linyuan Lu, and Van Vu. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 100(11):6313–6318, 2003.

[154] Yingjie Hu. Citation map: Visualizing the spread of scientific ideas through space and time, 2013. retrieved January, 2016 from: `http://www.escholarship.org/uc/item/0k09p3zw`.

[155] Easley David and Kleinberg Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[156] Marcus P. Zillman. Deep web report 2009, 2009. retrieved 2016, December from: `http://www.llrx.com/2008/12/deep-web-research-2009`.

[157] Wiki-java : A java wiki bot framework. retrieved August, 2014 from: Github `https://github.com/MER-C/wiki-java`.

[158] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *In 6th Int'l Semantic Web Conference, Busan, Korea*, pages 11–15. Springer, 2007.

[159] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[160] Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[161] *Wikibooks*, 2015 (accessed February 2, 2015). `http://en.wikibooks.org/wiki/Main_Page`.

[162] *Wiktionary*, 2015 (accessed February 2, 2015). `http://en.wiktionary.org/wiki/Wiktionary:Main_Page`.

[163] *Wikipedia:Tools/Alternative Browsing*, 2015 (accessed February 2, 2015). `http://en.wikipedia.org/wiki/Wikipedia:Tools/Alternative_browsing`.

[164] Boeker Arne. Reconstruction of complex networks based on event time series. *Martin-Luther-Universität Halle-Wittenberg*, 2012. unpublished Bachelor thesis.

[165] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[166] J. O'Madadhain, D. Fisher, S. White, and Y. Boey. The JUNG (Java Universal Network/Graph) Framework. Technical report, UCI-ICS, October 2003.

[167] Jure Leskovec and Rok Sosič. Snap.py: SNAP for Python, a general purpose network analysis and graph mining tool in Python. `http://snap.stanford.edu/snappy`, June 2014.

[168] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[169] Alexander Hinneburg, Frank Rosner, Stefan Peßler, and Christian Oberländer. Exploring document collections with topic frames. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 2084–2086, 2014.

[170] Geert Hofstede. *Culture's consequences: International differences in work-related values*, volume 5. SAGE Publications, 1984.

[171] Michael D. Ekstrand and John T. Riedl. rv you're dumb: Identifying discarded work in wiki article history. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 4:1–4:10, New York, NY, USA, 2009. ACM.

[172] Felipe Ortega, Jesús M. González-Barahona, and Gregorio Robles. The top-ten wikipedias - A quantitative analysis using wikixray. In *ICSOFT 2007, Proceedings of the Second International Conference on Software and Data Technologies, Volume ISDM/EHST/DC, Barcelona, Spain, July 22-25, 2007*, pages 46–53, 2007.

[173] Jeff Stuckman and James Purtilo. Measuring the wikisphere. In *Proceedings of the 2009 International Symposium on Wikis, 2009, Orlando, Florida, USA, October 25-27*, 2009.

[174] Arnaud Gorgeon and E. Burton Swanson. Organizing the vision for web 2.0: a study of the evolution of the concept in wikipedia. In Dirk Riehle and Amy Bruckman, editors, *Int. Sym. Wikis*. ACM, 2009.

[175] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar 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, 08 2011.

[176] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[177] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 09 2013.

[178] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9):938–949, 2015.

[179] George Kampis, Jan W. Kantelhardt, Kamil Kloch, and Paul Lukowicz. Analytical and simulation models for collaborative localization. *Journal of Computational Science*, 6:1 – 10, 2015.

[180] Martin Wirz, Tobias Franke, Daniel Roggen, Eve Mitleton-Kelly, Paul Lukowicz, and Gerhard Tröster. Inferring and visualizing crowd conditions by collecting gps location traces from pedestrians' mobile phones for real-time crowd monitoring during city-scale mass gatherings. In *Collaborative Technology for Coordinating Crisis Management (CT2CM) track of WETICE-2012*, 2012.

[181] Martin Wirz, Tobias Franke, Eve Mitleton-Kelly, Daniel Roggen, Paul Lukowicz, and Gerhard Tröster. Coenosense: A framework for real-time detection and visualization of collective behaviors in human crowds by tracking mobile devices. In *Proceedings of European Conference on Complex Systems*. Springer, 2012.

[182] K. Kulakowski and M. Nawojczyk. Sociophysics - an astriding science. *ArXiv e-prints*, May 2008. `https://arxiv.org/abs/0805.3886`.

[183] Joao Gama Oliveira and Albert-Laszlo Barabasi. Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251–1251, 10 2005.

[184] Sauro Succi. *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond (Numerical Mathematics and Scientific Computation)*. Numerical mathematics and scientific computation. Oxford University Press, 1 edition, August 2001.

[185] Mirko Kämpf. Poster: Simulation of information flow on dynamic interlinked networks. available online: `http://www.slideshare.net/mirkokaempf/information-spread-in-the-context-of-evacuation-optimization`.

[186] Mirko Kämpf. Presentation: Simulation of information flow on dynamic interlinked networks. available online: `http://de.slideshare.net/mirkokaempf/dpg-berlin-soe-18-talk-v124`.

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

[188] Lailil Muflikhah and Baharum Baharudin. Document clustering using concept space and cosine similarity measurement. In *Computer Technology and Development, 2009. ICCTD'09. International Conference on*, volume 1, pages 58–62. IEEE, 2009.

[189] Petre Caraiani. Using complex networks to characterize international business cycles. *PLoS ONE*, 8(3):e58109, 03 2013.

[190] Pierre-Olivier Amblard and Olivier J. J. Michel. On directed information theory and granger causality graphs. *Journal of Computational Neuroscience*, 30(1):7–16, 2011.

[191] A. P. Masucci, A. Kalampokis, V. M. Eguíluz, and E. Hernández-García. Extracting directed information flow networks: An application to genetics and semantics. *Phys. Rev. E*, 83:026103, Feb 2011.

[192] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luis A. Nunes Amaral, Thomas Guhr, and H. Eugene Stanley. Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 65:066126, Jun 2002.

[193] Philip Ball. *Why Society is a Complex Matter: Meeting Twenty-First Century Challenges with a New Kind of Science*. Springer Berlin Heidelberg Springer Berlin Heidelberg Springer Berlin Heidelberg, 2012.

[194] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[195] R. Cohen and S. Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, 2010.

[196] Mark E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[197] Xin Liu and Tsuyoshi Murata. Detecting communities in k-partite k-uniform (hyper) networks. *Journal of Computer Science and Technology*, 26(5):778–791, 2011.

[198] Tsuyoshi Murata. Detecting communities from tripartite networks. In *Proceedings of the 19th international conference on World wide web*, pages 1159–1160. ACM, 2010.

[199] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[200] Benjamin Renoust, Guy Melançon, and Tamara Munzner. Detangler: Visual analytics for multiplex networks. In *Computer Graphics Forum*, volume 34, pages 321–330. Wiley Online Library, 2015.

[201] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2012.

[202] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[203] Wael H. Gomaa and Aly A. Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18, April 2013. Full text available.

[204] Samer Hassan and Rada Mihalcea. Semantic relatedness using salient semantic analysis. *AAAI Conference on Artificial Intelligence*, 2011.

[205] A. Islam and D. Inkpen. Second order co-occurrence pmi for determining the semantic similarity of words. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1033–1038, 2006.

[206] Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills. *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*. O'Reilly Media, 2015.

[207] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[208] Diego R. Amancio, Osvaldo N. Oliveira Jr., and Luciano da F. Costa. Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Physica A: Statistical Mechanics and its Applications*, 391(18):4406 – 4419, 2012.

[209] Markus M. Geipel. Self-organization applied to dynamic network layout. *International Journal of Modern Physics C*, 18(10):1537–1549, 2007.

[210] Dror Y. Kenett, Yoash Shapira, Asaf Madi, Sharron Bransburg-Zabary, Gitit Gur-Gershgoren, and Eshel Ben-Jacob. Index cohesive force analysis reveals that the us market became prone to systemic collapses since 2002. *PLoS ONE*, 6(4):1–8, 04 2011.

[211] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.*, 105:158701, Oct 2010.

[212] Lawrence Mitchell and Michael E Cates. Hawkes process as a model of social interactions: a view on video dynamics. 43(4):11, Jan 2010.

[213] Xitong Li and Lynn Wu. Herding and social media word-of-mouth: Evidence from groupon. *Available at SSRN 2264411*, 2014. `http://dx.doi.org/10.2139/ssrn.2264411`.

[214] Dieter Nautz. Herding in financial markets: Bridging the gap between theory and evidence. Working Papers 2013002, Berlin Doctoral Program in Economics and Management Science (BDPEMS), 2013.

[215] Honglin Yu, Lexing Xie, and Scott Sanner. The lifecyle of a youtube video: Phases, content and popularity. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 533–542, 2015.

[216] P. G. Lind and H. J. Herrmann. New approaches to model and study social networks. *New Journal of Physics*, 9(7):228, 2007.

[217] D. Sornette. *Critical Phenomena in Natural Sciences*. Springer-Verlag, Berlin Heidelberg, 2006.

[218] Diego Rybski, Sergey V. Buldyrev, Shlomo Havlin, Fredrik Liljeros, and Hernán A. Makse. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences*, 106(31):12640–12645, 2009.

[219] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[220] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68:065103, Dec 2003.

[221] Gábor Csányi and Balázs Szendrői. Structure of a large social network. *Phys. Rev. E*, 69:036131, Mar 2004.

[222] Sergi Valverde and Ricard V. Solé. Self-organization versus hierarchy in open-source social networks. *Phys. Rev. E*, 76:046118, Oct 2007.

[223] Lazaros K. Gallos, Chaoming Song, and Hernán A. Makse. Scaling of degree correlations and its influence on diffusion in scale-free networks. *Phys. Rev. Lett.*, 100:248701, Jun 2008.

[224] Mária Ercsey-Ravasz and Zoltán Toroczkai. Centrality scaling in large networks. *Phys. Rev. Lett.*, 105:038701, Jul 2010.

[225] M. Argollo de Menezes and A.-L. Barabási. Fluctuations in network dynamics. *Phys. Rev. Lett.*, 92:028701, Jan 2004.

[226] A. Grabowski, N. Kruszewska, and R. A. Kosiński. Properties of on-line social systems. *The European Physical Journal B*, 66(1):107–113, 2008.

[227] D. Markovic and C. Gros. Vertex routing models. *New Journal of Physics*, 11(7):073002, 2009.

[228] Juliette Stehlé, Alain Barrat, and Ginestra Bianconi. Dynamical and bursty interactions in social networks. *Phys. Rev. E*, 81:035101, Mar 2010.

[229] V. Zlatic, M. Bozccevice, H. Stefancice, and M. Domazet. Phys. *Rev. E*, 74:016115, 2006.

[230] Lev Muchnik, Royi Itzhack, Sorin Solomon, and Yoram Louzoun. Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies. *Phys. Rev. E*, 76:016106, Jul 2007.

[231] Myshkin Ingawale, Amitava Dutta, Rahul Roy, and Priya Seetharaman. The small worlds of wikipedia: Implications for growth, quality and sustainability of collaborative knowledge networks. 2009.

[232] Rut Jesus, Martin Schwartz, and Sune Lehmann. Bipartite networks of wikipedia's articles and authors: A meso-level approach. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 5:1–5:10, New York, NY, USA, 2009. ACM.

[233] A. Lih. *The Foundations of Participatory Journalism and the Wikipedia Project*. AEJMC (Toronto), Canada, 2004.

[234] A. Lih. Wikipedia as participatory journalism: Reliable sources? In *Metrics for evaluating collaborative media as a news resource*, Texas, 2005. Proc. 5th Internat. Symposium on Online Journalism (Austin).

[235] Sonya Lipczynska. Power to the people: the case for wikipedia. *Reference Reviews*, 19(2):6–7, 2005.

[236] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon. Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Phys. Rev. Lett.*, 93:228701, Nov 2004.

[237] F. Deschâtres and D. Sornette. Dynamics of book sales: Endogenous versus exogenous shocks in complex networks. *Phys. Rev. E*, 72:016112, Jul 2005.

[238] R. Lambiotte and M. Ausloos. Endo- vs. exogenous shocks and relaxation rates in book and music "sales". *Physica A: Statistical Mechanics and its Applications*, 362(2):485 – 494, 2006.

[239] Jiang, Z.-Q., Guo, L., and Zhou, W.-X. Endogenous and exogenous dynamics in the fluctuations of capital fluxes - an empirical analysis of the chinese stock market. *Eur. Phys. J. B*, 57(3):347–355, 6 2007.

[240] D. Sornette and S. Utkin. Limits of declustering methods for disentangling exogenous from endogenous events in time series with foreshocks, main shocks, and aftershocks. *Phys. Rev. E*, 79:061110, Jun 2009.

[241] D. Sornette, V. I. Yukalov, E. P. Yukalova, J.-Y. Henry, D. Schwab, and J. P. Cobb. Endogenous versus exogenous origins of diseases. *Journal of Biological Systems*, 17(02):225–267, 2009.

[242] Lawrence Mitchell and Michael E Cates. Hawkes process as a model of social interactions: a view on video dynamics. *Journal of Physics A: Mathematical and Theoretical*, 43(4):045101, 2010.

[243] A. I. Saichev and D. Sornette. Generation-by-generation dissection of the response function in long memory epidemic processes. *The European Physical Journal B*, 75(3):343–355, 2010.

[244] A. Bunde, J. Kropp, and H. J. Schellnhuber. In *The science of disasters*. Springer, Berlin, 2002.

[245] Lorenz-M. Stadler, Bogdan Sepiol, Bastian Pfau, Jan W. Kantelhardt, Richard Weinkamer, and Gero Vogl. Detrended fluctuation analysis in x-ray photon correlation spectroscopy for determining coarsening dynamics in alloys. *Phys. Rev. E*, 74:041107, Oct 2006.

[246] Zhu Xiao-Yan, Liu Zong-Hua, and Tang Ming. Detrended fluctuation analysis of traffic data. *Chinese Physics Letters*, 24(7):2142, 2007.

[247] J. J. Wu, H. J. Sun, and Z. Y. Gao. Long-range correlations of density fluctuations in the kerner-klenov-wolf cellular automata three-phase traffic flow model. *Phys. Rev. E*, 78:036103, Sep 2008.

[248] Aicko Y. Schumann, Ronny P. Bartsch, Thomas Penzel, Plamen Ch. Ivanov, and Jan W. Kantelhardt. Aging effects on cardiac and respiratory dynamics in healthy subjects across sleep stages. 07 2010.

[249] J. F. Eichner, E. Koscielny-Bunde, A. Bunde, S. Havlin, and H.-J. Schellnhuber. Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. *Phys. Rev. E*, 68:046133, Oct 2003.

[250] D. Vyushin, I. Zhidkov, S. Havlin, A. Bunde, and S. Brenner. Reply to comment by R. Blender and K. Fraedrich on "Volcanic forcing improves atmosphere-ocean coupled general circulation model scaling performance". *Geophysical Research Letters*, 31, 2004.

[251] Jan W. Kantelhardt, Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H. Eugene Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, 316(1–4):87–114, 2002.

[252] Josef Ludescher, Mikhail I. Bogachev, Jan W. Kantelhardt, Aicko Y. Schumann, and Armin Bunde. On spurious and corrupted multifractality: The effects of additive noise, short-term memory and periodic trends. *Physica A: Statistical Mechanics and its Applications*, 390(13):2480 – 2490, 2011.

[253] Aicko Y. Schumann and Jan W. Kantelhardt. Multifractal moving average analysis and test of multifractal model with tuned correlations. *Physica A: Statistical Mechanics and its Applications*, 390(14):2637–2654, 2011.

[254] Gao-Feng Gu and Wei-Xing Zhou. Detrended fluctuation analysis for fractals and multifractals in higher dimensions. *Phys. Rev. E*, 74:061104, Dec 2006.

[255] Boris Podobnik and H. Eugene Stanley. Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Phys. Rev. Lett.*, 100:084102, Feb 2008.

[256] Kun Hu, Plamen Ch. Ivanov, Zhi Chen, Pedro Carpena, and H. Eugene Stanley. Effect of trends on detrended fluctuation analysis. *Phys. Rev. E*, 64:011114, Jun 2001.

[257] Zhi Chen, Plamen Ch. Ivanov, Kun Hu, and H. Eugene Stanley. Effect of nonstationarities on detrended fluctuation analysis. *Phys. Rev. E*, 65:041107, Apr 2002.

[258] Didier Delignieres, Sofiane Ramdani, Loïc Lemoine, Kjerstin Torre, Marina Fortes, and Grégory Ninot. Fractal analyses for 'short' time series: A re-assessment of classical methods. *Journal of Mathematical Psychology*, 50(6):525 – 544, 2006.

[259] M. S. Santhanam and Holger Kantz. Return interval distribution of extreme events and long-term memory. *Phys. Rev. E*, 78:051113, Nov 2008.

[260] N. R. Moloney and J. Davidsen. Extreme value statistics and return intervals in long-range correlated uniform deviates. *Phys. Rev. E*, 79:041131, Apr 2009.

[261] Mikhail I. Bogachev, Jan F. Eichner, and Armin Bunde. Effect of nonlinear correlations on the statistics of return intervals in multifractal data sets. *Phys. Rev. Lett.*, 99:240601, Dec 2007.

[262] M. I. Bogachev, J. F. Eichner, and A. Bunde. The effects of multifractality on the statistics of return intervals. *The European Physical Journal Special Topics*, 161(1):181–193, 2008.

[263] M. S. Santhanam and Holger Kantz. Long-range correlations and rare events in boundary layer wind fields. *Physica A: Statistical Mechanics and its Applications*, 345(3–4):713–721, 2005.

[264] Plamen Ch. Ivanov, Ainslie Yuen, Boris Podobnik, and Youngki Lee. Common scaling patterns in intertrade times of u. s. stocks. *Phys. Rev. E*, 69:056107, May 2004.

[265] Kazuko Yamasaki, Lev Muchnik, Shlomo Havlin, Armin Bunde, and H. Eugene Stanley. Scaling and memory in volatility return intervals in financial markets. *Proceedings of the National Academy of Sciences of the United States of America*, 102(26):9424–9428, 2005.

[266] Boris Podobnik, Davor Horvatic, Alexander M. Petersen, and H. Eugene Stanley. Cross-correlations between volume change and price change. *Proceedings of the National Academy of Sciences*, 106(52):22079–22084, 2009.

[267] Mikhail I. Bogachev, Igor S. Kireenkov, Eugene M. Nifontov, and Armin Bunde. Statistics of return intervals between long heartbeat intervals and their usability for online prediction of disorders. *New Journal of Physics*, 11(6):063036, 2009.

[268] M. I. Bogachev and A. Bunde. On the occurrence and predictability of overloads in telecommunication networks. *EPL (Europhysics Letters)*, 86(6):66002, 2009.

[269] Shi-Min Cai, Zhong-Qian Fu, Tao Zhou, Jun Gu, and Pei-Ling Zhou. Scaling and memory in recurrence intervals of internet traffic. *EPL (Europhysics Letters)*, 87(6):68001, 2009.

[270] Da-Hai Tang, Xiao-Pu Han, and Bing-Hong Wang. Stretched exponential distribution of recurrent time of wars in china. *Physica A: Statistical Mechanics and its Applications*, 389(13):2637 – 2641, 2010.

[271] Wiki: Xray. retrieved December, 2011 from: `http://meta.wikimedia.org/w/index.php?title=WikiXRay&oldid=1585263`.

[272] Wikistatistics. retrieved December, 2011 from: `http://en.wikipedia.org/wiki/Wikipedia:Statistics`.

[273] Wikipedia page counters. retrieved December, 2016 from: `https://dom.as/2007/12/10/wikipedia-page-counters/`.

[274] Wikimedia downloads. retrieved December, 2014 from: `http://dumps.wikimedia.org`.

[275] Jørgen Vitting Andersen, Andrzej Nowak, Giulia Rotundo, Lael Parrott, and Sebastian Martinez. "price-quakes" shaking the world's stock exchanges. *PLoS ONE*, 6(11):1–8, 11 2011.

[276] A. J. Morales, J. C. Losada, and R. M. Benito. User structure and behavior on an online social network during a political protest. *Physica A*, 391:5244–5253, 2012.

[277] Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165 – 4180, 2012.

[278] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[279] Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1:37–37, 2007.

[280] Andrei Kirilenko, Mehrdad Samadi, Albert S. Kyle, and Tuzun Tugkan. The flash crash: The impact of high frequency trading on an electronic market. *Journal of Finance, Forthcoming. Available at SSRN*, 2011. Electronic copy available at: `http://ssrn.com/abstract=1686004`.

[281] Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, 3:1684 EP –, 04 2013.

[282] Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific Reports*, 3:3578 EP –, 12 2013.

[283] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[284] Anderson Jenny and Dash Eric. Struggling lehman plans to lay off 1,500. retrieved August, 2009 from: `http://dealbook.nytimes.com/2008/08/28/struggling-lehman-plans-to-lay-off-1500/?_r=0`, Archived from the original on September 1, 2008, The New York Times.

[285] Afp: Lehman brothers in freefall as hopes fade for new capital. retrieved July, 2010 from: `http://www.lankabusinessonline.com/news/lehman-brothers-in-freefall-as-hopes-fade-for-new-capital/583886777`.

[286] Dror Y. Kenett. *Physics inspired investigation of Finacial markets*. PhD thesis, Raymond and Beverly Sackler School of Physics and Astronomy, 2012.

[287] Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, and Ingmar Weber. Web search queries can predict stock market volumes. *PLoS ONE*, 7(7):e40014, 07 2012.

[288] Yahoo! financial services. retrieved November, 2014 from: `http://finance.yahoo.com/`.

[289] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D. Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):1–16, 2012.

[290] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports*, 3:1801 EP –, 05 2013.

[291] Jan F. Eichner, Jan W. Kantelhardt, Armin Bunde, and Shlomo Havlin. Statistics of return intervals in long-term correlated records. *Phys. Rev. E*, 75:011128, Jan 2007.

[292] Boris Podobnik and H. Eugene Stanley. Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Phys. Rev. Lett.*, 100:084102, Feb 2008.

Part IV.

Appendix: Further Applications

The last part of this work contains four applications of the discussed methods from multiple disciplines.

First, we show details of a market study, focused on the emerging Big Data market in the time range 2009 to 2011. Final results were published in [4]. Because the article is available as full text online it is not reproduced here, but rather additional unpublished (preliminary) results are shown in chapter 15. We used the time-resolved relevance index to study phases in the life cycle of economic systems, and the relation between financial markets and social media applications.

The conference paper, submitted to and presented at the European Conference for Complex Systems (ECCS 2012) is reproduced in chapter 16.

In chapter 17 we apply DFA and RIS to simulation results obtained from a crowd simulation using the social force model.

Finally, in chapter 18, cross-correlation analysis and DFA are used to study the traffic flow on a highway and for comparison of real world data with simulation results. This allows us to identify limitations of the simulation approach and to identify traffic phases without predefined rules.

One common aspect among all the applications is the fact, that based on many individual time series the collective properties of a complex system can be derived and studied. Different ways of combining such information (from primary data) by averaging without or with utilization of topological information, lead to a new type of data, called secondary data. Secondary data is not measured or collected directly, but rather the result of previous analysis steps. Especially, based on categorical variables, such as life cycle phases, it is possible to integrate this kind of data with many different systems from multiple disciplines using classification and labeling algorithms.

15. Social Media Driven Market Studies

This chapter contains unpublished material, which was created during the work on the article: *'The detection of emerging trends using Wikipedia traffic data and context networks'*published in the journal PLOS ONE (see [4] in my publication list).

The global financial crisis, which started in 2008 with the bankruptcy of *Lehman Brothers*, can be seen as one of the reasons why systemic risk has received much more awareness in recent years. Andersen *et al.* [275] emphasize the fact that *"the excessive risk taking by major financial institutions pushed the world's financial system into what many considered a state of near systemic failure in 2008."* Furthermore, they argue that *"the IMF for example in its yearly 2009 Global Financial Stability Report acknowledged the lack of proper tools and research on the topic."* This leads to the question, how such disruptions can be identified before they are propagated across globally connected financial markets.

Further, a description of such a process' dynamics is still an important but unsolved problem. Methods from complex systems research are considered to be appropriate, especially as interconnected financial markets can be modeled as complex networks.

This allows combined studies using empirical data to derive analytical models, and simulations also known as '*stress test*', which are conducted using computational and simulation techniques to test several influencing factors like changes in unemployment rates, interest rates, or GDP[1]. There is a variety of research questions which can be addressed via less specific and less complicated approaches, e.g., market studies based on customer surveys or publicly available statistical data. More and more companies use such data sets for market analysis.

Market studies can be done in many different ways with different focus and different data collection techniques. Two different types are considered here. The first example is based on one selected economic sector, which is defined by entities of different types, like companies, technologies, products, and related community projects. All those entities define a market, in our case "the emerging Big Data market". The second example primarily consists of entities which are all of the same type, in this case companies for which stocks are traded in international stock exchanges and market indices, which can be seen as groups of companies.

[1] GDP: According to Wikipedia, the gross domestic product is a monetary measure of the market value of all final goods and services produced in a period (quarterly or yearly) in a region.

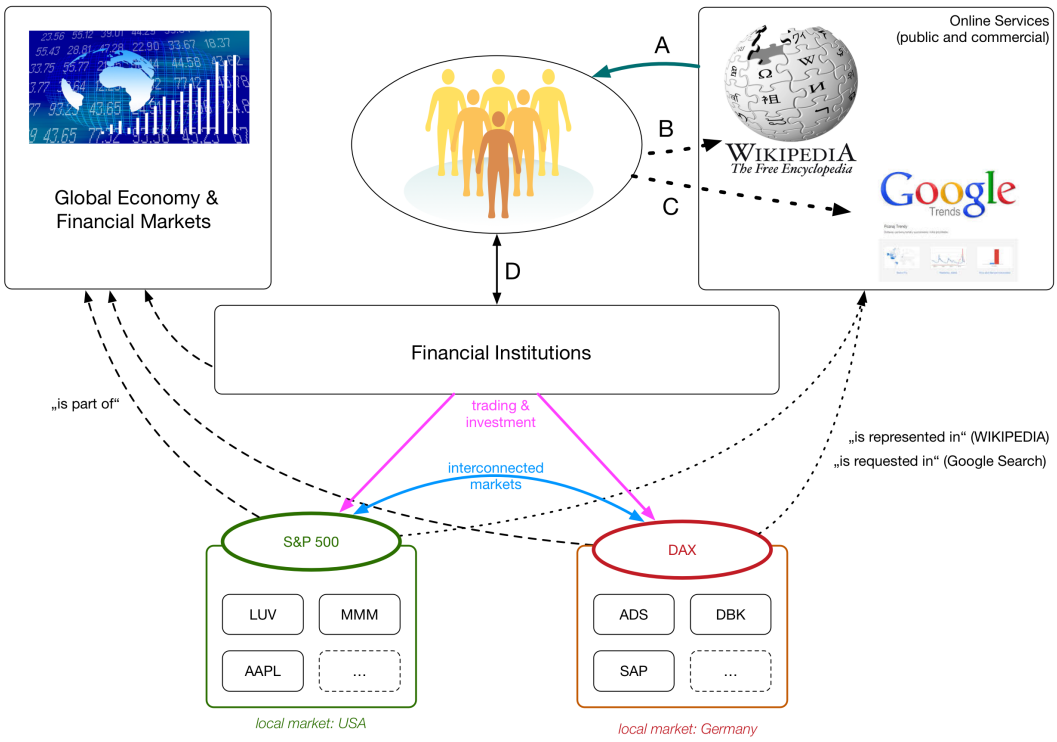

Figure 15.1.: **Simplified schema of interconnected international financial markets**. The network of networks, which is formed by international financial stock markets consists of many different but highly interconnected components, which are represented by stock indices. Stock indices are logical groups of traded stocks, which represent a certain facet of a geographically localized market. According to Kenett *et al.* [23, 210] the structure of interconnected international financial markets can be reconstructed from trading data like price and trading volume. Correlation measures can be calculated from trading data like closing prices or log return of prices (see section 5.2.5). In addition, we use access-rate time series from online services such as Wikipedia.

In our case, we want to find out, how well several companies are represented in Wikipedia. This allows a qualitative interpretation of collected access-rate data, which for itself should be used as an indicator for user interest in the pages, representing the topic. Large-scale studies from international markets can be based on this data collection approach only if the influence of the languages can be understood and if data can be normalized.

Dynamic properties like size, growth rate, connectivity between markets, or even public representation in several media channels can be derived form public data sets. Can data from Wikipedia or Google Trends be used as a reliable source for market models? This problem is investigated here. Methods from previous chapters, like, e.g., calculation of time-resolved relevance index (see Eq. 11.5 and 11.6 in section 11.3) are applied. The results address known limiting factors. A new approach for data set reliability assessment is provided.

15.1. What Wikipedia & Google Trends Tell About Market Dynamics

Financial markets can be seen as examples of highly interconnected systems with a variety of obvious facets or observable variables and also many hidden layers of non obvious relations. Many studies have investigated either structural [276, 60, 37, 277] or dynamical properties [278, 279] also using mathematical models and tools which have been developed in the context of climate research [26, 28, 132] and physiological research [22]. Nowadays, as more and more large data sets are publicly available to researchers, a convergence of such methods can be recognized and a set of key measures has been established, but those measures often over emphasize the structural properties of underlying networks (see figure 15.1), and ignore their complex dynamics and also their embedding into other surrounding networks, e.g., communication networks, as shown in figure 3.4 in chapter 3.

Financial market networks do not only contain internal links or connections. Very important are also relations to external systems like media companies, resource markets, technology companies, and maybe most importantly to the society in general. Such markets cannot function without information flow. Trading decisions are usually based on information, which is available to traders. A trader can be a person, acting on a time scale of minutes down to seconds, but not much faster. Since stock market transactions can also be done via automatic or semiautomatic

processes, therefore trading activity can be measured on a time scale on the order of milliseconds. This kind of automatic trading is called *'high frequency trading'*, and it is out of the scope of this work. A study by Kirilenko *et al.* [280] analyses the market events on May 6-th, 2010 which is called: *'The Flash Crash'*. Even if (according to Kirilenko *et al.* [280]) the high frequency trading cannot be seen as the reason for this crash, one has to notice that such automatic trading systems can probably influence markets, especially as long as not much detailed information about applied trading strategies is available publicly.

The majority of people, no matter what there role or position is, consume a lot of information via peer to peer communication, nowadays based on Internet applications like messaging services, provided by different complementary communication networks like Email services, Twitter, Facebook, Google+, commercial financial service portals, or public web pages like the encyclopedia Wikipedia. By analyzing the representation of stock markets in Wikipedia and measuring correlations between stock market data (like trading volume, or volatility) and the access-rates to corresponding groups of Wikipedia articles one can study the role of a protruding social network in the economic cycle. Can a measurable increase of interest in a special topic represented by a news article or a Wikipedia page be used as an indicator for changes in demand in financial markets? Questions like this one are important for individual decision makers but can we also identify a global state of markets and their dynamics based on such time series analysis?

A recent study by Preis *et. al.* [281] showed that following query volume for financial search terms on Google could predict stock market movement. Another study by Alanyali *et. al.* [282] demonstrates a significant correlation between the daily mentions of companies in the Financial Times in the morning and how much they were traded on the stock market during the day. Those results support the hypothesis of an existing mutual influence between financial markets and the news which is also illustrated schematically in figure 15.2.

Furthermore, it was previously shown that increases in the number of searches for a company name made on Google are correlated with increases in trading volume for that stock [281]. More importantly, it has been demonstrated that during the period of eight years (2004 to 2011), increases in searches for financially related terms tended to be followed by decreases in the Dow Jones Index Average [281]. This finding is in line with the proposal that Google search data may provide insight into the process of traders seeking information to help them determine optimum future decisions.

Since data from Google Trends is heavily used also in other domains, such as analysis of epidemic dynamics - see Google Flu Trends (GFT)[2] - one can clearly see the value of such type of analysis. But on the other hand a critical reflection and a quality discussion is more important than just processing more and more data. Lazer *et al.* [283] discuss typical problems and show that typical mistakes like overfitting a small number of cases, temporal auto-correlation, which leads to non randomly distributed errors, and a lag of stability of the applied method are critical factors, which all lead to wrong models. For example, in case of Google Flu, the flu prevalence has been overestimated in 100 weeks out of 108 starting in August 2011 (see figure 1 in [283]).

However, other online data sources may also provide insight into trader information gathering processes. Whilst many Internet users rely on Google to locate a range of different useful information sources online, the online encyclopedia Wikipedia is a widely-used central reference source for information across a number of subjects. As such, one can consider Google as a provider of data that gives insights into what information Internet users are looking for, whereas Wikipedia data provides insights into what information Internet users in fact use. Thus, we have investigated whether changes in frequency of views of certain Wikipedia pages also anticipate subsequent changes in stock market prices.

A few key practical differences exist between data on Wikipedia usage and data on Google search keyword usage. Firstly, some search terms have multiple meanings. For example, the term "Apple" is widely recognized both as the fruit and as the technology company. Google data, as retrieved from Google Trends[3], provides little insight into which meaning was of interest to the Internet user. Recent versions of Google Trends software provides a list of related searches, but a clear semantic context is not provided. In contrast, a Wikipedia page, other than those designed specifically for disambiguation, is about one topic only. In some cases a pages like the page with title "SOLR" (from English Wikipedia project) redirects to another one, here to a page with title '*Apache SOLR*'. Even if a user is not aware of the specific title, its interest will be counted and the relevant information is provided. The influence of keyword selection will be discussed and illustrated in a later section of this chapter. Disambiguation pages make it possible to consider changes in the number of views of the page about Apple the company separately to changes in the number of views of the page about apple the fruit. Redirect pages provide an implicit aggregation of several click trajectories.

Secondly, while data on Google usage relates to per-week changes in search volume, we are able to access data on hourly changes in Wikipedia usage. Thirdly, Wikipedia data describing access to each page across the Wikipedia encyclopedia since 2007 is freely available, whereas some restrictions exist on accessing large volumes of Google usage data.

We are interested in the role of Wikipedia as a public crowd based source for news in the context of market activity and we are especially focused on the readers side, because the number of article downloads reflects the

[2] `https://www.google.org/flutrends/about/`
[3] `https://www.google.de/trends/`

state of a larger part of the society which can be less influenced by the opinion of a single publisher using a shiny picture or provocative headline on page one of a newspaper. Readers select articles intentionally and they are not flooded by topics, which just sell well because Wikipedia is not a commercial system.

15.1.1. Description of the Analysis Procedure

The proposed method for a data-driven social media based market analysis has been built on the assumption that there is a direct relationship between trends in emerging markets especially in technology oriented markets, or between movements in financial markets and changes in the user activity, e.g., in Wikipedia or Google search volume. The approach combines a qualitative description and a quantitative analysis of connected information flows, which are manifested in Wikipedia content, stock market prices, and in usage data, which is derived from server logs by converting individual page request log messages into access-rate time series. We apply time series analysis paired with network analysis. Such systems and their possible interactions are illustrated in figure 15.2 where system **A** represents Wikipedia, which consists of many linked articles (*i*,*j*, and *k*) organized hierarchically in categories (which are omitted here for simplicity). Wikipedia works as a data collection stub and represents the process of personal information retrieval. Our current approach covers two types of systems about which people collect information via Wikipedia (system B, and system C). System **B** represents an emerging market, which has not yet a well defined nor a stable structure. Typically, the amount of information about an emerging market is expected to grow over time, and it is also subject of discussions. System **C** is the stock market, represented by stocks, traded on different exchanges. The stocks, indices, and exchanges are for itself represented in Wikipedia as wiki pages (system **A**).

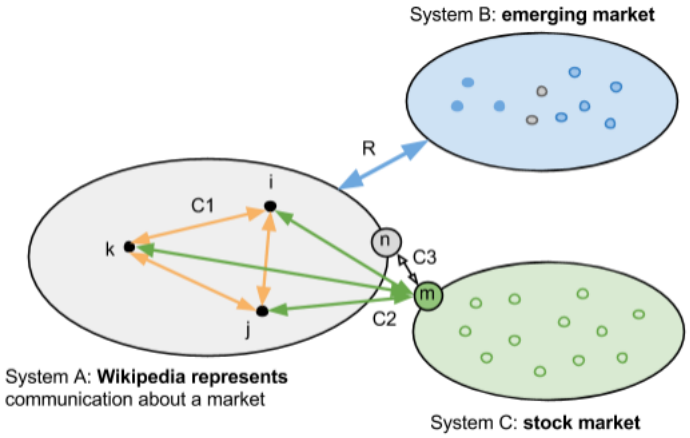

Figure 15.2.: **Markets and social media applications (SMA) as interconnected complex systems** - Information flow can be analyzed based on a generic stub, e.g., represented by Wikipedia content and server log data. Wikipedia allows an implicit selection of elements belonging to an unknown emerging market (blue) and it provides access to pages linked to categories and list pages, which describe well defined markets (green). How well a market is represented in Wikipedia is quantified by a text and structure based representation and relevance measure (R). Different types of correlation measures can than be applied, e.g., intra-correlations, which are expressed in a correlation matrix or by dependency networks (C1). Inter-correlations between Wikipedia access data and market data are calculated by partial correlations (C2). Meta correlations are applied on a group level in order to compare intra-correlations in corresponding systems (C3).

Preparation: Define the Scope

An appropriate selection of representative Wikipedia pages is crucial and influences the analysis results. Depending on the subject and objectives of the research project one has to select and to characterize the data sources carefully. On the other hand, it is relatively easy to inspect the data and to assess the quality in order to identify biases caused by the growth of the underlying system or dissimilar distributions of properties like text volume per page or number of links per page, which are typical for heterogeneous networks like the content networks we study in this work.

Step 1: Structural Analysis

The first step is a qualitative analysis, which is based on Wikipedia page content. It shows, how well a market - or even more general: a topic - is represented in Wikipedia. Therefore we apply the relevance analysis (introduced

in section 11.3). According to the intermediate results one has to refine the selection of nodes. Finally the bias, which might be introduced by connections to influencing nodes in the close neighborhood, can be evaluated and eliminated.

Step 2: Temporal Relevance Analysis

As a second step a more quantitative analysis is applied to pre-processed and cleaned time series data. For all selected central nodes - which define the scope of the study - one calculates the time-dependent relevance index (see chapter 11.4.2). This allows an identification of ranges in time during which obvious changes in public interest can be recognized. This approach goes beyond extreme event analysis as it does not need a definition nor an extraction procedure for events. The time-resolved relevance index is a relative measure which allows extraction of large changes regarding the local neighborhood, no matter if the system is in an equilibrium. Especially in emerging markets this robustness is an important aspect.

Step 3: Internal Correlations

Directed permanent connections between nodes are expressed by page links, which have been used for an implicit group definition. Temporal correlations (see C1 and orange arrows in figure 15.2) reveal additional information, especially in the presence of short-term activities, which do not have a direct influence on the underlying link structure.

Step 4: External Correlations

Finally external correlations between interlinked systems are analyzed by partial correlation analysis (see C2 and green arrows in figure 15.2) and by meta correlation analysis (see C3 and black arrow in figure 15.2).

In this work primarily cross-correlation as well as event-synchronization have been used as measures to define link strengths in a (re)constructed adjacency matrix that represents the underlying complex system. This allows us to apply several filter techniques and a characterization of dynamic properties of those networks over time. Our approach is in line with existing time series analysis methods, which also depend on the correct filtering of raw data. Our results show, that these aspects can not be handled in a single unified method - it depends on the properties of the measured data and the kind of effects that should be analyzed. As a consequence we introduce two categories of systems for which either only the internal analysis steps 1, 2, and 3 (System B) or all four steps (System C) of the proposed procedure can be applied as illustrated in figure 15.2.

15.2. Case Study I: Identifying Driving Forces in an Emerging Market

First, we study the *Hadoop ecosystem* in order to measure the public recognition of this emerging technology, using Wikipedia as a source for information. In this way, Wikipedia connects economical and sociological perspectives via social media data. In our case, the Wikipedia page content, page links, and click count data represent these two perspectives. User interest in a topic, which can be scientific, political, or even a commercial entity like a company or a product can easily be tracked by collecting web page usage statistics and also by sales statistics. The number of sold products or even the trading volume of stocks at stock exchanges are examples. In many cases such trading data is available only internally in companies, but in case of the stock markets a public trading data set is available.

To represent the Hadoop market we selected a set of 42 Wikipedia pages from six categories. Those groups represent multiple facets of the Hadoop market. The data set contains 5 of the global players in IT business (GP), 3 hardware vendors, two of which have relevant offerings around the data processing platform Hadoop and one without (HV), and also 6 important early adopters of the new technology (EA). The majority of selected nodes represents 17 fast growing young start-up companies (NC) which seem to drive this emerging market. The goal is to find out: how those entities are related to each other and if they can be grouped or clustered using social media based metrics. We add 5 more pages to the corpus in order to cover the probably influencing core technologies (TEC). This is important because the whole field around the Hadoop market is embedded into the existing database market. This explains why strong interactions should be expected. Because most innovations come form open source software projects we add 7 Wikipedia pages about software projects, which are managed by the Apache software foundation (SW). Some of those projects have been started already ten years ago and some others are really young but fast growing projects. Table 15.1 shows all page names used in this study.

Why are we interested in this specific topic? Apache Hadoop has been a so called ”game changer” in IT industry during the last 10 years. As the central software project, beside Apache Solr, Apache HBase, and many others it is also a synonym for Big Data platforms. The company Cloudera offers a Hadoop distribution and was the first company with commercial support around Hadoop. Other companies started very early with Hadoop related projects as early adopters. Global players on the other side are affected by this emerging market, it's opportunities and the new competitors. Some new but highly relevant companies like Talend or LucidWorks have been selected

Group	Page name (in English Wikipedia project)
GP	EMC_Corporation, IBM, Intel, Microsoft, Oracle_Corporation
HV	Dell, HP, Silicon_Graphics_International
EA	Amazon.com, Facebook, Google, National_Science_Foundation, The_New_York_Times, Yahoo!
NC	Datameer, BMC_Software, Cloudera, GoPivotal, Hortonworks, Karmasphere, LucidWorks, MapR, Penthao, Sematext, Splunk, Sqrrl, Syncsort, Talend, Teradata, TIBCO, WANdisco
TEC	Java, EMC_Isilon, Sun_Grid, Plattform_Computing, Condor_High-Throughput_Computing_System
SW	Apache_SOLR, Apache_Nutch, Apache_ZooKeeper, Apache_UIMA, Apache_Lucene, Apache_Mahout, Apache_Hadoop

Table 15.1.: **Selection of Wikipedia pages about topics with a direct relation to the emerging Hadoop market**. The list contains only pages from English Wikipedia which have been selected for a data driven study about the emerging Big Data or Hadoop market.

because their obvious commitment to the open source idea - they support open licenses, and because they are publicly recognized as driving forces in the developer communities for years. We expect to identify patterns, which indicate how interest and recognition of topics changes over time.

Since the details of our study have been published already, the focus of the next sections is on additional aspects with a potential relevance for market analysis procedures in general.

15.2.1. Trend Analysis and Dynamic Relevance

It was not a surprise, that for some of the selected concepts (or topics) no Wikipedia page was available prior to 2012. We had to identify pages about central technologies, which can be seen like seeds around which the marked emerges. Pages from global players, which define some kind of boundary conditions have been used instead as a stub for a time-resolved relevance analysis and for comparison within a broad context. The time-resolved relevance index (see Eq. 11.5 and 11.6) was calculated for nine pages as shown in figure 15.3 for a period of three years.

As a reference, companies like Oracle, the New York Times, and Capgemini are considered. They look like stable factors in the market and can be seen as a kind of reference while software projects have far less stability since they can be highly dynamic. This is why they are in our focus too.

We started with traditional trend analysis based on data from Google Trends and single page click-count data

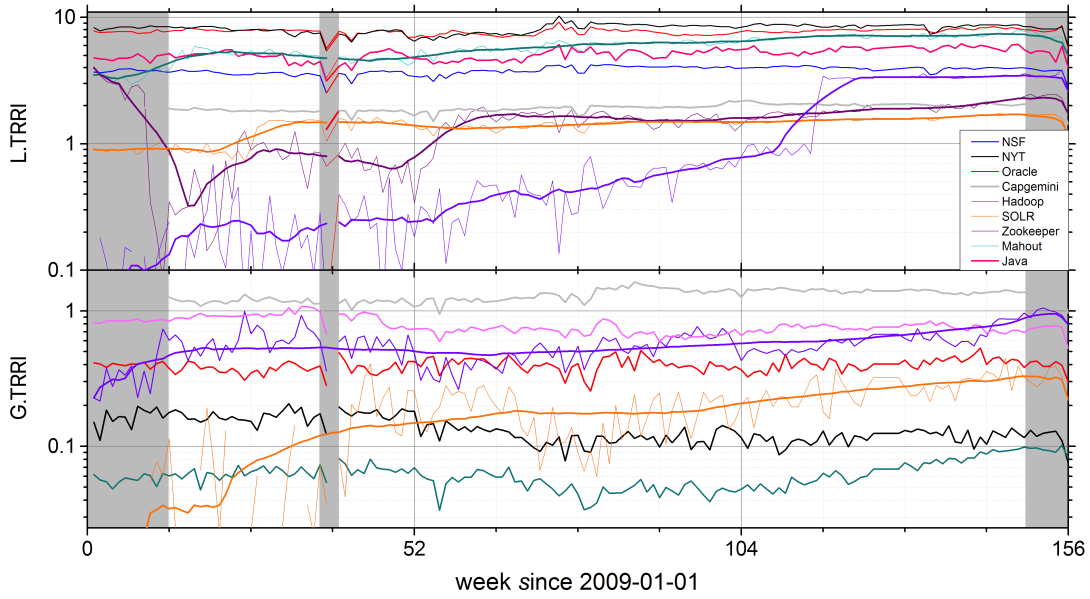

Figure 15.3.: **Time resolved relevance index for entities of an emerging market**. The L.TRRI is shown for a period of three years, starting in 2009 with a monthly resolution (thin lines) and quarterly sliding averages (bold lines). For global players we find stationary time series with comparable fluctuations. The "trending topics" which form the emerging market show higher fluctuations in the first year. As public interest in these topics increases, the fluctuations decrease. The relevance index increases significantly as soon as it is close to one. This leads to the conclusion that a relevance index close to one might be an indicator for a change in the entities life cycle.

from Wikipedia as shown in figure 3 in [4]. The figure illustrates major problems and weaknesses of both approaches if no detrending and normalization is applied.

The increasing interest in the topic Hadoop seems to be much stronger in keyword usage data than one can find in Wikipedia data. Trend analysis based on single time series is an error prone approach, because the neighborhood is not taken into account. Therefore the context sensitive relevance analysis seems to offer more reliable results, which also consider the embedding of the topics into their local neighborhood. In order to get reliable results a contextual normalization of the data was done. Figure 15.3 shows the local compared to the global time-resolved relevance index calculated for daily access-rates for the years 2009 to 2011. The highest representation index was measured for Apache Mahout followed by Apache Solr and Apache Hadoop. The first two are directly related to Apache Lucene, a Java based search library, which was started in 1999, and became an Apache top level project in February 2005.

Figure 15.3.b shows the G.TRRI values for some of the pages (NYT, Java, Mahout), which already have many non-English pages in Wikipedia. In general those topics have a very low relevance in other languages than English.

In the case of Apache Solr we can clearly see a significant change in the third quarter in 2009 where the average value increased by one order of magnitude and became stable at this level. For Hadoop and Zookeeper, an increasing trend can be recognized as well. More data has to be evaluated before a final statement is possible. Currently I assume, that such a clear grow pattern, up to a plateau can be found in many other topics. In our case the trends are changing at a scale larger than one year.

Not only the time scale, also the language of the selected central node is essential. An interesting example for language dependency is the page about the French company Capgemini which illustrates the impact of a wrong context selection. The local relevance index for the English page is rather low, but the global relevance index is high, caused by a high value for the French page. This allows us to conclude, that the context has to be adjusted according to the topic, otherwise an unknown bias still exists. The best context selection is given if the difference between local and global relevance index is maximized (see figure 11.6).

15.2.2. Discussion

From a marketing and business development perspective it is interesting to compare the new companies, which are improving the new and innovative technology, all together and most importantly with established global players. Companies with a strong commitment to Apache Hadoop, like Cloudera, MapR, Hortonworks, and Pivotal are the most important young companies in the market but regarding their representation in social media, especially in Wikipedia they are far away from early adopters which use the Hadoop ecosystem since the beginning.

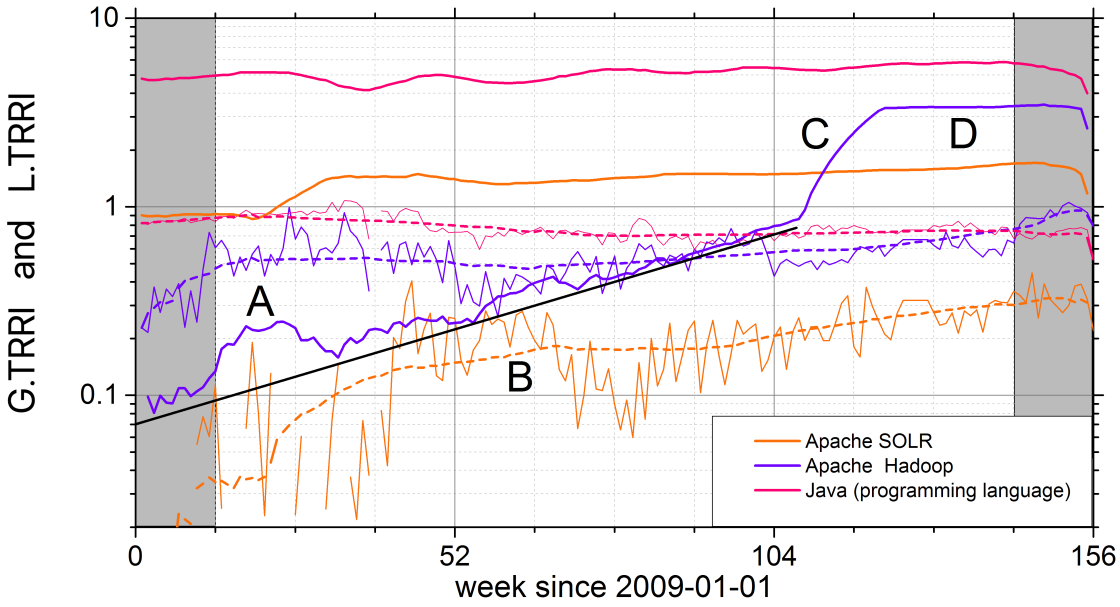

Figure 15.4.: **Project life cycle phases derived from Wikipedia usage data**. The black line shows a monotonous increasing trend which overlaps with a temporary increase (ranges A,B) which fades out after 3 months in the years 2009 and 2010. In 2011 this increase leads to a much stronger growth of the relevance index (range C) until a stable saturation level is reached (D). During the same time the dashed violet line shows a significant weaker trend for non English pages and a saturation can not be identified in the available data. Increasing interest during time ranges A and B can be explained with the ”conference season” each spring. Phase C illustrates the ”break through” of the topic in public recognition.

Because Google, Yahoo!, and Facebook use the public open source software Hadoop (or a comparable closed source technology in the case of Google) for their own services they should not be considered as direct competitors in the Hadoop market. One question remains: Who is the global player which might attract the most Big Data experts? This question is important not only for investors but also for people who want to enter this market, either by running own services or by looking for a valuable position.

Figure 15.4. compares L.TRRI (line) and G.TRRI (dashed line) for three exemplary topics. For the first topic *'Java (programming language)'* we find minor changes but not heavy fluctuations which indicates long-term stability. Such a long-term stability is reached in the third quarter 2009 for Solr (orange), and about two years later in the second quarter 2011 for Hadoop (purple).

In figure 15.5. we can identify sectors in the relevance plot which might be used for classification of topics. Especially groups A, C, and E seem to be related to typical phases in the life cycle of a company or product. For now we cannot generalize those results. More market studies like this one have to be conducted in other markets, like the automotive, mobile communications, pharmaceutical, or energy supply sector.

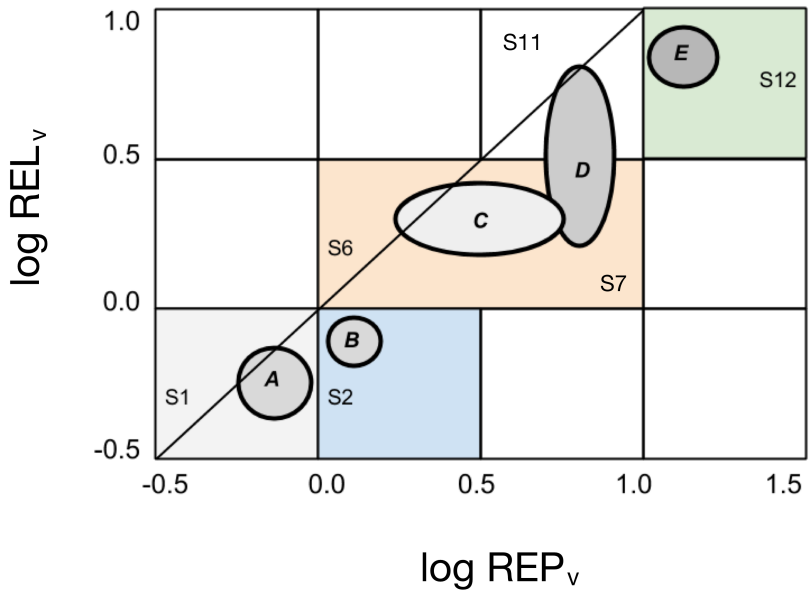

Figure 15.5.: **Clusters of Wikipedia pages correspond to common properties**. Wikipedia pages about companies, products, technologies, or open source projects form clusters, which represent properties they have in common. We use this approach to illustrate relationships between entities for which no obvious relation can be determined. We define the following classes: (S1) not well represented, group A, Start-Up companies and new topics; (S2) well represented but less relevant, group B, established projects with less attraction; (S6) attracting topics with average visibility, group C, companies with B2B focus; (S7 and S11) highly visible topics, group D, global players, and group E (S12), companies with strong end user focus.

This part of the study illustrates how Wikipedia allows tracking of public attention and public recognition of emerging topics based on content, contributed by a public crowd, which consists of self motivated editors in a self-organized process - which sometimes might be influenced by very active and focused editors in a business context - and by access-rate statistics which can be considered as a reliable data source. This approach allows to calculate the time delay between the increase of relevance of several contexts, e.g., the local and global context. There seems to be a critical value which can be an indicator for a transition in the projects life cycle, e.g., a break through in public acceptance.

We have drawn a clear picture of a very young fast growing market. The relevance plot and also time-dependent measures are in line with statements from domain experts and public recognition derived from several information channels which unfortunately cannot be used for a quantitative comparison.

Our approach can easily be generalized and extended to analyze content relevance and public recognition in arbitrary types of social communication and content networks. One of the most important technical requirements is availability of content together with editorial history beside access-rate data with at least daily resolution. If those different types of data sets would be available on all web servers than the proposed approach could also be extended to any type of web resources. Web servers would have to provide content, collect metadata and publish such metadata in a reliable way. Especially in the context of the growing linked data cloud it seems to be very useful and promising because it would allow to investigate many more processes in a well connected society.

15.3. Case Study II: The Movement of Interest in Financial Market Data During the Global Crisis

This case study is based on data from multiple financial markets, each represented by one stock index.

In order to connect to economic and historical facts we review some important aspects, which might have been hidden in the flow of events during 2008, when Lehman Brothers has been in trouble.

What happened in 2008? During the first half of the year, the value of Lehman Brothers' stocks decreased by 73% [284]. This was leading to an official announcement about the future strategy of the company. The number of employees should now be decreased by 6% which means 1500 people should loose their job. This kind of bad news can be see as a direct connection between business activity and social life or even society. When the Korean government announced plans to buy the tumbling company Lehman Brothers, the value increased by 5% on one day, and even by 16% during the period of a week, but as soon as problems had been reported, the value decreased by another 45% [285]. This numbers illustrate the influence of news, and even if the message is wrong or incomplete, an immediate negative effect can be recognized. The strong relation between companies in financial markets is highlighted by the impact, the losses had on the large stock indices. As a result of the missed deal with Korea not only the companies value decreased, also the S&P 500 index lost 3.4% on one day and the Dow Jones lost 300 points ($\approx 2.6\%$), some days later the Dow Jones lost 4.4% on one day, which was the highest loss of the stock index, which covers the 30 most important American companies, since the attacks on September 11, 2001 schematically.

How such a disruption of the financial sector can be spread via several network channels is illustrated in figures 3.2, 3.4, and figure 15.1.

Why are stock indices so important? In general, stock indices represent subsets of markets. Typically, they combine individual stocks of important companies in a region from multiple branches. Furthermore, additional differentiation between different index types is possible. Variants of the DAX, such as MDAX, SDAX, TechDAX, and ÖkoDAX all refer to the same region, Germany. MDAX and SDAX contain companies like the DAX but smaller. The TechDAX and ÖkoDAX are specialized to specific branches. TechDAX contains the 30 largest German companies from the technology sector. The ÖkoDAX includes the top ten companies in the renewable energy sector.

One stock index is not enough to analyze the economical properties of a country, but because an index includes multiple companies it is a convenient way of averaging. On the other hand, using Wikipedia categories has a comparable effect. In order to get more control on the composition we developed the local neighborhood networks. This allows a specific aggregation of multiple time series to increase the amount of available information. Finally we conclude, that comparison of such groups (stock indices and local neighborhood graphs) is more reliable than the analysis of individual stocks and single Wiki pages.

15.3.1. Preparation of Wikipedia Data

A set of Wikipedia pages has been selected instead of single pages about a specific topic. Such list-pages about a stock index contain several links to special pages about all companies included in the specific index. Following the inter-wiki links guides the crawler to list-pages about the same index in different languages using implicit semantic meaning.

Id	Page name	Id	Page name
1	Indice_de_Precios_y_Cotizaciones	12	Swiss_Market_Index
2	S&P_500	13	Athex_Composite_Share_Price_Index
3	S&P_Africa_40_Index	14	MICEX
4	Nikkeiv_225	15	Hang_Seng_Index
5	SSE_Composite_Index	16	All_Ordinaries
6	DAX	17	NZX_50_Index
7	FTSE_350_Index	18	KOSPI
8	BSE_Sensex	19	Taiwan_Capitalization_Weighted_Stock_Index
9	CAC40	20	Straits_Times_Index
10	Austrian_Traded_Index	21	IBOVESPA
11	AEX	22	TA-100_Index

Table 15.2.: **Selection of international stock indices with representation in Wikipedia**. The list shows 22 international stock markets, selected for a representation and relevance study.

Here we have a slightly different situation compared to case study one in the previous section. Because list-pages and their related nodes are covering the same topic, the embedding in an off-topic neighborhood is weaker. The links from normal pages usually have a stronger embedding into off-topic content, which forms the nearest neighborhood. Table 15.2 shows the page names of all list-pages for this part of the study.

15.3.2. Language Dependency of Market Representation in Wikipedia

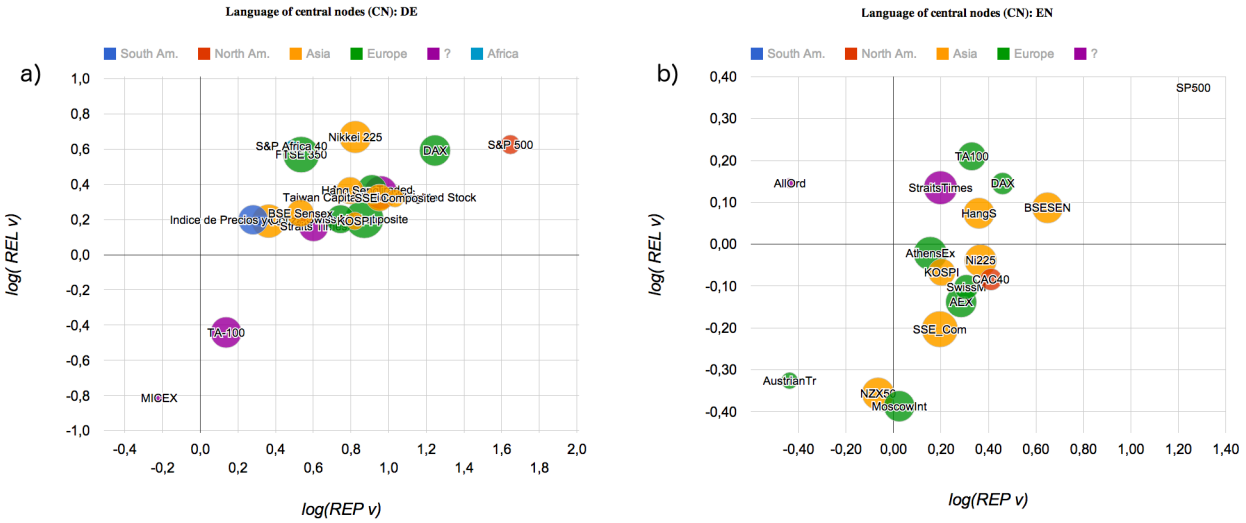

Figure 15.6.: **Language dependence of relevance-plots**. International markets should be analyzed based on data from multilingual sources, such as Wikipedia or Google Trends. But the representation index of a Wikipedia page strongly depends on the selected language. This kind of inconsistency has to be taken into account in more advanced analysis procedures.

Such a purely data-driven extraction method uses implicit properties included in the structure of local networks. This lowers the barriers especially in a multilingual global research context, which can now easily be analyzed using data from a multilingual environment like Wikipedia. In order to achieve a high level of accuracy the data has to be verified. As a first result the relevance-plot is shown in figure 15.6.

Before the multilingual data set can be used in an appropriate way to study dependencies and correlations between international financial markets, one has to understand, what language specific properties regarding topic representation exist. Figure 15.6 shows a relevance-plot for pages in German language in panel a) and in panel b) for the corresponding English pages for selected stock market indices shown in table 15.2. The average relevance index for pages in German language is higher than for English pages. This means that two different results can be found, depending on the selected language. In order to study and compare the interest of users by language such a context differentiation is useful and has to be repeated for all languages included in the study. Instead of context sensitive segregation a global aggregation would help to collect more individual time series per topic (in this case per company, included in an index). A context differentiation can be done based on Wikipedia categories or one can also use the second neighborhood, which usually is defined by off topic pages. Instead of a simple correlation analysis for each individual time series one can now use the time-resolved relevance index for each individual market.

Figure 15.7 shows a language specific difference in the long-term properties of the time-resolved relevance measure derived from Wikipedia access-rate data. While for the local index L.TRRI (pages in English language) only a short shock can be recognized during the time, when the international financial crisis was recognized (see also red area in figure 15.7) much stronger trends can be found in the global index G.TRRI. A local time-resolved index would have to be calculated for all languages to localize the reason for the trends.

The comparison of a dependency network and a Wikipedia page network illustrates how important a qualitative investigation of network properties is, and how it supports the interpretation of final results. In some cases a language specific segregation is not useful, especially in case of a global financial market, in which people and traders from all countries use all languages at the same time. Measuring the impact of a global shock in a particular language specific context is a good example for using the new relevance measure.

15.3.3. Markets as Networks Based on Different Data Sets

How are those local market networks inter-connected and is the link structure between Wikipedia pages reflecting the market connections, which have been found by Kenett *et al.* [286]? To analyze this, the two different network types are shown in figure 15.8 for comparison. An additional quantitative analysis has to be done, using network profiles.

Figure 15.8.b shows four groups of local page networks from Wikipedia, which are obviously intensively used as a source for information about financial markets. Multiple Wikipedia projects in several languages are inter-connected by so called inter-wiki links. The network shows also the first neighborhood around four central nodes

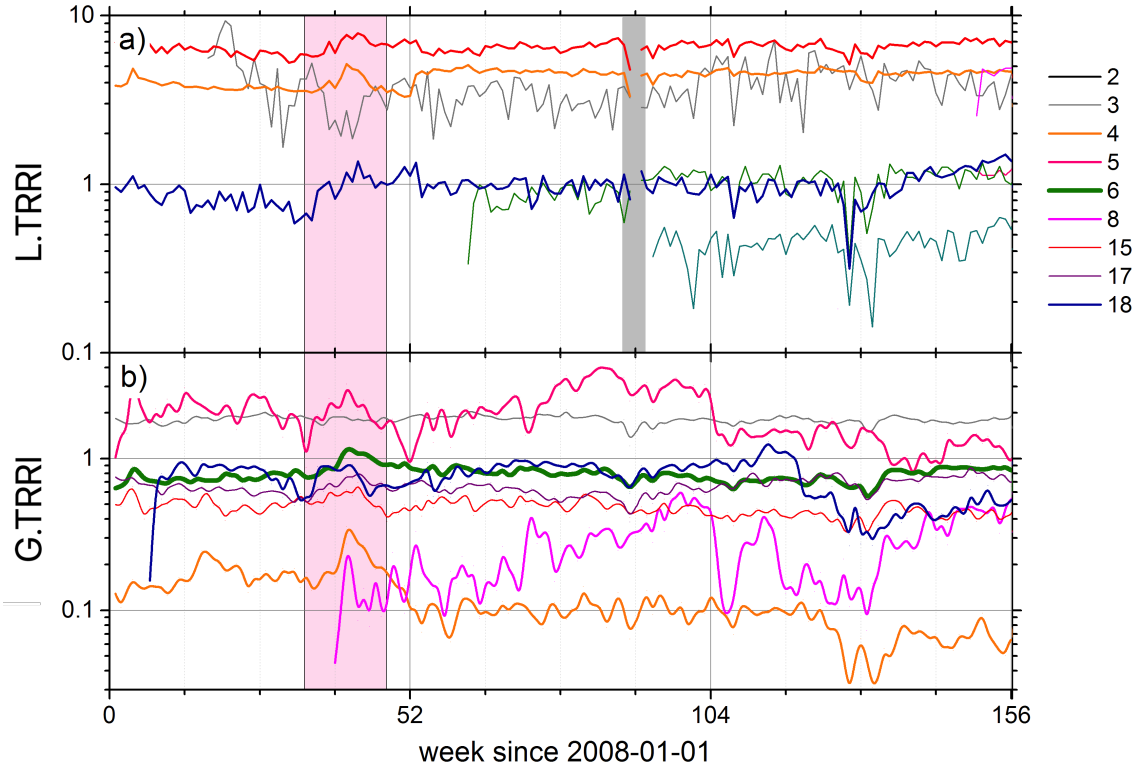

Figure 15.7.: **Time resolved relevance index for international financial markets**. The L.TRRI and G.TRRI are shown for a period of three years, starting in 2008 with smoothing (running averages) over one month. The red background indicates the time, when the international financial crisis was recognized. During this time, a clear impact on the relevance index for the German stock index DAX (dark green) and also for the Japanese index (orange) was found. One can also see a more stable local relevance index compared to a changing global relevance index, which means the impact is stronger for local languages. This could mean that people use the pages in their local language as a response to news from media channels, which is a reason for the changes in panel (b), while panel (a) corresponds to the business sector, in which the English pages are more relevant. Numbers in the legend are as in table 15.2.

for the stock market indices Nikkei 225 (Japan, red), DAX (Germany, blue), NASDAQ 100 (U.S. green), and BSE 200 (India, orange). The highest link density is found in the local network around the Japanese index Nikkei 225. The list-pages for the two Asian markets (red: Japan, orange: India) are linked directly via one intermediate page while the pages for the German index DAX (blue) and the American index NASDAQ 100 are not connected directly to each other.

Such a high-level inspection already shows structural aspects, which exist in both networks (clusters with different interconnection properties), but at the same time both networks show very different details and thus they might not be comparable to each other directly.

15.4. Case Study III: Correlations in Stock-Trading Time Series and Wikipedia Access-Rate

High market volatility and spontaneous shocks in stock markets might be related to increasingly heavy news coverage. Such activity can be considered as one reason for increasing interest in financial topics also in Wikipedia. Time series correlation analysis can be applied to study the coupling between markets and social media applications. Even though a causal dependency analysis is usually not possible we expect to find indicators for an existence of significant correlations or event synchronization during different market phases, such as phases of high market volatility, during shocks, or during stable market periods.

This third case study summarizes preliminary results. An obvious time dependence of cross-correlation link strengths in bi-partite functional networks has been found. The time delay between the two time series for which the cross-correlation is maximized also shows a significant difference if compared to randomly shuffled data. It is important to note that link strength distributions are not stable over time because the underlying process is non-stationary.

One can assume that during periods of massive price changes in stock markets, the correlations in measured

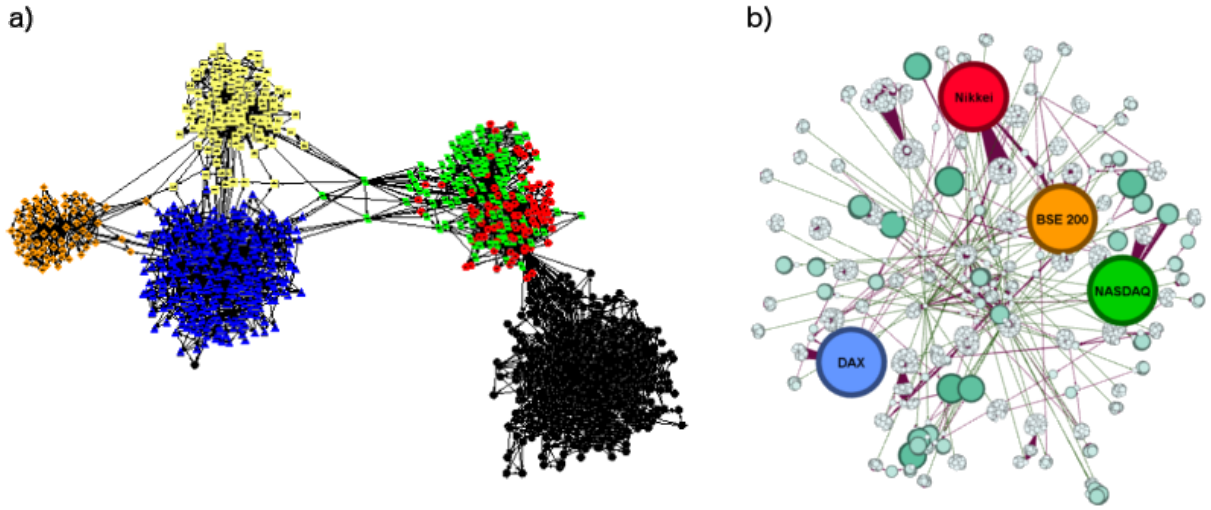

Figure 15.8.: **Page link network from Wikipedia pages about financial markets**. Panel a) was taken from [286] and shows a dependency network of dependency network of stocks belonging to six markets. The different markets are represented as follows: U.S. — black, Germany — red, U.K. — green, Japan — blue, India — yellow, and China — orange. Arrows indicate the direction of influence, from source to target. It is possible to observe that the U.S., China, India and Japan mainly influence themselves, while the U.K. and Germany have a mixed influence, as stocks belonging to these two markets form one large cluster. Furthermore, stocks belonging to the U.K. have an influence on all markets. Such interlinked subsystems can also be found in the Wikipedia page links network as shown in panel b). It was created from Wikipedia data, collected in January 2014. The large colored nodes represent selected list pages which have links to pages which for itself are also linked by pages from other markets.

Wikipedia access activity also increases because people use the content network more often to lookup background information. Since many different media channels, especially news channels might be considered to play the role of mediators in a global communication process (see figure 3.4), a time delay of at least one up to several days can be expected. A delay shorter than one day cannot be studied for technical reasons: the resolution of available financial data is limited. Strong daily patterns would have to be removed according to time zones, but click count data contain no details about the locations of the users.

We want to relate Wikipedia pages about companies included in stock indexes to the financial time series of those indexes. If the network structure of the internal dependency network of a market is known, the Wikipedia neighborhood networks can even be taken into account. This allows a differentiation between the elements involved in a coupling process and those that are loosely connected in the neighborhood but not directly involved in information flow. Even if the average correlation between the time series data sets is weak, one can show a systematic change in internal correlations compared to inter-correlations as shown in figure 15.11.(a,c,e). The core of the local context network attracts significantly more traffic compared to the surrounding neighborhood networks. This leads to a stronger correlation between core pages.

We adopted the concept of the "Index Cohesive Force" (ICF), introduced by Kenett *et al.* [286]. In section 9.2.3.c we introduced the "Context Cohesive Force" (CCF) based on a comparison of the cross-correlation within core and hull of local neighborhood networks. This allows a differentiation between a local and a global communication context defined within the local neighborhood network and seems to highlight properties of the underlying information flow. Preliminary results are shown in figure 15.11.(b,d,f) for Wikipedia pages around three financial markets.

What is the time shift between stock market activity and related Wikipedia access? Using a sliding window technique we could find the strongest peak at a delay of six days for short window lengths. For longer windows the peak becomes less pronounced but the location is stable. This shows clear differences between Wikipedia usage and Google keyword search statistics. According to Bordino *et. al.* [287] there is a delay of one day between the Google keyword time series and financial data. These results lead to the conclusion that ad-hoc search via web-search engines and retrieval of background information from Wikipedia are different processes on time scales in the range of one day to one week.

15.4.1. Data Analyzed and Methodology

For two pairs of data sets, each containing one subset from the social network Wikipedia and one from historical stock market data, we reconstruct two functional bi-partite networks. The two sub data sets for this case study

are the access-rate time series from Wikipedia and the trading data of the German stock index DAX, and the American index Standard and Poor's ($S\&P500$) recorded daily over a period from 1st of January 2009 to 30th of September 2009. The DAX data set contains time series for 30 companies and the S&P 500 data set contains data series from 500 companies. In this initial study we considered only Wikipedia pages with titles matching the company name. Such pages are not available in all languages because of the representation-bias, found in many topics. The embedding of pages into their neighborhood was thus only taken into account for the Wikipedia internal correlation analysis as reported in figure 15.11.

Several time series types are available form stock market trading activity via Yahoo! Financial Services [288]. Beside trading volume (TV) the logarithm of daily returns (LRP) has been used in this work. The absolute value of LRP (|LRP|) of traded stocks has been used as a stub for stock volatility.

Which is the most reliable metric to be used for a correlation or dependency analysis, and what time scales are relevant? According to Krings *et al.* [289] the selection of an appropriate window size is a critical factor in sliding window techniques. A good selection should contain time windows of a length, which fits to the assumed process and also such, for which no significant correlation is expected to show contrary properties. Seasonal trends like the weekly patterns in Wikipedia access-rate data can cause artifacts. Such misleading results can be identified by a comparison of results obtained from multiple time scales. Time series episodes have to be of the same resolution and length and they have to be aligned, which means the start time must be equal. This is a very important requirement and can be achieved either by removing values from both series or by filling in missing values to hide gaps. Such gaps, or missing data can be caused by technical problems or in case of trading data, by the nature of the underlying trading processes. During bank holidays and weekends no data is available. Removal of such data points influences the natural properties of time series. The frequency of an underlying oscillation will be increased and the pattern is also changed. This leads to artificial frequencies in the data series. Here we calculated the average cross-correlation links strength s_{p} and the standard deviation for the link strength distribution for non-overlapping episodes of length $l \in \{20, 40, 60, 80, 100\}$ days as a function time for a delay $\tau \pm 5$ days. The selected delay is related to a standard trading week of 5 days. For comparison with results presented by Bordino *et al.* in [287] we extended the range of delay values to ± 20 days and the length of episodes to $l \in \{50, 100, 150, 200, 250\}$.

15.4.2. Preliminary Results and Discussion

So far, only a weak indicator for a significant correlation between Wikipedia pages about financial topics and stock markets was found. Preiss *et al.* [290] reported a kind of an indirect connection between Wikipedia and stock markets in a recent study using Wikipedia user activity as an influencing factor to a trading strategy. The strategy is called 'Google Trends strategy' and seems to be much more successful than the traditional 'random investment strategy'.

One reason for the low quality of our current results was a technical limitation. The amount of data was not sufficient in the beginning. An even more important negative influence was identified as the *'hidden bias'*, also called representation-bias. This bias is caused by non-comparable representations of topics in Wikipedia such as companies, stocks and stock indexes. Representation of Wikipedia pages varies by topic and even more by language. A solution to this problem was developed in this work. It is illustrated in figure 15.6 in the previous section.

The first problem can easily be addressed in future projects, because Wikipedia access-rate data is available at an hourly resolution since December 2007. The data set is updated each month. The representation-bias can be analyzed as shown in chapter 11. Such a representation-bias is visible in a representation plot as a wide-spread cloud of bubbles of different size. If all bubbles are within a small area, and of a comparable size, no such bias exists or the bias is negligible.

15.4.2.a. Comparison of link strength distributions for different raw data types

In chapter 9.2 we investigated several computational approaches for the creation of functional networks. Here we compare the link-strength distributions for the Pearson correlation s_{p} (Eq. 9.8) and for the normalized correlation s_{norm} (Eq. 9.11) with delays τ up to ± 5 and later also ± 20 days for a bi-partite network from stock trading time series and Wikipedia access-rate time series to illustrate the hurdles towards a meaningful interpretation of such results.

Figure 15.9 shows the link strength distributions for episodes of different length (n=20, blue line; n=60, black dots) from raw data and from surrogate data (blue and black filled area respectively) for three different financial data series (TV, LRP, |LRP|) from German stock index DAX.

We found that the distribution of link strengths from surrogate data is very stable over time while the link strengths from raw data change as a function of time. This indicates that information regarding the dynamics of the coupling process might be extracted from this kind of data. Even if this approach gives a first indication about information possibly included in the data, it is no final result yet. It is known, that a plain cross-correlation analysis is not very stable. Especially the influence of peaks - as discussed in section 10.5.1 and 15.4.2.b - has

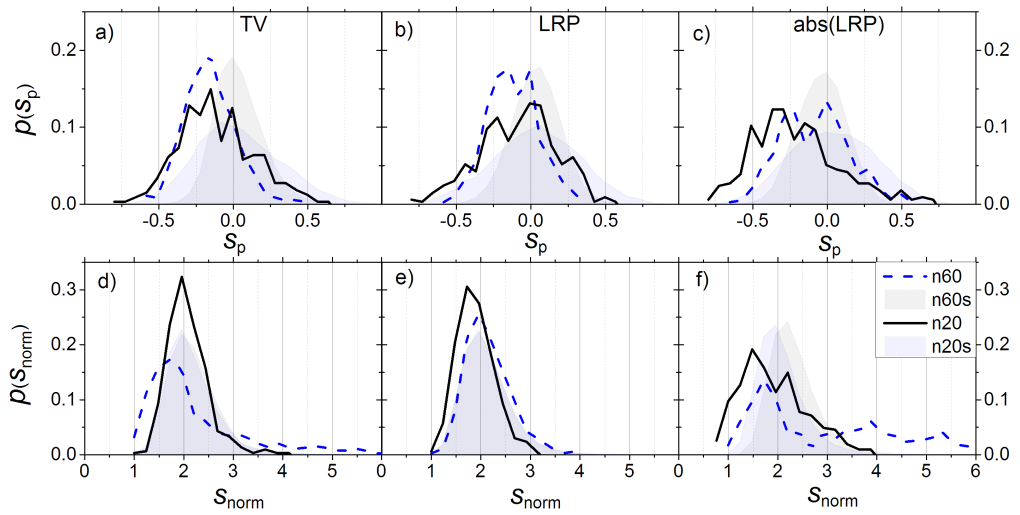

Figure 15.9.: **Graphical comparison of link strength distributions**. Link strength distributions for episodes of different length (n=20, blue line; n=60, black dots) calculated from raw data and from surrogate data (blue and black filled area respectively) for three different financial data types (TV, LRP, *abs*(LRP)) from German stock index DAX. Values s_{p} have been calculated with formula (Eq. 9.8). The values s_{norm} shown in the bottom row have been calculated via formula (Eq. 9.11)

to be taken into account. Another improvement was achieved by using the median of the cross-correlation value for a set of 30 episodes with hourly resolution to represent the correlation during a month - as shown in [134]. A meaningful adaptive filtering, e.g., based on a time-dependent threshold as shown in section 10.6.2 can also improve the quality of the analysis as it might produce more stable results, independent of unknown influences, which systematically change the average link strength.

Because the distributions of link strengths s_{p} can not be interpreted easily without additional statistical significance tests (such as Kolmogorov-Smirnov test) we use the results obtained for s_{norm}. As figure 15.9.d and 15.9.f suggest, we choose the data types trading volume (TV), and absolute log returns (|LRP|) for further analysis because of the clear differences between correlations calculated from real data and from surrogate data. For TV (see dashed blue line and blue shaded areas in 15.9.d) the difference is clearly visible for long episodes (dashed

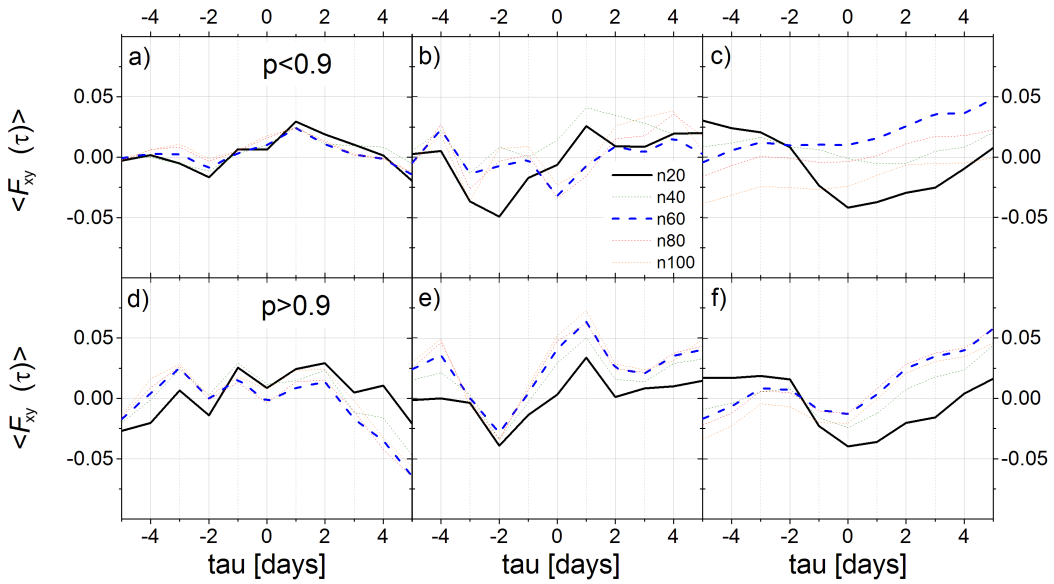

Figure 15.10.: **Influence of strong peaks on average correlation function for three financial time series types**. The p value of the Shapiro-Wilk test, applied to raw time series of three metrics for Wikipedia pages related to the German stock market index DAX is used as a static filter threshold. Average correlation-functions (see Eq. 9.11) are shown for five different window lengths for TV (a,d), LRP (b,e), and |LRP| (c,f) based on all time series with $p < 0.9$ in top row, and for time series with $p > 0.9$ in the bottom row.

line). The data type LRP (see panel (e)) shows such a difference for short episodes only, and in panel (f) we can see such a difference for short and long episodes. The differences are not the same for short (black line) and long (dashed blue line) episodes. These distributions already indicate a strong influence of the episode length. We investigate this aspect more in the following subsections.

15.4.2.b. Influence of Strong Peaks on Cross-Correlation

Next, we investigate the influence of strong peaks on correlation functions (see figure 15.10). One requirement for the application of cross-correlation analysis is that the distribution of all the values has to be consistent with a Gaussian distribution. The Shapiro-Wilk test together with a filter threshold $p_t = 0.9$ is used to check this supposition. For a high p value ($p_{\mathrm{SW}} > p_t$) the criteria is fulfilled and we can use the time series, in case of smaller p values ($p_{\mathrm{SW}} < p_t$) we put the considered time series into another group.

For trading volume time series with small p values the correlation is slightly higher for delays above $\tau = 2$. The overall shape is not changed by the filter in all three types of data. Figure 15.10 allows the conclusion that the influence of strong peaks in raw data series is visible and has to be taken into account. Figure 15.13 shows a comparison of average correlation functions for both groups in the DAX data set.

One can use the filter threshold to split the time series into two parts, a set with continuous time series and an event time series set for further analysis. All values higher than a variable threshold t are extracted and stored in an event time series. If the p value for the restricted time series is still smaller than p_t one decreases t and extracts more values which are stored in the event time series iteratively until the filter criteria matches. The two time series groups describe two different aspects of the underlying process and can be analyzed independently.

15.4.2.c. Correlated Information Flows in Financial Markets

In this case study we investigate two different types of connections, which are considered to be results of information flows. The first one is the internal correlation for pairs of access-rate time series within Wikipedia and the second one is the correlation between Wikipedia access activity and stock market data. We start with a comparison

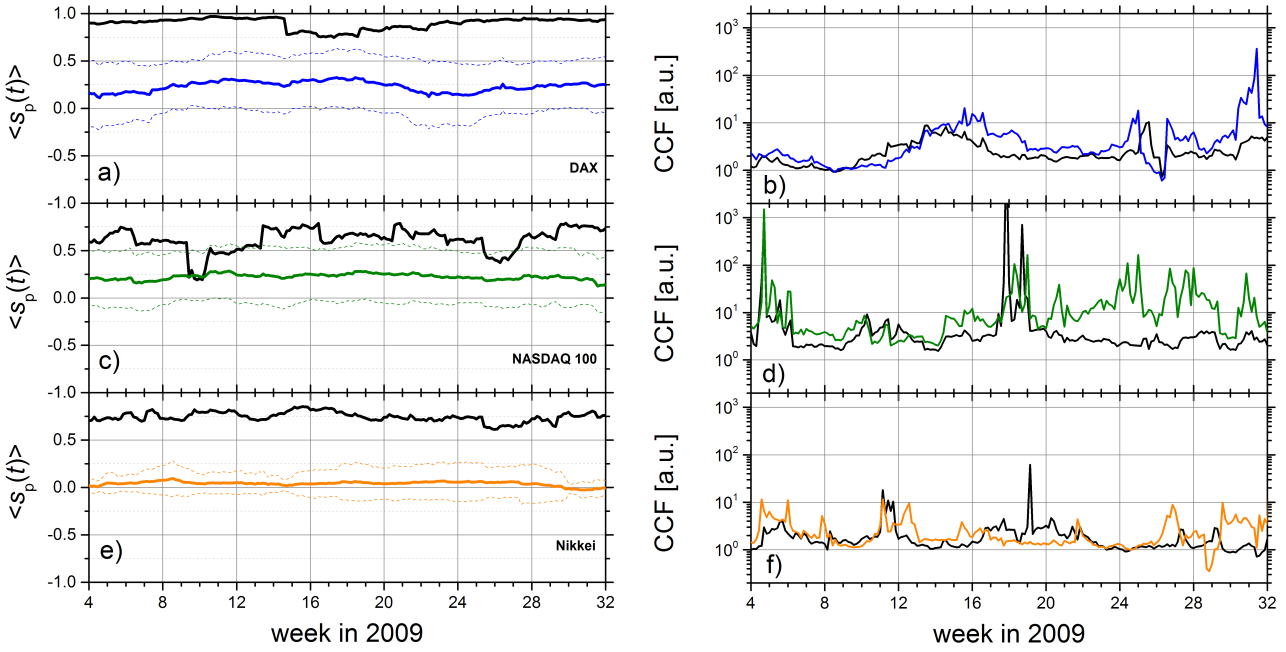

Figure 15.11.: **Average correlation strength s_{p} (Eq. 9.8) and context cohesive force (CCF) (Eq. 9.14) for Wikipedia page networks about three financial markets**. Panels (a,c,e) show the average correlation strength for access-rate time series from German Wikipedia pages DAX (top), NASDAQ100 (middle), and Nikkei (bottom). The black curve in each tile shows $\langle s_{\mathrm{p}} \rangle$ for the core and the colored lines show $\langle s_{\mathrm{p}} \rangle$ for the hull of the local neighborhood networks around the chosen pages. Dashed lines show $\pm\sigma$ for a Gaussian distribution of s_{p} values. Panels (b,d,f) compare the CCF for local communication context (black lines) and for global communication context (colored lines).

of average link strength as a function of time (for episodes of length $l = 28$ days). Figures 15.11 (a,c,e) show the time evolution of average correlation strengths within sub networks regarding Wikipedia pages around the stock indices DAX (DE, top), NASDAQ100 (US, middle), and Nikkei (JA, bottom).

The context cohesive force (CCF) is calculated (see equation 9.14) and presented in figure 15.11 (b,d,f) for the local communication context (black line) and the global communication context (colored line). This comparison allows an interpretation of correlation properties, which seem to be related to events in the real world. The different shapes of the black and colored curves in figures 15.11 (b,d,f) illustrate differences in the local and global properties. Local and global contexts are defined by the chosen languages. Especially in (b,d,f) local means all pages in German language and global refers to the group of all other languages in Wikipedia. This curve allows also a comparison of market or topic specific properties. The figure shows a first step towards a measure, which allows a comparison of markets such as financial markets (in Wikipedia represented as page groups, categories or individual articles), based on correlation properties. How those properties are related to structural properties and to real world events has to be analyzed in future projects using the context cohesive force as a generic metric, which expresses a collective property of an ensemble of nodes as a time-dependent value taking the individual properties of all elements into account.

15.4.2.d. Influence of Time-Delay τ on Cross-Correlation

From all cross-correlation functions $F_{xy}(\tau)$ for episodes of different length (20, 40, 60, 80, 100) we calculated the average correlation function $\langle F_{xy}(\tau)\rangle$ as shown in figure 15.12 for the DAX data set (a,b,c) and the S&P500 data set (d,e,f).

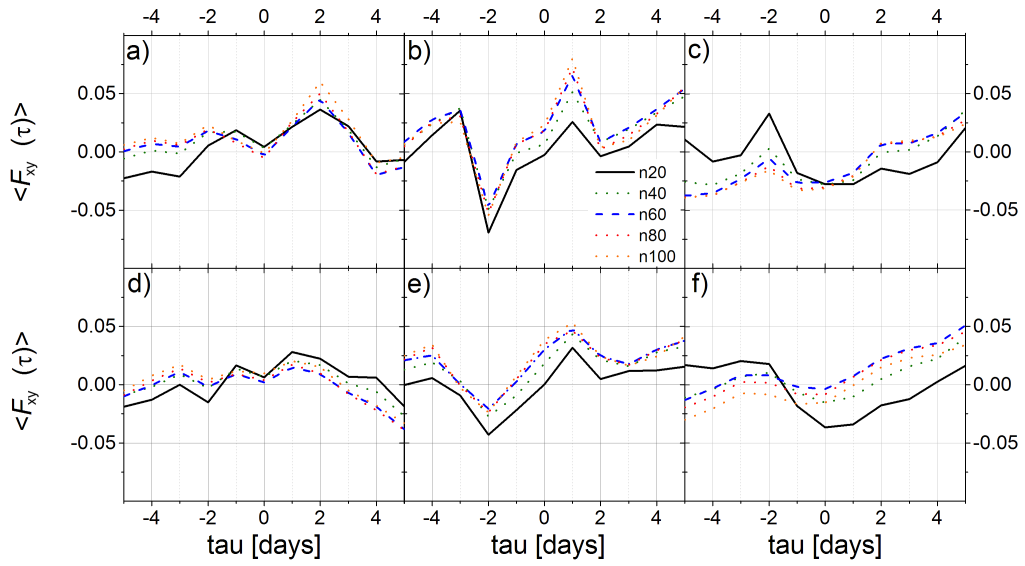

Figure 15.12.: **Comparison of correlation functions for two different markets**. The average correlation-functions from three financial metrics and access activity for Wikipedia pages related to the German stock market index DAX (top row) and the American market S&P500 (bottom row) are compared for different window lengths from 20 to 100 days.

Data was filtered based on value distribution properties of the input series. Only if the values of the raw time series were approximately Gaussian distributed and $t_{rmSW} > 0.9$ - the correlation function contributes to the final result, otherwise it is skipped. This filter approach allows one to identify artifacts, if they are caused by input data with properties non suitable for cross-correlation analysis.

We extended the range of analyzed time delays to ± 20 days to test the hypothesis that Wikipedia acts as a long-term memory. We assume this because it takes some time until Wikipedia pages reflect changes in reality. This might be because people look up background information not instantly but after a period of time. Than they come back and read more details. A second reason might be the media life cycle. It takes some time to transfer information and if people react on information which arrived late they contribute to such a time delay as well. What is the time range for such a delay?

Figure 15.13 shows this effect for five different window lengths with a clear maximum at $\tau = 6$ days for |LRP|, not for TV. The longer the time episodes are, the weaker the maximum is for metric |LRP| (orange and green line). For the index DAX (green line) with fewer companies this effect appears already for episodes of length 150 days. For the large index S&P500 only the intensity is decreased while the overall shape of the curve, especially the location of the maximum, is stable. This shows that for smaller groups this method can only be used with short episodes ($l <= 100$ days).

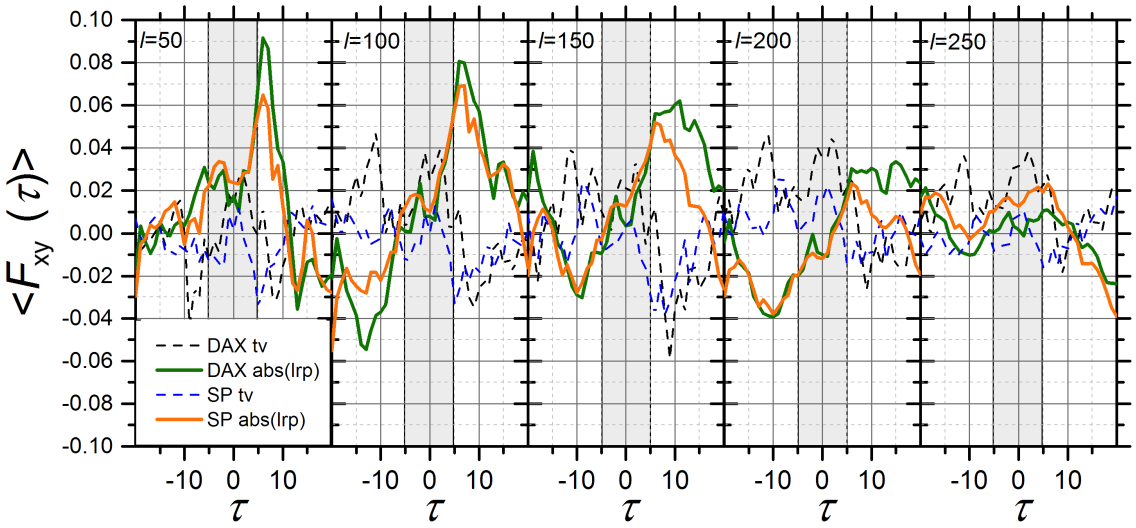

Figure 15.13.: **Average correlation functions (see Eq. 9.8) for sliding windows of variable length**. We show the average correlation function for the volatilities and Wikipedia access activity for delays τ in the range of ± 20 days for DAX (green line) and S&P500 (orange line). The dashed lines are for trading volume and Wikipedia access activity. We find the maximum correlation at a delay of six days. A local minimum at delay zero indicates two different components in the communication process, one with a positive delay and a weaker one with a negative delay. Market volatility seems to be a better measure than trading volume (black dashed line for DAX, blue dashed line for S&P500).

15.4.2.e. Discussion

This delay of six days (indicated by a maximum in the average cross-correlation function) is much longer than for Google keywords usage. Google is obviously used as an ad-hoc method for information retrieval, while Wikipedia is more of a long-term memory. Sometimes Wikipedia is considered to be a global brain. Our results show, if such a global brain exists, it might be build from multiple complementary parts. Such a complex global brain contains long-term memory, which is accessed via contextual associative access strategies such as web-pages, web-portals, Wikipedia, blogs, etc., and short-term memory. This short-term memory supports a kind of fast random access to content, like web-search engines and personal information management systems do.

Reasonable results can be found for the |LRP| data. We can conclude, that Wikipedia access-activity is about six days behind the stock market activity. Surprisingly, the local minimum at $\tau = 0$ and the small local maximum at $\tau = -2$ indicate also a certain amount of access activity with an opposite behavior. Those contributions precede the stock market activity. This does not mean, that Wikipedia influences the stock market directly, but we can see at least a connection between both systems on two different time scales with different strengths.

Furthermore, this case study shows that for the relation between information access on Wikipedia and a change in stock market trading some characteristic time delays exist. One has to be careful with this preliminary result. Especially using more Wikipedia categories with obvious relations to financial topics and also such with no direct relations should be considered to strengthen the conclusion based on more systematic comparisons. Such tests also known as AB testing. How such data sets could be selected for a future study is explained in the final subsection.

15.4.3. Further Improvements

Our initial questions: "Has Wikipedia an influences on financial markets?" and: "Is information flow via Wikipedia related to market dynamics?" are interesting but not specific enough to be research questions. As a result of this work I have to rephrase those initial questions. Furthermore, I suggest an extended data preparation procedure. We plan to collect and extract more data about historical events, such as strategic activities of large companies, political events like elections, the beginning and the end of wars, or cultural events like the presence of a song in music charts, a premiere of a movie, the Olympic games, or the world championship of several sports disciplines to define episodes. Such episodes must contain data from two phases, e.g., some time before and after the event. One can define such events automatically based on extreme event detection. An automatic detection of extreme events was successfully applied to the Wikipedia access-rate time series as well as to other social media applications like YouTube and Twitter. Such an approach is more general and independent from any particular topic.

Correlation analysis can be applied to such episode pairs and results can be compared across disciplines or domains. Using statistical tests one can now find if properties like link strength distributions, the time-resolved relevance index, or any other property is significantly different for both time frames.

Currently the significance test is based on randomly shuffled time series, because all correlations are destroyed by the shuffling procedure. Randomly chosen time series from other topics, which are assumed to be independent in the given research context, can also help to identify a systematic bias, which might exist. If a change in correlation-strength can be detected for the manually selected data set and also for some sets of randomly selected pages, one has found an counter example. This means a causal dependency can not be identified in this way for the data about the chosen topic.

Instead of an expensive calculation of long time-dependent properties during long stable phases an event driven approach is more efficient. It allows a higher number of episodes which show specific properties - an extreme event in this case - no matter what the reason or the meaning of the event is. Such an approach was already used to analyze the fluctuation properties of Wikipedia access-rate data (see section 13.2) for two different groups. The time series of one group were considered to have stationary access-rate time series. Time series containing extreme events have been grouped together in the second group.

As a conclusion of this work we recommend a normalization of the access-rate time series data. Instead of the average correlation function for all pages one should calculate the correlation between the time-resolved relevance index and the financial data series. The major benefit from this approach is a larger input data set consisting of all access-rate time series for all pages about the same topic in all available languages and the direct neighborhood - also from all available languages - as a reference. This allows a relative measurement instead of a biased absolute measure for which no calibration exists.

A more detailed quantitative analysis and a more systematic interpretation is required. The methods developed in this work will help to automatize the software tools for advanced studies on larger data sets. The focus of such a study will be on the role of Wikipedia as an important element in global communication processes. Besides the passive role as a global multilingual memory, Wikipedia has proven to be a useful stub to measure interest of people in a variety of topics. Several studies illustrate how extreme events in Wikipedia access activity, no matter if endogenous or exogenous, are caused by real world events.

In a future study the new method shall be applied to the same set of pages but instead of the daily access-rate, a normalized data set, the time-resolved relevance index shall be used. Because the new methods take language diversity into account and because they enable fine grained context selectivity, more robust results can be expected in future projects.

16. From Time Series to Co-Evolving Functional Networks: Dynamics of the Complex System 'Wikipedia'

This chapter is a reproduction of the conference paper submitted to the ECCS 2012[1] in Brussels[2]. Figures 13.1 to 13.6 would have been duplicated in this chapter. They were taken out of this section and references point back to chapter 13.

16.1. Introduction

Internet-based social networks (as novel information and communication platforms) often reflect the dynamics of changing interests and activities in society by characteristic usage patterns. Here we study the dynamics of user access-rate time series and edit-interval time series for all articles in the online encyclopedia "Wikipedia" with access-rates exceeding 255 per hour at least once. While other research on social networks mainly focuses on the development of their structure, we also study the usage of the elements (Wikipedia articles) for information spread. In particular, we characterize the fluctuation behaviour of the both, access-rate and edit-interval time series. For describing the reoccurrence of bursts exceeding certain thresholds we investigate the statistics of return intervals between these bursts. We find stretched exponential distributions of return intervals with identical parameters for all thresholds in access-rate time series, while edit time series show a simple exponential distribution of return intervals. To characterise the fluctuation behaviour of the access-rate time series we apply – after removing the daily and weekly periodicities – the detrended fluctuation analysis (DFA) method. We find that most access-rate time series are characterized by long-term correlations with fluctuation exponents $\alpha \approx 0.9$. To understand the complex processes underlying these different dynamics of access-rates and edit intervals, we characterize and compare three organizational and dynamical networks associated with 'Wikipedia' in the second part of the work: (i) the network of direct links between Wikipedia articles, (ii) the usage network as determined from cross-correlations between access-rate time series of many pairs of articles, and (iii) the edit network as determined from co-incident edit events. The major goal is to find correlations between components of these three networks that characterize the dynamics of information spread in the complex system. The network reconstruction is done by two different approaches. For access-rate time series, we use the cross-correlation coefficient at time delay zero between both time series of a selected group of nodes, linked to a central node, in combination with statistical significance tests. The link strengths for the corresponding edit time series are determined by the event synchronisation between all pairs of articles. We find that – even though the dynamics of article access-rate and edit-interval time series are characteristically different – there are indications of a co-evolution of the corresponding dynamic functional networks. Obvious differences between both reconstructed networks are also shown. The results help in understanding the complex process of collecting, processing, validating, and distributing information in self-organised social networks.

16.2. Dataset

We study Wikipedia access-rate and edit-interval time series recorded during a period of 10 months in 2009, focusing on all articles with access-rates exceeding 255 per hour at least once. Our dataset is a collection of log records of all edit activities (in an SQL database, 1 second time resolution) and the hourly counted number of accesses (downloads) of each page (in a binary database, time resolution occess rates is 1 hour). We begin with characterizing the properties of each single page (article) [1]. We observe daily and weakly activity patterns in the hourly access-rates for most Wikipedia pages in addition to apparently random fluctuations and bursting activity, see Figure 13.1. These weekly trends are removed from the raw data of access times. To characterize the properties of extreme events in the access-rate data we extract the width of pronounced bursts that exceed the average access-rate be more than a factor of two. The average durations of each burst before (t_{before}) and after (t_{after}) the maximum are compared to each other, see Figure 13.2(a). We find two regimes of different scaling behaviour for short events with length of less than 2 h and long events with a length of more than 10 hours. This property is independent from the selected threshold. The bursts are characterized by power-law or exponential

[1] `http://eccs2012.ulb.ac.be/program.php`
[2] The original document is available here: `http://www.physik.uni-halle.de/Fachgruppen/kantel/100-12-ECCS_Proc.pdf`

increases and decreases of activity, and they can be classified as 'endogenous' (with significant precursory activity) or 'exogenous' (extrinsically caused) events [1, 2, 3].

16.3. Characterization of Single-Article (Node) Properties

To characterize the fluctuation behaviour of the access time series we applied – after removing the daily and weekly periodicities – the detrended fluctuation analysis (DFA) method [4, 5]. We also compare results for different detrending orders to see if there are relevant effects of trends or nonstationarities (such as bursts). We find that most article access time series are characterized by longterm (power-law) correlations, see Figure 13.4(a) to (c). The histograms in Figs. 13.4(d) and (e) show that the power-law scaling behaviour, $F(s) \sim s^{\alpha}$, is quite universal with fluctuation scaling exponents rather narrowly distributed around $\alpha \approx 0.9$ for articles with quite stationary access-rates and rather nonstationary, bursty fluctuations. Note that α is related to the correlation exponent γ characterising the power-law decay of the auto-correlation function: $C(s) \sim s^{-\gamma}$ by $\gamma = 2 - 2\alpha$. There is only a weak dependence of α on the total number of accesses (i.e., the "importance" of the articles), see Fig. 13.4(f).

For describing the reoccurrence of bursts exceeding certain thresholds we investigate the statistics of the return intervals between these bursts [6, 7], see Figure 13.5.

We find stretched exponential distributions of return intervals with identical parameters for all thresholds in access-rate time series. Edit time series, on the other hand show a simple exponential distribution of return intervals, see Figure 13.6. The results are also compared regarding different languages in Wikipedia. The results for article access-rates are in full agreement with DFA results shown in Fig. 13.4. For more details we refer to [1]. After the detailed characterisation of the properties of single pages, we want to understand the causes of the different dynamics of the edit and access processes. Therefore we use a network representation of the interrelations of the articles to study these different processes.

16.4. Construction of Functional Networks

As human interaction and information spread via on-line networks is becoming increasingly important for our contemporary technological society we should not regard the Wikipedia system as a collection of independent web pages. We therefore reconstruct and compare three organizational and dynamical network structures associated with Wikipedia in the following second part of our work. The analysis of the static link network is just one aspect of the whole system. By looking at a dynamic link structures, we can obtain a second aspect or a second subsystem. The whole Wikipedia community uses the system, while it is edited and while it changes its underlying properties. Because of this, we want to isolate the different views of the interconnected processes of growing, changing and using the network. In particular, we study (i) the network of the direct links between Wikipedia articles of various languages, (ii) the usage network as determined from cross-correlations between click-count time series of many pairs of articles, and (iii) the edit network as determined from co-incident edit events, see Figure 16.1. The major goal is to find correlations between components of these three subsystems which can be seen as networks which characterize the dynamics of information spread in the complex system.

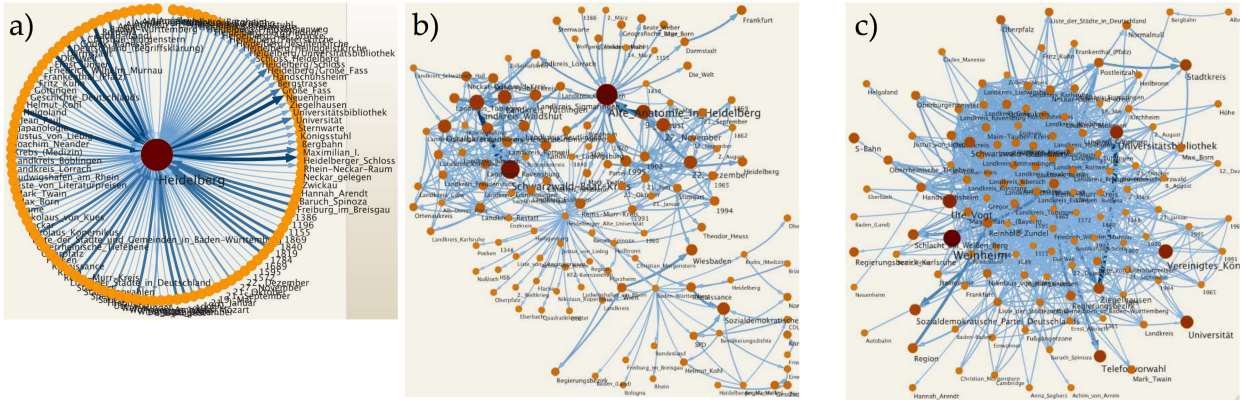

Figure 16.1.: Comparison of (a) the static link network, (b) the correlation network based on access activity for the whole recording period and (c) the correlation network of edit activity for a subnet of about 120 Wikipedia pages linked to the page with the name "Heidelberg". The figures were generated using the map.equation tool [8].

The process of reconstruction is done by two different approaches. For access-rate time series, we use the cross-correlation coefficient at time delay zero [9] between both time series of a selected group of nodes, linked to a central

node, in combination with statistical significance tests. The link strengths for the corresponding edit time series are determined by the event synchronisation between all pairs of articles [10, 11]. Obvious differences between both reconstructed networks are apparent in the exemplary functional networks, see Figure 16.1. In addition, we can observe dynamic changes in the reconstructed networks, which reflect changes of interest focus in society. These presentations (as shown in Figure 16.2) help in understanding the complex process of collecting, processing, validating, and distributing information in self-organised social networks.

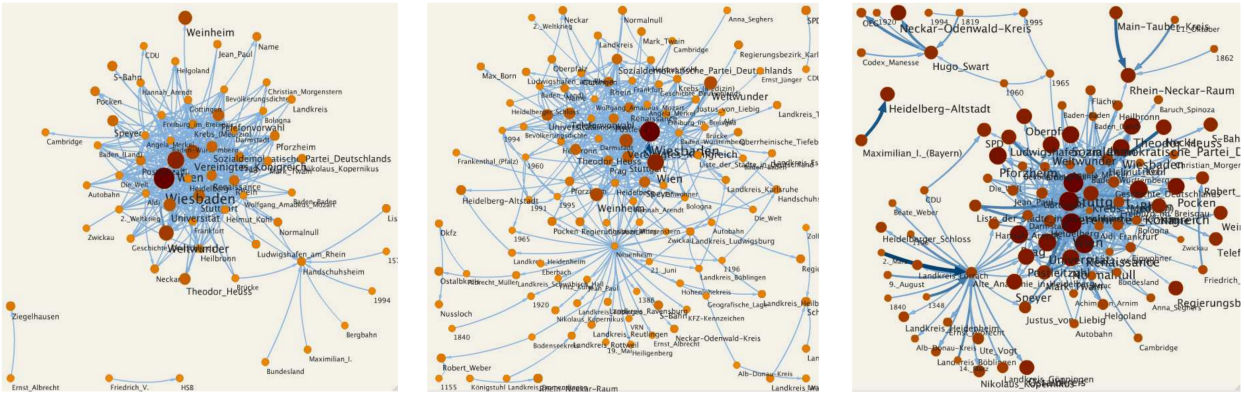

Figure 16.2.: For one selected central node (Wikipedia page for the city of Heidelberg) the time series for all linked nodes are extracted and access-rate cross-correlation link strengths are calculated for three different time frames. One can see clearly, that the correlation between single nodes in the context of a central node changes in time. The figures were generated using the map.equation tool [8].

16.5. Outlook

A deeper analysis of the complex processes underlying the Wikipedia system will be possible as soon as we have a generic method for generating such networks based on measurable data sets. The properties of the data recordings also have an influence on the used algorithms, e.g. the number of edit events is much smaller than the access events. Such differences lead to several variations of the definition of the link or correlation strength. Dependent on the properties of each single subsystem different algorithms have to be used and adapted. We have to select or define useful measures, for example the clustering coefficient, the degree distributions or the average path length of the calculated networks. Based on these properties we may see, what external influences could lead to a phase transition in the underlying system. A study of dependencies between properties of correlation networks, calculated from measured time series, will allow new approaches in the research field of socio-technical-complex-systems. A study of the relations or interactions between connected subsystems also leads to the emerging field of “network of networks” [12, 13].

References

[1] M. Kämpf, S. Tismer, J. W. Kantelhardt, and L. Muchnik; Burst event and return interval statistics in Wikipedia access and edit data, submitted to Physica A (2011).

[2] L. Mitchell, M. E. Cates; Hawkes Process as a model of social interactions : a view on video dynamics; J. Phys. A: Math. Theor. 43 (2010) 045101.

[3] R. Crane, D. Sornette; Robust dynamic classes revealed by measuring the response function of a social system; PNAS 105 (2008) 15649-15653.

[4] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger; Mosaic organization of DNA nucleotides; Phys. Rev. E 49 (1994) 1685.

[5] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley; Multifractal detrended fluctuation analysis of nonstationary time series; Physica A 316 (2002) 87.

[6] J. F. Eichner, J. W. Kantelhardt, A. Bunde, and S. Havlin; Statistics of return intervals in long-term correlated records; Phys. Rev. E 75 (2007) 011128.

[7] A. Bunde, J. F. Eichner, J. W. Kantelhardt, and S. Havlin; Long-Term Memory: A Natural Mechanism for the Custering of Extreme Events; Phys. Rev. Lett. 94 (2005) 048701.

[8] M. Rosevall, D. Axelsson, C. T. Bergstrom; The map equation, Eur. Phys J. Special Topics 178 (2009) 13-23.

[9] J. F. Donges, Y. Zuo, N. Marvan and J. Kurths; Complex networks in climate dynamics : Comparing linear and nonlinear network construction methods, Eur. Phys J. Special Topics 174 (2009) 157-179.

[10] R. Q. Quiroga, T. Kreuz, and P. Grassberger; Event synchronization: A simple and fast method to measure synchronicity and time delay patterns, Phys. Rev. E 66 (2002) 041904.

[11] N. Malik, B. Bookhagen, N. Marwan, and J. Kurths; Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks, Clim Dyn (in press 2011), DOI 10.1007/s00382-011-1156-4

[12] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin; Catastrophic cascade of failures in interdependent networks; Nature 464 (2010) 1025-1028.

[13] S. Havlin, D. Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J. W. Kantelhardt, J. Kertesz, S. Kirkpatrick, J. Kurths, Y. Portugali, and S. Solomon; Challenges of network science: Applications to infrastructures, climate, social systems and economics, Eur. Phys. J. ST (in print, 2012).

17. Evacuation in the Social Force Model is not Stationary

This chapter is a reproduction of an article which was written in the context of the SOCIONOCAL project and published by P. Gawroński, M. Kämpf, J. W. Kantelhardt, and K. Kułakowski in the Journal Acta Physica Polonica (see [13] in my publication list). This research project is partially supported by the European Union within the FP7 project SOCIONICAL, No. 231288.

17.1. Abstract

An evacuation process is simulated within the Social Force Model. Thousand pedestrians are leaving a room by one exit. We investigate the stationarity of the distribution of time lags between instants when two successive pedestrians cross the exit. The exponential tail of the distribution is shown to gradually vanish. Taking fluctuations apart, the time lags decrease in time till there are only about 50 pedestrians in the room, then they start to increase. This suggests that at the last stage the flow is laminar. In the first stage, clogging events slow the evacuation down. As they are more likely for larger crowds, the flow is not stationary. The data are investigated with detrended fluctuation analysis and return interval statistics, and no pattern transition is found between the stages of the process.

17.2. Introduction

A human crowd is a specific system, which is of interest for various specialists for different reasons. A physicist is willing to treat a crowd as a gas or a fluid of interacting particles [1], a psychologist can concentrate on the process of self-categorization in crowds [2], and a sociologist asks for emergence of norms in a crowd [3]. In an interdisciplinary approach, these perspectives overlap. In the Social Force Model (SFM), designed by Dirk Helbing and coworkers [4, 5] in the 90's, physical interactions are combined with action of the social norm of keeping distance to unknown persons [6]. Although the differential equations used there can be considered as computationally complex, the SFM seems to describe properly the collective effects in crowd, which appear in particular in emergency situations, as an evacuation. In simpler techniques, such as cellular automata or lattice gas models, a prescribed area is reserved for each pedestrian, and the influence of other pedestrians is reduced to short range interactions. Even if some specific effects such as clogging and arching are reproduced (as, for example, in [7, 8]), dynamics of a pedestrian in these techniques is fully determined by her/his local environment. On the contrary, in most of crowd disasters the crowd size was essential [9]. For reviews on the SFM and other techniques and a list of literature we refer to [9, 10, 11, 12, 13].

Here we are interested in forces exerted by masses of pedestrians, when physical interactions accumulate and the crowd size does matter. Namely, we intend to investigate how the number of pedestrians in a room influences the flow through an exit. Therefore we designed a numerical experiment as follows. A number of pedestrians is waiting, crowded, at the exit. At $t = 0$, the exit is opened and the crowd is pushing towards it. Let us denote the number of pedestrians who remain in the room at time t as $n(t)$. In this setup, an experiment with N pedestrians provides data on experiments for all $n < N$, because there is no stage when people gather at the exit. Provided that the flow at the exit depends on the number of pedestrians in the room, we should observe this dependence by just measuring the curve $n(t)$. Alternatively, we can measure the time gaps between successive crossings of exit. One can write

$$n(t) = N - \sum_{i=1}^{N} \Theta(t - t_i) \tag{17.1}$$

where t_i is the time instant when the number of pedestrians in the room changes from $i + 1$ to i. We are going to concentrate on the series of the time lags $\Delta_i = t_{i+1} - t_i$. Is it stationary? Are there long-term correlations and/or regimes with characteristically different behavior?

Up to our knowledge, this question was not posed directly in the literature, but the shape of the function $n(t)$ was obtained several times by different authors. In the next section we gather the results obtained by other authors which are directly close to our specific interest. For reasons explained above, we do not refer to simulations done with the cellular automata and lattice gas model. In the third section the SFM is briefly described. In Section IV we describe the technique of the data analysis. Finally, we show our numerical results (Section V) and discuss them (section VI).

17.3. What is Known

In [4], the model equation of motions of pedestrians were formulated. Among other results, an effect of pressure of crowd was demonstrated; out of two groups attempting to cross the door in opposite directions, the larger group was prevailing until the larger group became smaller. In this paper, a noise term is included to the equation of motion, hence the term 'Langevin equations'. In [5], the same SFM equations were used without the noise term. There, the desired velocity was associated with the level of panic. Also, the evacuation time dependence on the desired velocity was found to display a minimum. Also, when the desired velocity was increased, a change of the process from a laminar to a clogging mode of the crowd behavior was observed. As the authors remarked, the effect was less pronounced for wider exits. Note that, as noted in [14], the experimental data collected in planes do not show abrupt changes of the effectivity of evacuation when the exit width is changed from 0.6 m to 1.8 m. As shown in [12] with more experimental data, the relation between the bottleneck width and the flow of pedestrians does not show any threshold.

In [15], three curves are shown, obtained by simulations with using the SFM, on the number of pedestrian who left the room against time. The curves were obtained for 200 pedestrians and three values of the desired velocity: 0.8, 2.0 and 6.0 m/s. First curve (0.8) shows that at the last stage of evacuation the flow decreases. This effect exists also, but is weaker, for the second curve (2.0), but not for the third one (6.0). Instead, the latter curve was found to be particularly noisy. In Fig. 3 of [15], the distribution of clogging delays is shown for up to 160 people and the three above given values of the desired velocity. Each curve shows a clear maximum between 0.2 and 0.4 s. In this and subsequent paper [16], the cluster size distribution is also investigated, where a cluster means a group of people in physical contact between them. For the laminar and the turbulent flow, this distribution is found to be qualitatively different.

In [17], the formalism of optimization, developed by authors for other purposes, has been applied to the evacuation problem. Both clogging and arching have been observed in the simulation. The evacuation time dependence on the number of people was found to decrease in a non-linear way, but no minimum of this curve was found. In [18], the influence is investigated of the desired velocity on the evacuation time, the latter being a measure of panic. The mode of motion when the evacuation time increases with the desired velocity has been classified as turbulent. The effect of wider exit was investigated directly: the desired velocity where the evacuation time displays a minimum was shifted towards larger value with the exit width. In [19] and references cited therein, an experiment performed in a wardroom with a group of 70 soldiers is described. It was found among other results that the clogging is more likely if the number of persons is larger than 45.

For completeness we remark also our two recent papers [20, 21], where the SFM was used to investigate the chances that persons in the crowd can decide about themselves. The stationarity of the process of evacuation was not investigated there.

17.4. The Model and Simulation

The simulation is based on the model of crowd dynamics, described by Helbing et al. [5]. In this model, the equation of motion of a person i of mass m is written as

$$m\frac{d\mathbf{v}_i}{dt} = m\frac{\mathbf{v}(\mathbf{r}_i) - \mathbf{v}_i}{\tau} + \sum_{j(\neq i)} \mathbf{f}_{ij} + \sum_{W} \mathbf{f}_{iW} \tag{17.2}$$

where the first term on the right hand side is the tentative acceleration of a person i who intends to have the velocity $\mathbf{v}(\mathbf{r}_i)$, (its length commonly termed as the desired velocity) dependent on the coordinates $\mathbf{r}_i$. As a rule, the vector $\mathbf{v}$ points to the exit center (large distance from the person to the exit) or to the closest point of the exit (small distance). Further, τ is the characteristic time of this acceleration, $\mathbf{v}_i$ is the actual velocity of i-th person, $\mathbf{f}_{ij}$ is the force exerted on i-th person by j-th person, and $\mathbf{f}_{iW}$ is the force exerted on i-th person by a wall W. The force $\mathbf{f}_{ij}$ contains three components; 'social' interaction which describes the tendency of i and j to keep distance between each other, and two physical interactions between their bodies: radial force and slide friction. The social part of interaction is also adapted from [5]. It is given by

$$f_{ij}^{psych} = A_i \exp((2R - \left\|\mathbf{r}_i - \mathbf{r}_j\right\|)/B) \tag{17.3}$$

where A_i and B are constants, R is the mean 'radius' of the vertical projection of the human body, and $\mathbf{r}_i$ is the position of i-th agent. The parameters of the simulation are adapted from [5]: the amplitude of the social force $A_i = 2000N$, the constant B which is responsible for the spatial dependence of the social force is $0.08m$, the radii of agents $R = 0.3m$, their masses $m = 75kg$, the characteristic time of acceleration is $\tau = 0.5s$ and the absolute value of the desired velocity is $|\mathbf{v}| = 3m/s$. As remarked in [5], these values allow to reproduce experimental interpersonal distances and flows through bottlenecks. Also, the same values are assumed for all persons, to minimize the number of parameters. The instant values of the velocities $\mathbf{v}_i$ allow to update the positions $\mathbf{r}_i$ as

well. The equations of motion are solved with the Runge-Kutta method of 4-th order.

The simulation was performed as follows. $N = 10^3$ pedestrians were gathered at a the closed exit of width of 1 m, which was opened at $t = 0$. The time lags $\Delta_i = t_{i+1} - t_i$ were measured between crossing of the exit by subsequent pedestrians. The simulation was repeated 100 times.

17.5. Data Analysis

In our analysis procedure, we split the data of each run into ten non-overlapping parts corresponding to 100 persons leaving the room. Each part is analyzed independently, but averages over all 100 simulation runs are calculated to improve statistics.

Quantitatively, correlations between time lags Δ_i separated by s people are defined by the (auto-) correlation function,

$$C(s) \equiv \frac{1}{L-s} \sum_{i=0}^{L-s-1} (\Delta_i - \bar{\Delta})(\Delta_{i+s} - \bar{\Delta}), \tag{17.4}$$

where L is the length of the considered data part and $\bar{\Delta}$ is the average time lag in this part. If the time lags are uncorrelated, $C(s)$ is zero for s positive. If correlations exist up to a certain number of people $s_\times$, the correlation function will be positive up to $s_\times$ and vanish above $s_\times$. For the relevant case of long-range correlations, $C(s)$ decays as a power law,

$$C(s) \sim s^{-\gamma}, \quad 0 < \gamma < 1. \tag{17.5}$$

A direct calculation of $C(s)$ is hindered by the non-stationarities and trends in the data, since $\bar{\Delta}$ is not constant. We thus apply return interval statistics and detrended fluctuation analysis to study short-term and long-term correlations in the data, respectively.

Usually, return interval statistics (RIS) study the time intervals between 'extreme events' that exceed a given threshold [22, 23, 24, 25, 26, 27]. In a sequence of uncorrelated values ('white noise'), these return intervals are also uncorrelated and distributed according to a Poisson distribution,

$$P_q(r) = (1/R_q) \exp(-r/R_q), \tag{17.6}$$

where R_q is the mean return interval $\langle r \rangle$ for the given threshold q. For long-correlated data, on the other hand, a stretched exponential distribution

$$P_q(r) = \frac{a_\gamma}{R_q} \exp[-b_\gamma (r/R_q)^\gamma] \tag{17.7}$$

has been observed [22, 23, 24, 25], where the exponent γ is the correlation exponent from Eq. (17.5), and the parameters a_γ and b_γ are independent of q [24, 25]. If, on the other hand, the data is nearly deterministic (and not random), all return intervals will fluctuate weakly around the typical value R_q, giving rise to, e. g., a Gaussian distribution,

$$P_q(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(r - R_q)^2/(2\sigma)], \tag{17.8}$$

for $r > 0$ with the small standard deviation $\sigma \ll R_q$.

Here we consider each event of a person leaving the room as an extreme event, so that the return intervals r are identical with the time lags Δ_i, and R_q is identical with $\bar{\Delta}$. There is thus no threshold q for extreme events, but we get much more statistics. Our RIS focus on short-term correlations (between successive persons).

Detrended fluctuation analysis (DFA) [28] has become a widely-used technique for the detection of long-range correlations in noisy, nonstationary time series [29, 30, 31]. The DFA procedure consists of four steps. First we determine the 'profiles' $Y(j) \equiv \sum_{i=0}^{j} (\Delta_i - \bar{\Delta})$, $j = 1, \ldots, L$. Secondly, we divide $Y(j)$ into $L_s \equiv \mathrm{int}(L/s)$ non-overlapping segments of equal length s. Thirdly, we calculate the local trend for each segment by a least-square fit of the data. Linear, quadratic, cubic, or higher order polynomials can be used in the fitting procedure (conventionally called DFA1, DFA2, DFA3, ...) [32]. Then we determine the variance $F_s^2(\nu)$ of the differences between profile and fit in each segment ν. Fourthly, we average $F_s^2(\nu)$ over all segments and take the square root to obtain the fluctuation function $F(s)$. Since we are interested in how $F(s)$ depends on the time scale s, we have to repeat steps 2 to 4 for several s. Apparently, $F(s)$ increases with increasing s. If data Δ_i are long-range power-law correlated according to Eq. (17.5), $F(s)$ increases, for large values of s, as a power-law,

$$F(s) \sim s^\alpha, \quad \alpha = 1 - \gamma/2. \tag{17.9}$$

To determine the asymptotic scaling behavior of this fluctuation function we plot $F(s)$ as a function of s on double logarithmic scales and calculate the slope α by a linear fit in the regime $10 < s < 100$. This way, short-term correlations affecting less than 10 persons subsequently exiting the room are ignored in the analysis.

17.6. Results

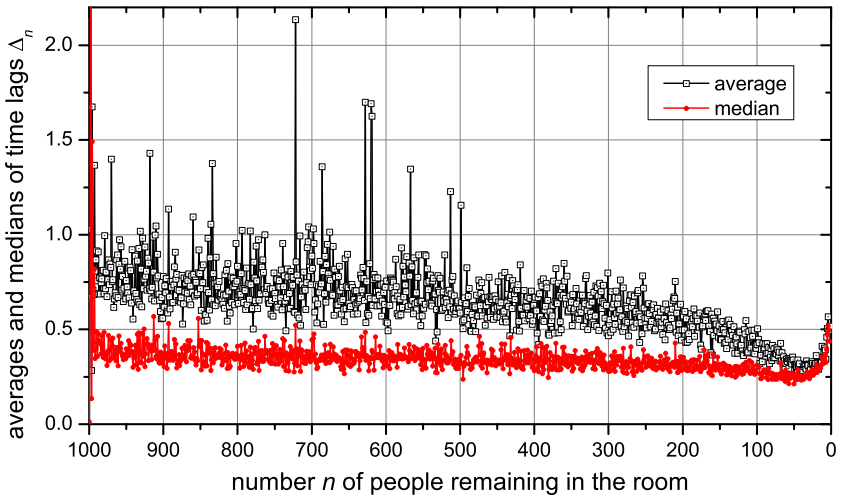

Figure 17.1.: Mean values and medians of the time lags Δ_n against the number n of pedestrians remaining in the room. Data from 100 simulations with $N = 1000$ people are included. The strong fluctuations at the beginning of the simulations (n close to 1000) are transient effects due to the opening of the exit. One can distinguish two regimes (approximately for $n > 50$ and $n < 50$) when comparing averages and medians.

In Fig. 17.1, results are shown for the average values and medians of Δ_n against the number n of pedestrians remaining in the room. As we see, two parts of the process can be distinguished. In the first regime from $n = 990$ to ≈ 50, averages and medians differ significantly and the fluctuations of Δ_n are rather strong. However, the fluctuations seem to decrease gradually, and so do the measured values. In the second regime from $n \approx 50$ to 0, where fluctuations are small, averages and medians are nearly identical, and the measured values increase slightly. This change of behavior can be interpreted as a cross-over from a stage with temporal cloggings to a laminar stage. In the latter case, the crowd behind pedestrians at the exit is large enough to push them out but not large enough to cause clogging. The appearance of the second stage of evacuation agrees with the character of two out of three discharge curves, presented in Fig. 2 of [15], for small and moderate desired velocity. We refer to Helbing et al. [33] (Section 2) regarding the applicability of the terms 'phase transition' or 'pattern transition' to finite non-equilibrium systems such as the one considered here.

In Fig. 17.2(a), we show the distributions of return intervals $r = \Delta_n$, i. e. $P(r)$ gathered in shorter parts of the data, where the departure from stationary flow can be approximately neglected. However, comparing the statistics, we see the differences. Initially, for large numbers n of people in the room, there is a large exponential tail of the distribution, formed by the clogging events and corresponding to the Poisson distribution Eq. (17.6) with a modified prefactor. In addition, there is a distinguished maximum near the time lag $r = \Delta_n \approx 0.2$ s, which is approximately described by a Gaussian distribution Eq. (17.8), also with a modified prefactor. The center of this peak agrees approximately with the results shown in Fig. 3 of [15]. We see that there are apparently two distinct components in the time lag distributions: typical short time lags around 0.2 s (probably due to persons successively exiting without delays) and exponentially distributed longer time lags (probably due to interruptions in the flow of exiting people because of clogging or arching effects).

For smaller numbers n of people in the room the exponential tail decreases, to nearly vanish during the last stage, i. e. for $n = 50$ to 0. Simultaneously, the nearly Gaussian peak for short time lags is hardly changing. While the behavior of the short time lags is well characterized by the nearly constant median of Δ_n (see Fig. 17.1), the changing averages of Δ_n, i. e. $\bar{\Delta}$, characterize the exponential behavior for long time lags. This is confirmed by the plot of scaled distributions shown in Fig. 17.2(b).

The application of DFA to the data from each of the ten parts yields fluctuation exponents α very close to 0.5, which proves the absence of relevant long-term correlations during all stages of the evacuation procedure. Specifically, we obtain $\alpha = 0.55$ and 0.54 for the first two parts (between 1000 and 800 people in the room) and values between 0.50 and 0.53 for all other parts. Since the systematic error of such fluctuation exponents is around 0.05 for time series of just 100 values [29], all of these numerical results are fully consistent with the null hypothesis of only short-term correlations in the data.

Finally, we are interested in the scaling behavior of the so-called discharge curve, i. e. the number of people remaining in the room against time. However, as we observe in Fig. 1, for the last approximately 50 persons the character of the curve changes. Then, for each numerical experiment $j = 1, ..., 100$ we determine the time t_{50}^j when $m = 50$ persons are left in the room in $j - th$ experiment, and we investigate the dependence of $n(t_{50}^j - t) - 50$ on t for $t < t_{50}^j$. Again, $n(t)$ is the number of pedestrians in the room. This dependence is averaged over all 100

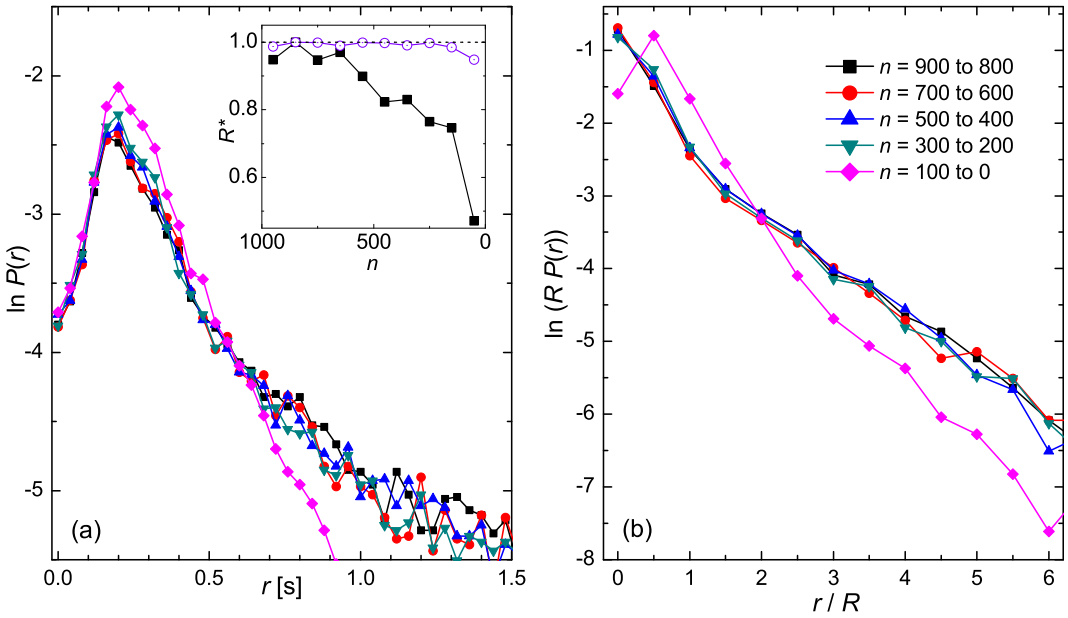

Figure 17.2.: (a) Distributions $P(r)$ and (b) scaled distributions $R \cdot P(r/R)$ of the time lags $r = \Delta_i$ with mean $R = \bar{\Delta}$ for $n = 900$ to 800 (black squares), $n = 700$ to 600 (red circles), $n = 500$ to 400 (blue triangles up), $n = 300$ to 200 (green triangles down), and $n = 100$ to 0 (violet diamonds) persons remaining in the evacuated room. The unscaled distributions in (a) show that the nearly Gaussian peak for short time lags is hardly changing, while the exponential tail is decaying with decreasing n. The inset in (a) shows the fitted slopes R^* of the exponentials decays (black squares together with the Pearson correlation coefficients of the fits (blue open circles). The scaled distributions in (b) show that $R = \bar{\Delta}$ characterizes the exponential peak fairly well.

experiments. The result is shown in Fig. 17.3 in the log-log scale. Fitting the result to a straight line we get an exponent β, which indicates if and how the evacuation speed depends on the crowd size in the first stage of evacuation. A result $\beta = 1$ means lack of this dependence; if $\beta = 1$, the evacuation is stationary at least in its first stage. However, our result is that clogging events make the evacuation slower, and these events are more likely if the number of pedestrians in the room is larger. The latter effect agrees with the experimental result in [19].

Effectively, our exponent β is smaller than one; we get $\beta = 0.832 \pm 0.001$ for the fitting range 50s$< t <$ 500s and parameter $m = 50$ as cutoff point of minimal number of persons in the room. Fig. 3 also shows that the numerical value of β is depending on m, reaching a bit smaller values for smaller m and larger values for larger m. We admit that perhaps the scaling regime is not fully reached yet because the crowd may be too small. Still, the conclusion that β is close to 0.85 seems to be well grounded. This form of the scaling relation allows to extrapolate the results for larger crowds.

17.7. Discussion

The results indicate that the probability distribution of the time lags Δ_i changes in time. In other words, the evacuation process simulated here is not stationary. One of the consequences is that the total evacuation time depends on the number n of pedestrians in the room, and therefore it is not a good measure of the efficiency of the process. If the size of the crowd is large, effects of clogging appear which are absent for small numbers of pedestrians. Although the case of larger desired velocity is not investigated here, we can reasonably expect that our findings will be particularly important if large desired velocity happens to be combined with large crowd. Effects of victims, who become obstacles, can only enhance the nonstationary character of the process.

A straightforward interpretation of our result is that the SFM successfully describes the effect of cumulation of the physical forces between agents at the exit. Keeping the hydrodynamic analogy, the pressure at the exit increases with the crowd size. If this pressure exceeds some critical value, pedestrians at the exit are not able to move, even if they are close to the exit. This is the origin of the observed large values of the time lags Δ_i, and the dependence of the size of the exponential tail of the lag distribution on the size of the crowd.

On the other hand, the data analysis with the return interval statistics and detrended fluctuation analysis shows that there is no transition between the first stage, when the mean time lag $\bar{\Delta}$ decreases, and the second stage, when the mean time lag increases. Also, correlations between pedestrians leaving the room in subsequent times are broken by the clogging events. As a result, the observed tail of the probability distribution of Δ is Poissonian.

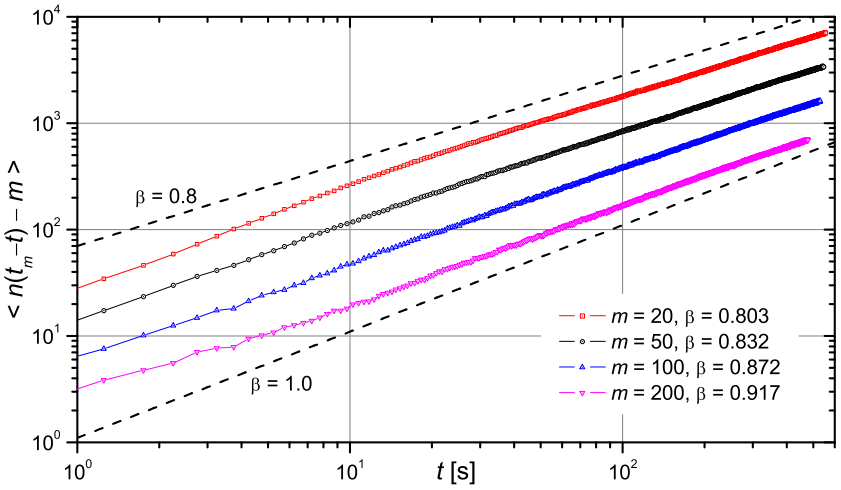

Figure 17.3.: The average discharge curve $\langle n(t_m - t)\rangle - m$ is shown versus the time t that measures the time interval till only m persons remain in the room. Curves are shown for $m = 20$ (red), 50 (black), 100 (blue), and 200 (violet). Each curve is fitted with t^β in the scaling regime $50 < t/s < 500$.

References

[1] D. Helbing, *A fluid dynamic model for the movement of pedestrians*, Complex Systems **6** (1992) 391.

[2] S. Reicher, *The psychology of crowd dynamics*, in M.A. Hogg and R.S. Tindale (Eds.), Blackwell Handbook of Social Psychology: Group Processes, Blackwell, Oxford 2001, pp. 182-208.

[3] R. H. Turner and L. M. Killian, *Collective Behavior*, Englewood Cliffs, N. J., Prentice-Hall, 1972.

[4] D. Helbing and P. Molnár, *Social force model for pedestrian dynamics*, Phys. Rev. E **51** (1995) 4282.

[5] D. Helbing, I. Farkas and T. Vicsek, *Simulating dynamical features of escape panic*, Nature **407** (2000) 487.

[6] T. S. Hall, *The Hidden Dimension*, Doubleday, Garden City, N.Y., 1966.

[7] A. Kirchner, K. Nishinari and A. Schadschneider, *Friction effects and clogging in a cellular automaton model for pedestrian dynamics*, Phys. Rev. E **67** (2003) 056122.

[8] R. Y. Guo and H. J. Huang, *A mobile lattice gas model for simulating pedestrian evacuation*, Physica A **387** (2008) 580.

[9] D. Helbing, I. J. Farkás, P. Molnár and T. Vicsek, *Simulation of pedestrian crowds in normal and evacuation situations* in *Pedestrian and Evacuation Dynamics*, edited by M. Schreckenberg and S. D. Sharma, Springer, Berlin 2002, pp. 21-58.

[10] A. Johansson and D. Helbing, *Crowd dynamics*, in *Econophysics and Sociophysics. Trends and Perspectives*, edited by B. K. Chakrabarti, A. Chakraborti and A. Chatterjee, Wiley-VCH, Weinheim 2006, pp. 449-472.

[11] F. Schweitzer, *Brownian Agents and Active Particles. Collective Dynamics in the Natural and Social Sciences*, Springer-Verlag, Berlin 2003.

[12] A. Schadschneider, W. Klingsch, H. Küpfel, T. Kretz, C. Rogsch and A. Seyfried, *Evacuation dynamics: empirical results, modeling and applications*, in Encyclopedia of Complexity and Systems Science, R.A. Meyers, ed., vol. 5, p. 3142-3176. Springer Science+Business Media, New York (2009).

[13] Z. Xiaoping, Z. Tingkuan and L. Mengting, *Modeling crowd evacuation of a building based on seven methodological approaches*, Building and Environment **44** (2009) 437.

[14] H. C. Muir, D. M. Bottomley and C. Marrison, *Effects of motivation and cabin configuration on emergency aircraft evacuation behavior and rates of egress*, The International Journal of Aviation Psychology **6** (1996) 57.

[15] D. R. Parisi and C. O. Dorso, *Microscopic dynamics of pedestrian evacuation*, Physica A **354** (2005) 606.

[16] D. R. Parisi and C. O. Dorso, *Morphological and dynamical aspects of the room evacuation process*, Physica A **385** (2007) 343.

[17] J. Izquierdo, I. Montalvo, R. Pérez and V. S. Fuertes, *Forecasting pedestrian evacuation time by using swarm intelligence*, Physica A **388** (2009) 1213.

[18] R. A. Kosiński and A. Grabowski, *Langevin equations for modeling evacuation processes*, Acta Phys. Pol. B Proc. Suppl. **3** (2010) 365.

[19] A. Seyfried, A. Portz and A. Schadschneider, *Phase coexistence in congested states of pedestrian dynamics*, LNCS **6350** (2010) 496.

[20] P. Gawroński, K. Saeed and K. Kułakowski, *Early warning of cardiac problems in a crowd*, LNAI **6071** (2010) 220.

[21] P. Gawroński and K. Kułakowski, *Crowd dynamics - being stuck*, submitted (arXiv:1009.1017).

[22] A. Bunde and J. F. Eichner and S. Havlin and J. W. Kantelhardt, *The effect of long-term correlations on the statistics of rare events*, Physica A **330** (2003) 1.

[23] A. Bunde and J. F. Eichner and J. W. Kantelhardt and S. Havlin, *Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records*, Phys. Rev. Lett. **94** (2005) 048701.

[24] E. Altmann and H. Kantz, *Recurrence time analysis, long-term correlations, and extreme events*, Phys. Rev. E **71** (2005) 056106.

[25] J. F. Eichner and J. W. Kantelhardt and A. Bunde and S. Havlin, *Statistics of return intervals in long-term correlated records*, Phys. Rev. E **75** (2007) 011128.

[26] M. S. Santhanam and H. Kantz, *Return interval distribution of extreme events and long-term memory*, Phys. Rev. E **78** (2007) 051113.

[27] N. R. Moloney and J. Davidsen, *Extreme value statistics and return intervals in long-range correlated uniform deviates*, Phys. Rev. E **79** (2009) 041131.

[28] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, *Mosaic organization of DNA nucleotides*, Phys. Rev. E **49** (1994) 1685.

[29] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde, *Detecting long-range correlations with detrended fluctuation analysis*, Physica A **295** (2001) 441.

[30] K. Hu, P. Ch. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley, *Effect of trends on detrended fluctuation analysis*, Phys. Rev. E **64** (2001) 011114.

[31] Z. Chen, P.Ch. Ivanov, K. Hu, and H. E. Stanley, *Effect of nonstationarities on detrended fluctuation analysis*, Phys. Rev. E **65** (2002) 041107.

[32] A. Bunde, S. Havlin, J. W. Kantelhardt, T. Penzel, J.-H. Peter, and K. Voigt, *Correlated and uncorrelated regions in heart-rate fluctuations during sleep*, Phys. Rev. Lett. **85** (2000) 3736.

[33] D. Helbing, M. Treiber, A. Kesting, and M. Schönhof, *Theoretical vs. empirical classification and prediction of congested traffic states*, Eur. Phys. J. B **69** (2009) 583.

18. Phases of Scaling and Cross-Correlation Behavior in Traffic

This chapter is a reproduction of an article which was written by Jan W. Kantelhardt, Matthew Fullerton, Mirko Kämpf, Cristina Beltran-Ruiz, and Fritz Busch and published in the journal Physica A (see [9] in my publication list). The work was supported by the European Union project SOCIONICAL (FP7 ICT, grant no. 231288).

Abstract

While many microscopic models of traffic flow describe transitions between different traffic phases, such transitions are difficult to quantify in measured traffic data. Here we study long-term traffic recordings consisting of ≈ 2900 days of flow, density, and velocity time series with minute resolution from a Spanish motorway. We calculate fluctuations, cross-correlations, and long-term persistence properties of these quantities in the flow-density diagram. This leads to a data-driven definition of (local) traffic states based on the dynamical properties of the data, which differ from those given in standard guidelines. We find that detrending techniques must be used for persistence analysis because of non-stationary daily and weekly traffic flow patterns. We compare our results for the measured data with analysis results for a microscopic traffic model, finding good agreement in most quantities. However, the simulations cannot easily reproduce the congested traffic states observed in the data. We show how fluctuations and cross-correlations in traffic data may be used for prediction, i.e., as indications of increasing or decreasing velocities.

18.1. Introduction

Transportation systems must usually be regarded as complex systems, since many agents are interacting with the infrastructure environment and with each other in many, often non-linear ways [1, 2, 3, 4]. In complex systems, non-stationary, intermittent, and non-linear oscillations and fluctuations occur that can often be described by self-organization and scaling relations (e.g., $1/f$ noise related with critical behavior) [5, 6]. Because of the non-stationary nature of traffic data, sophisticated linear and non-linear time series analysis techniques are needed for identification and quantification of scaling relations [7]. While many simulation model approaches for traffic flow have been developed by traffic engineers and also physicists, comparisons of model output with recorded long-term data are rather sparse, see, e.g., [2] for a recent comprehensive review. For the design and control of traffic infrastructure, traffic engineers have developed particular rules including definitions of traffic states and capacities of roads [8, 9, 10]. Physicists have also tried to characterize the stability of traffic states and to identify phase transitions between them by a dynamical analysis of their models [2, 4].

As most drivers are aware, the state of movement of traffic changes over time. In general it is expected that during periods of high demand (the so-called morning and evening peaks) the flow of traffic may experience a breakdown. Over almost 80 years, many researchers have attempted to develop and optimize models such that traffic states can be identified from the behavior of the models [2, 3, 4, 11]. However, such models, while important for a diverse range of needs (theoretical understanding, shock-wave theory, capacity estimation, and simulation calibration to name a few), do not yield a list of robust stylized facts that can be used to classify time series data into traffic states. Here, we proceed in the opposite direction and derive a data-driven (statistical) classification of traffic states based on stylized facts regarding the dynamical properties of the data, see e.g. [12] for a somewhat similar approach. Our results can be applied for a characterization of the transitions between traffic states and for traffic prediction. We note that the terms 'traffic state' or 'traffic phase' cannot be associated with standard thermodynamic phases [11], since traffic is an intrinsically non-equilibrium process [2]. According to Kerner [3], a *traffic pattern* is a distribution of traffic flow variables in space and time and can consist of different *traffic phases*. Traffic phases (e.g., 'free flow', 'synchronized flow' and 'wide moving jams') may again consist of different local 'traffic states', e.g. in stop-and-go traffic. Here, we completely avoid the term 'traffic phase' because of the non-equilibrium processes, and we use the term 'traffic state' to refer to traffic flows with similar local fluctuation and persistence behavior.

Simple approaches to state classification use a mapping (i.e. a set of thresholds) according to at least one traffic variable. In particular, tables of road density are used for the level-of-service definitions in Germany [9] and the USA [10]. Traffic speed and road density can be used together; examples include the Radio-Data-System Traffic-Message-Channel (RDS-TMC) traffic alert service [13] and the so-called MARZ process (German acronym: 'Merkblatt für

die Ausstattung von Verkehrs-Rechner-Zentralen') [8], which is the basis of motorway overhead sign alerts in much of Germany. State classification depends on good source data and appropriate data analysis. The work of [14] is particularly interesting in that in addition to macroscopic (i. e. averaged) values, microscopic properties were used to define some states. To delineate transitions, the magnitude of moving averages of velocity and flow were used. Cross-correlation between variables can also be used to define states [12]. Regarding state classification within the flow-versus-density diagram, clustering (K-means algorithm), was suggested [15]. In short, most approaches to state classification are based on rules that may or may not take past history into account. In this paper, we use empirical time-series relations to directly identify traffic states from dynamical properties irrespective of the current or average values of traffic parameters. We apply established statistical time-series analysis techniques, namely cross-correlation analysis and scaling analysis (using detrended fluctuation analysis, DFA [16, 17, 18]), to traffic flow in a part-time congested motorway corridor. This allows firstly a reliable examination of whether scaling relations, e.g., long-term persistent correlations, are present in traffic data. In addition, clustering of the method results allows us to observe an empirical set of traffic states based on stylized facts with potential applications for short-term prognosis. Finally, it allows us to test traffic simulation techniques commonly used by traffic engineers with respect to whether they can replicate the real-world stylized facts. The analysis results could also facilitate the development of an automated algorithm that can distinguish different states of traffic by analyzing data measured along a road or within a car.

The paper is structured as follows. First, the real-world data and the simulation approach are described (Sections 2 and 3). Then we present the results of variance analysis and cross-correlation analysis (Section 4), leading to an empirical traffic state classification scheme (Section 5). This is followed by the results of scaling (persistence) analysis (Section 6) and first results regarding a prognosis of traffic changes and phase transitions (Section 7). Section 8 discusses the results, in particular the properties of the states found and the types of transitions between them and concludes the paper by summarizing our separation between states and how it differs from other current approaches.

18.2. Long-Term Traffic Data and Flow-Density Diagram

Inductive detectors at 23 detection points along the Spanish M30 motorway (Madrid inner orbital motorway) were used for the measurement of traffic flow time series (number of vehicles per minute) and average velocity time series (in km/h). The data were digitally processed, aggregated over all lanes of the motorway at each detection point, and transmitted to the traffic control center at a time resolution of one minute. For this study, data is collected over four periods of four weeks and two periods of one week, covering three years (April 2008 till May 2011). This yields $(4 \times 4 + 2) \times 7 = 126$ full days of data recordings from each of the 23 detectors, i.e., 2,898 days (173,880 minutes) of data altogether.

We study time series of traffic flow Q_i (veh/min, multiplied by 60 to change the unit to the veh/h), velocity v_i (in km/h), and density ρ_i (in veh/km) at the time resolution of one minute. The density is obtained via dividing flow by velocity, $\rho_i = Q_i/v_i$. The problem of determining the empirical density via this formula is that it mixes a temporal average (the flow) with a spatial one (the density) [4] and gives too little weight to low velocities. Thus, a linear relationship between density ρ_i and occupancy p_i, defined as the portion of time when a vehicles is above the detector, holds only for low densities [12]. Since only 11 percent of our data include occupancy values, we estimate density by $\rho_i = Q_i/v_i$ in the following, but also use presentations that avoid densities.

Figures 1(a,b) show a flow-versus-density plot and a velocity-versus-flow plot of all measured traffic data. These plots are sometimes denoted as 'fundamental diagrams' [2]. The color of each dot is chosen according to the MARZ process [8], the current practical German definition of traffic states, see below. To include data from all detectors, flow and density values have been normalized to two-lane roads after determination of the MARZ states. I.e., flows and densities on a five-lane (four-lane, three-lane) road are normalized by a factor of 2/5 (2/4, 2/3). This corresponds to the assumption that the capacity of a road scales linearly with the number of lanes. Such an assumption is common in Spain and in the US [10]), although slight deviations from this rule are used in Germany [9]. However, these deviations are smaller than those related with road slope and type of traffic, so they can be disregarded here. The inset of Fig. 1(b) shows that average speeds are very similar along the M30 motorway during all recording periods, although the numbers of lanes vary between two and five and recording periods are during different seasons of the year. Similar results hold for flow and density. Specifically, Pearson cross-correlation coefficients between the number of lanes and average velocities, average flows, and average densities (normalized to a two-lane road) are $+0.27$, -0.24, and -0.30, respectively. Although these coefficients are not zero and data from different road widths are not fully equivalent, we can combine the data from all 23 detectors and from all six recording periods (18 weeks) to improve statistics.

The traffic state classification according to MARZ [8] used for Figs. 18.1(a,b) indicates free-flow traffic (state 1, black, 67.5 percent of all data points) in a quite narrow regime with velocities above 80 km/h and densities below 24-30 veh/km (depending on the number of lanes, figures re-calibrated to two lanes). At high speed (> 80 km/h), a larger density leads to the classification as dense traffic (state 2, red, 16.0 percent). Traffic states 3 and 4, viscous traffic ('zähfließender Verkehr', green, 9.1 percent) and traffic jam (blue, 7.4 percent), respectively, frequently occur

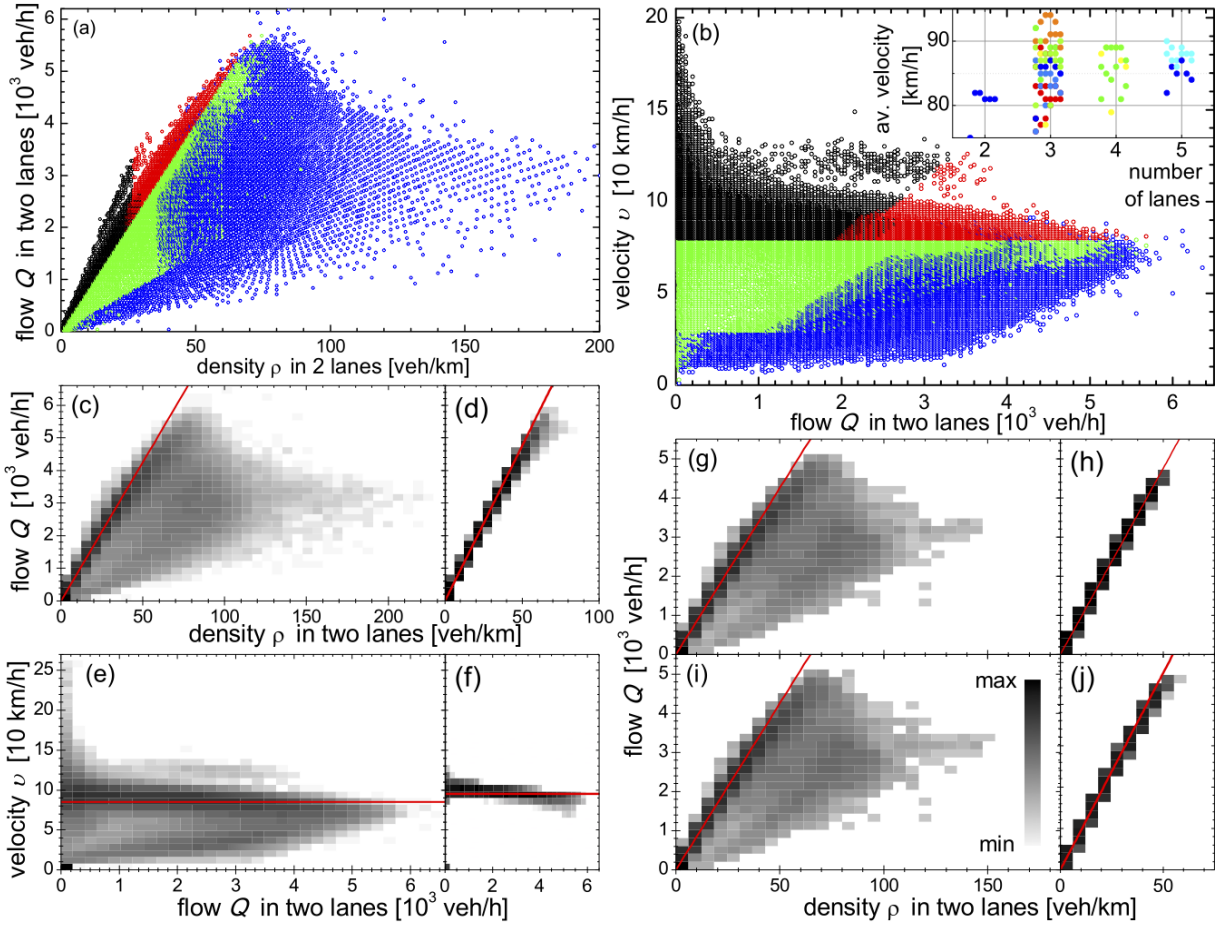

Figure 18.1.: **Fundamental diagrams:** (Color online) (a) Flow-density and (b) velocity-flow diagrams of all M30 data (≈ 2900 days). Each dot represents one minute of data; all data is converted to values for roads with two lanes. The dots are colored by the index of conventional traffic states [8]: free traffic (black), dense traffic (red), viscous traffic ('zähfließender Verkehr', green), and traffic jam (blue). The inset in (b) shows the number of lanes and the total average velocity for each of the 23 detectors in each of the six periods of data recording. The color of the dots corresponds to the positions of the detectors along the motorway. (c-j) Probabilities of occurrence (black – frequent occurrence, light gray – sparse occurrence, logarithmic gray scale is used) in both types of fundamental diagrams for (c,e,g,i) the full M30 data (see Section 2) and (d,f,h,j) our simulations (see Section 3). For (c-f), the original time resolution of 1 minute is kept, while the data have been averaged in non-overlapping segments of 15 minutes for (g-j). In (g,h) average density in each 15-minute segment is calculated by dividing average flow by average velocity, while averaging is performed over 1-minute densities in (i,j). The straight red lines correspond to velocities of (c,e,g,i) 85 km/h and (d,f,h,j) 95 km/h; they are shown for comparison.

in our recorded data with densities up to 200 veh/km in a traffic jam. In general, these two regimes are separated by a velocity threshold of 30 km/h and a density threshold of 36-60 veh/km according to MARZ. However, due to the varying number of lanes in our dataset and due to the history-dependent traffic state classification (see Section 8), there is a large overlap of the regimes associated with states 3 and 4 in the diagram. It is expected that the precise thresholds in the guidelines should be adjusted according to local conditions [8].

The frequencies of occurrence for each pair of flow and density values (or velocity and flow) in our data set are better visible in Fig. 18.1(c,e), where these frequencies are shown in a logarithmic gray scale coding for small boxes in the diagrams. One can see that data points corresponding to velocities around 85 km/h (red line) are most frequent, but also traffic jams with density around 80 veh/km and flow around 3000 veh/h often occur on the considered motorway. Data points with density above 150 veh/km are very sparse. Figure 18.1(e) shows that speeding (velocities above 120 km/h) is not uncommon at very low flow values. Figures 18.1(g,i) show flow-density diagrams for non-overlapping 15-minute averages. Here, no more data points with density above 150 veh/km occur, independent the way of density calculation. This indicates that traffic data with very high density are too sparse and not sufficiently stable, so that no reliable state classification (or even study of persistence properties) is possible for them. For comparison, Figs. 18.1(d,f,h,j) shows the corresponding frequencies of occurrence we achieved in our microscopic traffic flow simulations (see Section 3). They are centered around velocities of 95 km/h without

occurrences of traffic jams.

18.3. Microscopic Traffic Flow Model

In a related, though sometimes isolated area of research, microscopic traffic models represent traffic through many individual vehicles with individualized driver models all interacting in a common road network. The practical argument for their use is changes in e. g. road infrastructure at small levels that result in small changes in vehicle movements that will hopefully yield improvements at a larger level. But at a theoretical level much work has also been done to find microscopic models that either analytically (e.g. [19]) or otherwise (e.g. [20]) implicitly fit to a macroscopic model. In the implicit category are models that are highly complex and comparisons to macroscopic theory tend to be qualitative. Although it is standard practice to compare the numeric results with macroscopic data (calibration), this is usually performed for short periods of time (in the range of hours) and often for one traffic state relevant to the question at hand. Comparisons with real long-term data are rather sparse. Traditional methods of comparison (e.g. root-mean-square error) also do not examine the long-term properties of the data. To our knowledge, there are no guidelines or publications advocating calibration through indicators from long-term time-series analysis. This may have something to do with the time-intensive need for microscopic calibration in the first place: a different set of parameters is deemed necessary for different situations, hence considering wider timescales might be seen as counter-intuitive.

For our work, data for several points on a straight two-lane road are simulated for different traffic flows using the commercial software VISSIM 5.30 (PTV AG, Karlsruhe, Germany), which is commonly employed by traffic engineers. Separate simulations were run with input flows $\bar{Q} = 100$ veh/h, $\bar{Q} = 200$ veh/km, ... until VISSIM reported that not all vehicles could be input to the road. This occurred at $\bar{Q} = 4300$ veh/h. Each simulation lasted 1060 min., i. e. 1000 min. plus 1 hour 'warming up' time at the beginning (ignored in data analysis). The dataset is made up of 19 detectors, one every 250 m; the data is collected at one-minute resolution. The probabilities of occurrence for these data are shown in Figs. 1(c,e,h,j). As there are no bottlenecks or specially pre-programmed driver actions or road conditions that would induce a 'temporary' bottleneck, no points from the congested area of the diagram are present.

The model used in the VISSIM simulator (German acronym: 'Verkehr In Städten – SIMulationsmodell') is described in detail in [21]. Vehicle movements are governed by a car following model designed for motorway traffic and originally introduced in 1998 [21], a lane-change model, and by limiting vehicle accelerations and decelerations. The psycho-physical car-following model considers both physical and psychological limits in the driver's perception and action (e. g. perception of distance to the next vehicle or current speed). VISSIM driver-vehicle units are assigned desired speeds from a distribution, and seek to maintain these speeds (free driving) or maintain a safety distance to a vehicle ahead within perception thresholds, whereby the need to accelerate or decelerate is sensed. Drivers make lane changes to reach their desired speed. Various speed distributions centered on different values are provided. In this work we use the model in an unparameterized default state with a generic, fictional road network and a standard speed distribution between 85 and 120 km/h. Lane changing is governed by the right-side rule, i. e. vehicles do not have free lane selection but only change lane in order to near their desired speed. We note that promoting or optimizing the model is not a goal of our work. We rather want to compare the measured data with such a standard simulation that is typically used by traffic engineers.

18.4. Fluctuations and Cross-Correlations of Flow, Density, Velocity, and Occupancy

First we study the fluctuations of the main traffic parameters (flow, density, velocity) within short segments of the data to identify distinctions in traffic variations between different areas of the flow-density diagram (fundamental diagram). Figure 18.2 shows the average standard deviations of Q, ρ, and v, e. g.,

$$\langle \sigma_Q \rangle = \left\langle \sqrt{\frac{1}{15} \sum_{i=1}^{15} \left(Q_i^2 - \bar{Q}^2 \right)} \right\rangle , \tag{18.1}$$

where $\langle \ldots \rangle$ denotes averaging over all 15-minute data segments with average flow $\bar{Q} = \frac{1}{15} \sum_{i=1}^{15} Q_i$ and average density $\bar{\rho} = \frac{1}{15} \sum_{i=1}^{15} \rho_i$ in the corresponding bin; analogous for σ_ρ and σ_v. One can see that short-term flow variations (Fig. 18.2(a)) are small for very weak traffic (blue) and most pronounced in the lower central area (red, below the dashed line, for $Q < 3500$ veh/h). Short-term density variations (Fig. 18.2(b)) are very small for high velocities and low densities (above the continuous line) and large for large densities (i.e., in extreme traffic jams). Short-term velocity variations (Fig. 18.2(c)) are small at the borders of the diagram (very high or low velocities) and large in the center (between the dashed line and the continuous line, for $Q < 3500$ veh/h). Plots for 4-minute, 8-minute, and 30-minute segments (not shown) look very similar, although the corresponding average SD values

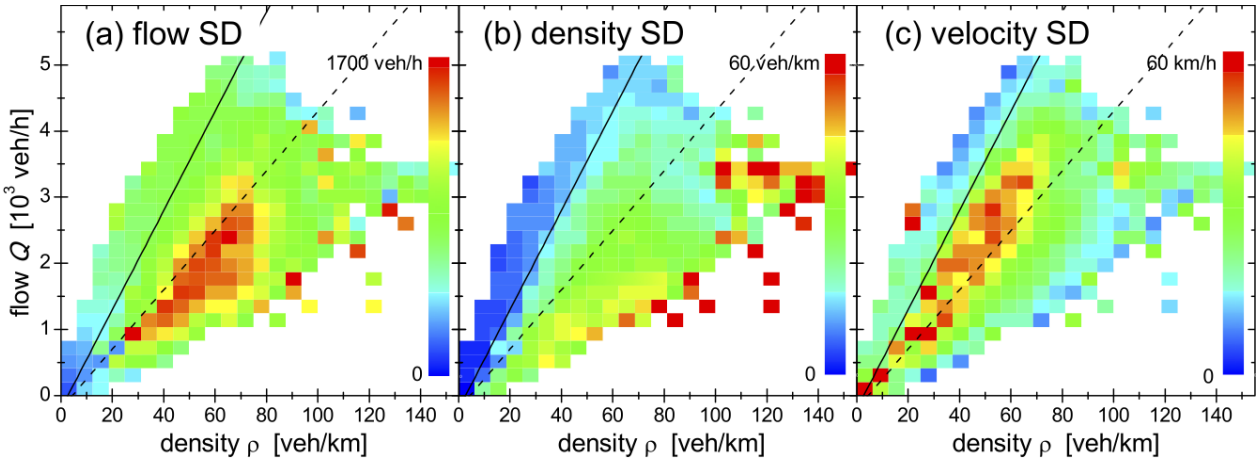

Figure 18.2.: (Color online) Fluctuations (i.e., average standard deviations, SD) of (a) flow Q, (b) density ρ, and (c) velocity v for M30 data, color-coded plots in fundamental diagrams, i.e. versus density $\bar{\rho}$ and flow $\bar{Q}$. Non-overlapping 15-minute segments have been considered to calculate $\bar{\rho}$ and $\bar{Q}$ (to identify the appropriate bin, see text) and to calculate the SD values σ_Q, σ_ρ, and σ_v. The SD values for each bin have been averaged and linearly color coded from 0 to (a) 1700 veh/h, (b) 60 veh/km, and (c) 60 km/h. The continuous line at $Q = (75\text{km/h})\rho - 200$ veh/h and the dashed line at $Q = (45\text{km/h})\rho - 200$ veh/h separate approximate regions of different behavior, see Section 5.

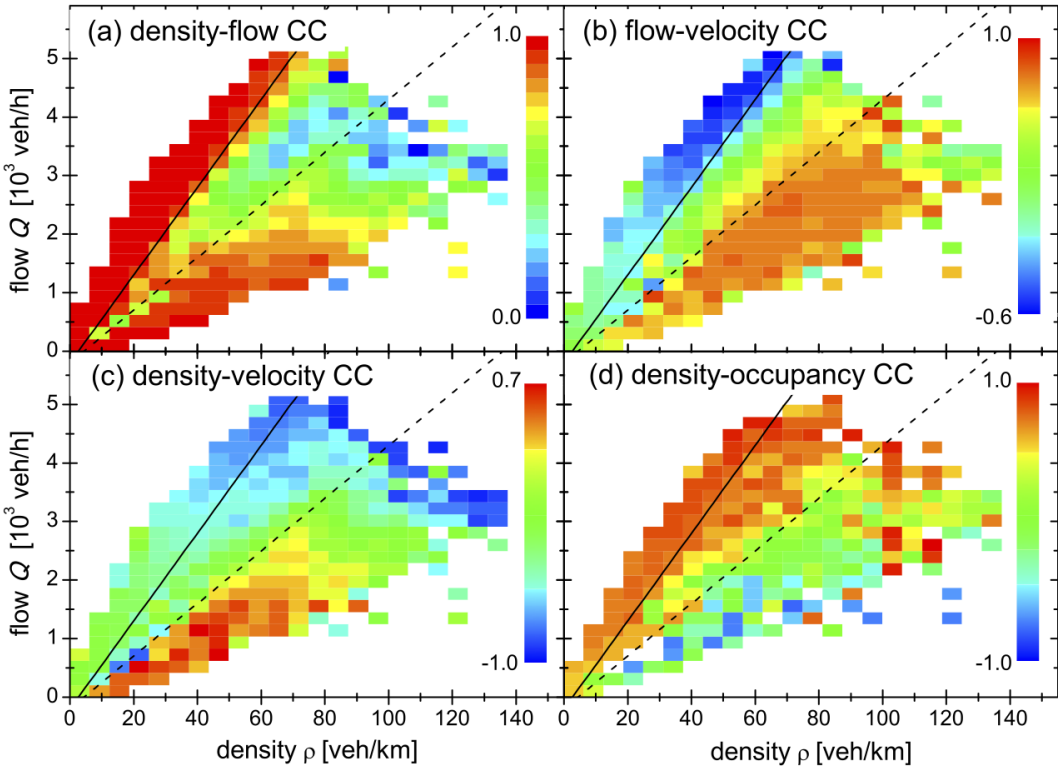

Figure 18.3.: (Color online) Cross-correlations (CC) of (a) flow and density, (b) flow and velocity, (c) density and velocity, and (d) density and occupancy for M30 data, color-coded plots in fundamental diagrams, i.e. versus density $\bar{\rho}$ and flow $\bar{Q}$. Again, non-overlapping 15-minute segments have been considered. The color code is different in each part and given in the corresponding legends. Data from 18 weeks have been used in (a-c), while only two weeks were available for occupancy (d). The continuous line at $Q = (75\text{km/h})\rho - 200$ veh/h and the dashed line at $Q = (45\text{km/h})\rho - 200$ veh/h separate regions of qualitatively different behavior, see Section 5.

are somewhat different. In addition, increasing the statistics by taking into account overlapping data segments does not change the plots, except for a few more results at very high density and low flow (velocities below ≈ 20 km/h). We cannot exclude an additional regime and/or state with increased fluctuations for these very sparse data.

Next we want to quantify the relations between the considered traffic parameters in order to obtain a data-driven

definition of traffic phases. For each 15-minute data segment the cross-correlation function for flow Q and density ρ is defined as

$$\langle C_{Q,\rho} \rangle = \left\langle \frac{1}{15\sigma_Q\sigma_\rho} \sum_{i=1}^{15} (Q_i - \bar{Q})(\rho_i - \bar{\rho}) \right\rangle; \tag{18.2}$$

analogous for $C_{Q,v}$, $C_{\rho,v}$, and $C_{\rho,p}$ with occupancy p. Figure 18.3 shows the type (positive or negative) and strength of these cross-correlations in the fundamental diagram. In interpreting these diagrams we have to keep in mind that flow, density and velocity are related by $Q = \rho v$. Due to the very small fluctuations of the velocity at high and low velocities (above the continuous line and for low flows; see Fig. 18.2(c)), $C_{Q,\rho}$ is very close to 1.0, i.e. strongly positively correlated, in these regions (red in Fig. 18.3(a)). Only in the intermediate regime and for $Q > 2000$ veh/h to the right of the continuous line we see a significant decrease of flow-density cross-correlations, which, however, always remain positive. The picture is drastically different for $C_{Q,v}$ and $C_{\rho,v}$ (Figs. 18.3(b,c)), which are both negative at high velocities (above the continuous line). This indicates that increases in velocity are usually associated with decreases in density and flow. Apparently, people drive faster, when there is more space around them on the motorway.

For $C_{\rho,v}$ the negative cross-correlations remain unchanged in the intermediate regime (between the continuous line and the dashed line) (Fig. 18.3(c)), while $C_{Q,v}$ changes to a positive sign there, however still remaining small (Fig. 18.3(b)). We thus see that flow and velocity become nearly uncorrelated in the intermediate regime in agreement with previous work [12], i.e., drivers' speeds are uninfluenced by traffic flow or traffic state. Large positive cross-correlations of density and flow with velocity are observed in the regime below the dashed lines. For $C_{\rho,v}$, this behavior seems quite paradox, since larger velocities occur when densities are also large (if $Q < 3000$ veh/h). Note, however, that this happens at fairly low velocities ($v < 50$ km/h), while densities are still in the regime of normal traffic flow.

Finally, Fig. 18.3(d) shows density-occupancy cross-correlations in a much smaller subset of the data (see Section 2). Although large and intermediate positive values are observed in the regimes with high velocities (above the dashed line), the coefficients become negative for lower velocities. This observation is surprising, since occupancy is often considered as a proxy for density. Clearly, such an assumption does not hold in the regime below the dashed line, where we can assume stop-and-go traffic (see Section 2, Appendix A of [12], and [4] for discussion). However, even in the free-flow regime at low flow values, density-occupancy cross correlation is fairly low (just around 0.6).

18.5. Empirical Traffic State Classification Scheme

In general, traffic state classification can be attempted either by looking at the current (or running average) values of major quantities like flow, velocity, density and/or occupancy, or by looking at the dynamic properties of the corresponding time series data. The first alternative (denoted by *static classification* in this paper) is chosen in the MARZ process, where thresholds for velocity and density are used to define four traffic states (see Section 2). In addition, the previous state might be taken into account in some cases. In other words, one or more functions of the major quantities define a line or lines that separate traffic states as *regimes* in the fundamental diagram. Such function(s) can also be seen as macroscopic traffic models.

The second alternative (denoted by *dynamic classification* in this paper) is used much less often, since (i) it requires more evolved data analysis techniques to define traffic states and (ii) more or less extended time segments of data are needed to quantify dynamic properties, so that time resolution of the state classification is reduced. The study of fluctuations and cross-correlations in the previous Section 4 of this paper, however, can be regarded as a systematic approach towards such a dynamic *state* classification.

In literature, it is universally agreed that there is a free-flow regime which exhibits an approximately linear relation between density and flow [2]. The simplest and earliest model for the fundamental diagram is one parabolic regime (or phase) [22] – the relation has a linear relation followed by an inversion at a maximum flow (road capacity) and a descent into jammed traffic. Many models operate under two or three regimes. Three-phase traffic theory, developed by Kerner [23], proposes a second linear regime of 'wide moving jams'. Since free flow and wide moving jams can be identified by looking at the values of flow and density, they represent a static classification. All other points around the regime of wide moving jams are termed 'synchronized flow' [3], and are for Kerner not part of a fundamental diagram approach because no unique relationship exists between the variables. Synchronized flow is thus a typical example of a dynamic state classification. Helbing has led criticism of Kerner's theory (see, e.g. [24]) arguing that all points in the latter two phases can be explained with one regime (i. e. the second phase). The scattering is, according to Helbing, due to variations in desired time-gaps between drivers, which, when applying simulation, can be resolved by utilizing varying parameters: the same form of function is at work but many times over according to different driver and vehicle types.

Our work is not based on any theory or pre-set number of 'phases' or traffic states. Instead, we suggest a dynamic traffic state classification based on our empirical observations of fluctuations and cross-correlations of the traffic variables reported in the previous section. Furthermore, our state definitions are based only on the properties of

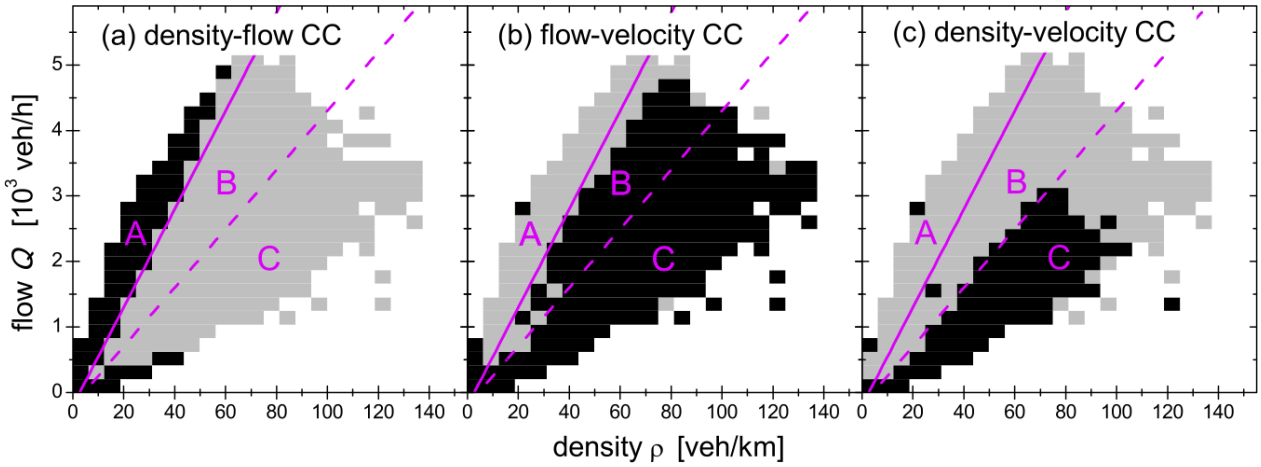

Figure 18.4.: (Color online) (a,b,c) Cross-correlations as in Fig. 18.3 with drastically simplified gray scale coding. In (a), $C_{Q,\rho} > 0.95$ is marked in black (gray for lower values). In (b) and (c), positive (negative) values of $C_{Q,v}$ and $C_{\rho,v}$ are marked in black (gray), respectively. The purple continuous lines at $Q = (75\text{km/h})\rho - 200$ veh/h and the purple dashed lines at $Q = (45\text{km/h})\rho - 200$ veh/h separate regions of qualitatively different behavior.

state	$C_{Q,v}$	$C_{Q,\rho}$	$C_{\rho,v}$	segment	fraction	in A	in B	in C
1	≤ 0	> 0.95	≤ 0	4 min.	54.0%	**57.0%**	25.1%	2.9%
2	> 0		≤ 0	4 min.	10.5%	8.8%	**32.4%**	32.6%
3	> 0		> 0	4 min.	29.2%	28.6%	18.8%	**59.3%**
other	≤ 0	≤ 0.95		4 min.	6.3%	5.6%	23.7%	5.2%
1	≤ 0	> 0.95	≤ 0	15 min.	59.5%	**63.8%**	9.7%	0.1%
2	> 0		≤ 0	15 min.	13.0%	10.8%	**42.0%**	38.9%
3	> 0		> 0	15 min.	21.8%	20.8%	14.8%	**59.9%**
other	≤ 0	≤ 0.95		15 min.	5.7%	4.6%	33.5%	1.1%
1	≤ 0	> 0.95	≤ 0	60 min.	52.8%	**57.3%**	3.2%	0.0%
2	> 0		≤ 0	60 min.	12.0%	8.4%	**55.0%**	46.2%
3	> 0		> 0	60 min.	30.1%	30.3%	15.5%	**52.7%**
other	≤ 0	≤ 0.95		60 min.	5.1%	4.0%	26.3%	1.1%
all					100%	92.7%	4.1%	3.2%

Table 18.1.: Definition of dynamic traffic states by cross-correlations (top part based on 4-minute data segments, center part for 15, and bottom part for 60-min. data segments) with frequency of these states in all data and in data falling into each of the three regions A, B, C of the fundamental diagram (static classification, see text).

the current data segment and do not regard particular sequences of traffic states. Figures 18.4(a-c), simplified versions of Figs. 18.3(a-c) clearly show that the observed characteristic sign changes of cross-correlations between the major quantities of traffic flow give rise to a natural (dynamic) definition of three traffic states (Table 18.1).

Figures 18.4(a-c) and Table 18.1 also show that these states 1, 2, 3 (the dynamic classification) can be approximately identified with three regions A, B, C (a static classification) in the flow-density diagram. While the states 1, 2, 3 are defined by the properties of the cross-correlations (Table 18.1), the regions are defined by separating lines at $Q = (75\text{km/h})\rho - 200$ veh/h (continuous lines) and $Q = (45\text{km/h})\rho - 200$ veh/h (dashed lines). We denote the regions by A (above/left of continuous line), B (intermediate region), and C (below/right of dashed line) in the following. We note that different separating lines for the regions will be required for other roads; in particular the slopes of the separating lines (i.e., the velocities 75 km/h and 45 km/h) must be chosen according to characteristic traffic velocities and speed limits. Here, the separating lines have been determined by two main criteria: (i) optimizing the results in Table 18.1, i. e., obtaining the best possible association of the states 1, 2, 3 with the regions A, B, C, i. e., the largest possible sum of the diagonal (fat) percentages, and (ii) making sure that the frequency of data points in regions B and C is sufficiently large, so that persistence analysis is possible for both of them (see Section 6). Our preliminary calculations for other data sets indicate that it may be appropriate to select $\langle v - 10\text{km/h}\rangle$ for the slope of the separation line between regions A and B (continuous lines in Fig. 18.4) and to select $\langle v + 5\text{km/h}\rangle/2$ for the slope of the separation line between regions B and C (dashed lines in Fig. 18.4). These rules yield 75 km/h and 45 km/h, respectively, when we insert the average velocity $\langle v\rangle = 85$ km/h that we found in the M30 data (inset of Fig. 1(b)).

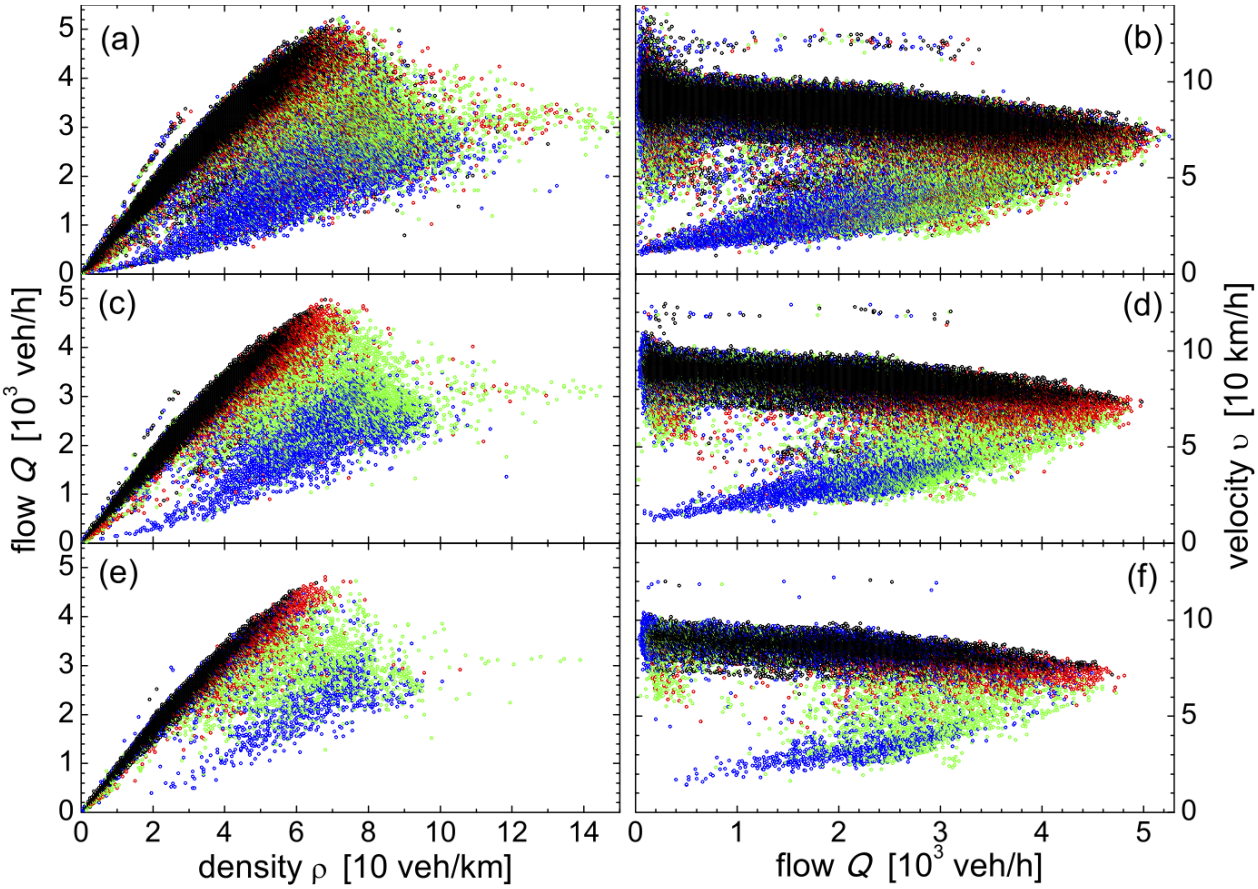

Figure 18.5.: (Color online) Fundamental diagrams of (a,c,e) flow versus density and (b,d,f) velocity versus flow with each dot representing (a,b) 4 minutes, (c,d) 15 minutes, and (e,f) 60 minutes of data (to be compared with Figs. 1(a,f)). The dots are colored according to the values of cross-correlations, see Table 1 — state 1: $C_{Q,v} \leq 0$ (black) $C_{Q,\rho} > 0.95$, state 2: $C_{Q,v} > 0$ and $C_{\rho,v} \leq 0$ (green), state 3: $C_{Q,v} > 0$ and $C_{\rho,v} > 0$ (blue), and remaining data points (red).

Table 18.1 and Fig. 18.5 show that our dynamic state definition is remarkably stable and works quite well also for rather short time segments (4 minutes) and rather long time segments (60 minutes). Note, however, that the identification of states and regions becomes somewhat worse for these extremes (Table 18.1), since 4 data values are very few for calculating reliable cross correlations in Eq. (18.2), and 60 data values often involve non-stationary behavior (and thus problems with the averages in Eq. (18.2)). We thus think that 15 data values are a good compromise. Note also that the high-flow, high-density part of region C is more similar to region B (state 2) than to the rest of region C (Figs. 18.4(c) and 18.5). In region B, a fairly large fraction of data points is classified as 'other' by our state definition (Table 18.1), but state 2 occupies the largest fraction of region B. Overall, a dynamic classification of each individual 15-minute data segment according to its respective cross-correlations (Figs. 18.4(c,d)) yields a fairly good separation of the data points in both fundamental diagrams.

However, there are still significant dependencies of fluctuations and cross-correlations on the flow Q within each region, although dependencies on density or velocity are nearly eliminated by the splitting of the diagram into the three regions. Figures 18.6 and 18.7 show the dependencies of average fluctuations ($\langle\sigma_Q\rangle$, $\langle\sigma_\rho\rangle$, $\langle\sigma_v\rangle$) and average cross-correlations ($\langle C_{Q,\rho}\rangle$, $\langle C_{Q,v}\rangle$, $\langle C_{\rho,v}\rangle$) on flow $\bar{Q}$ for each of the three regions (static classification). In addition, the corresponding results for our microscopic traffic model simulations (Section 3) are included (open symbols, for region A only). According to Fig. 18.6 the fluctuations of all quantities are lowest in region A. The model simulations describe the variations of flow and density very well, but variations of velocity are significantly underestimated (Fig. 18.6(c)). Probably, a slight constraining bottleneck or a broader set of driver parameters would be needed to increase velocity fluctuations in the simulated data. Fluctuations of flow and density are largest for region C (Fig. 18.6(a,b)) and go through maxima for flows between 1000 and 2500 veh/h for regions B and C. Contrarily, velocity fluctuations are largest in region B (Fig. 18.6(c)).

According to Fig. 18.7, cross-correlations of flow, density, and velocity drop linearly with increasing flow in region A (filled black symbols). The model simulations also show such a systematic decrease, but the dependence deviates from the linear form (Figs. 18.7(b,c), open black symbols). As already noted in the discussion of Fig. 18.3, $C_{Q,\rho}$ (Fig. 18.7(a)) is very similar in regions B and C, $C_{\rho,v}$ (Fig. 18.7(c)) is similar in regions A and B, and $C_{Q,v}$ (Fig. 18.7(b)) in region B is intermediate between the behaviors in region A and C. Figure 18.7(d) confirms that

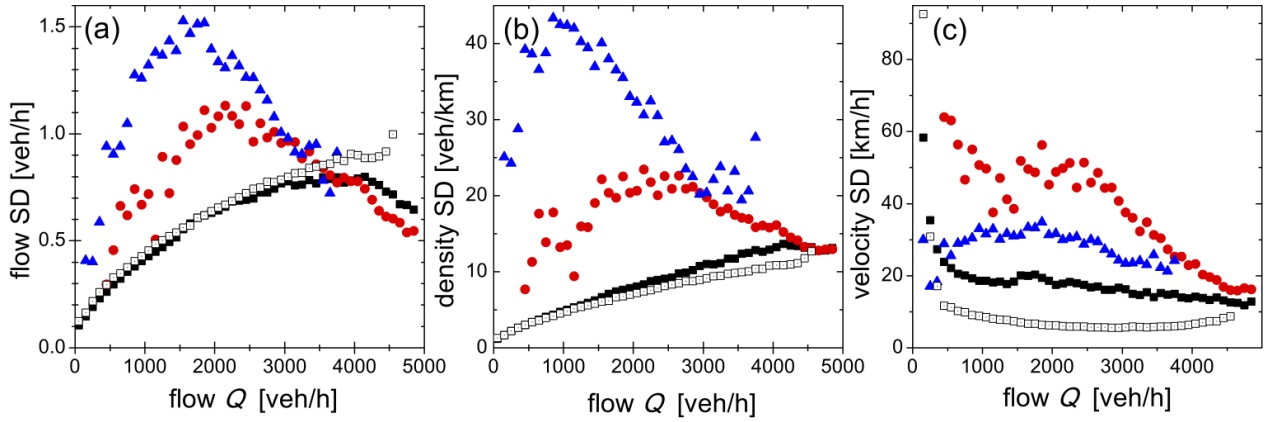

Figure 18.6.: (Color online) Average standard deviations (SD, fluctuations) of (a) flow, (b) density, and (c) velocity for M30 data in regions A (filled black squares), B (filled red circles), C (filled blue triangles), and simulation in region A (open black squares).

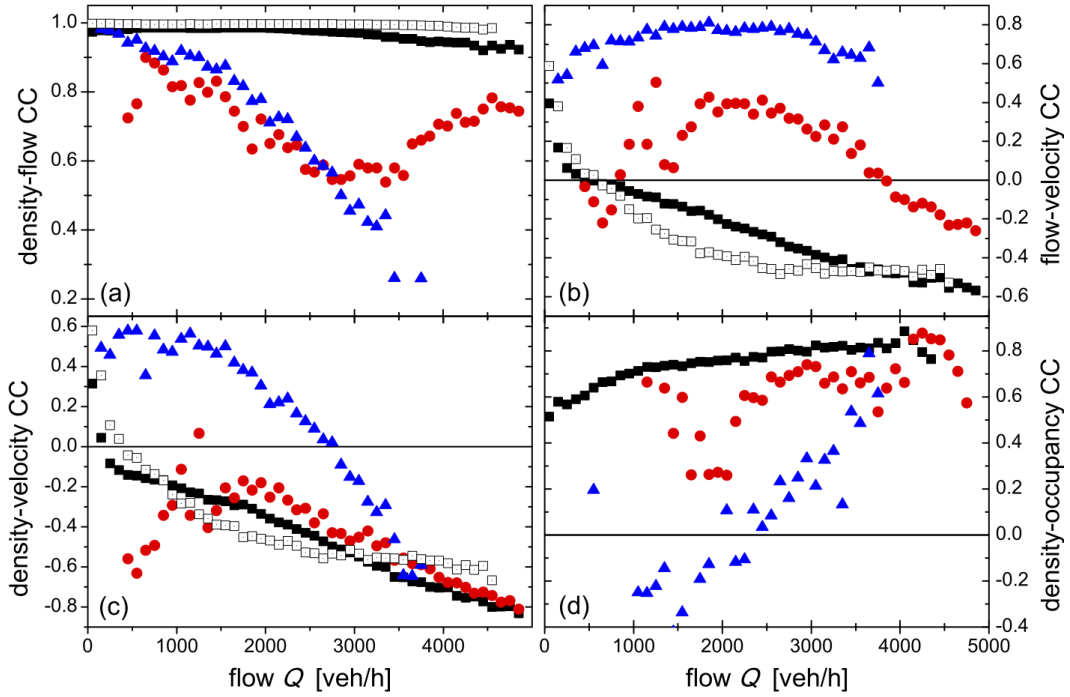

Figure 18.7.: (Color online) Average cross-correlations (CC) of (a) density-flow, (b) flow-velocity, (c) density-velocity, and (d) density-occupancy for M30 data in regions A (filled black squares), B (filled red circles), C (filled blue triangles), and simulation in region A (open black squares, not in (d)) versus flow.

$C_{\rho,p}$ is surprisingly small, particularly at low flow values.

18.6. Scaling Behavior of Flow, Density, and Velocity

Next we want to study the scaling behavior of the traffic variables in the three states (dynamic classification) and the three regions of the fundamental diagram (static classification). We apply Detrended Fluctuation Analysis (DFA) with first, second, third, forth, and fifth order polynomial fitting. The DFA method first introduced by Peng *et al.* [16] for studying DNA sequences has been intensely applied to study correlations in noisy, non-stationary time series. Bunde *et al.* improved it describing higher-order detrending [17]. It has been validated on surrogate (control) time series with correlations and trends [18, 25, 26]. A Detrended Cross-Correlation Analysis (DCCA) for studying long-term cross-correlations based on DFA has also been developed [27].

The method quantifies fluctuations on different time scales s (in minutes, corresponding to the available resolution). For each s the integrated (i.e., cumulated) time series of length N is split into non-overlapping pieces (segments) of length s. Within each segment an n-th order polynomial fit is subtracted, and the remaining mean-square fluctuations are averaged. Repeating the procedure for many time scales s yields the square of the

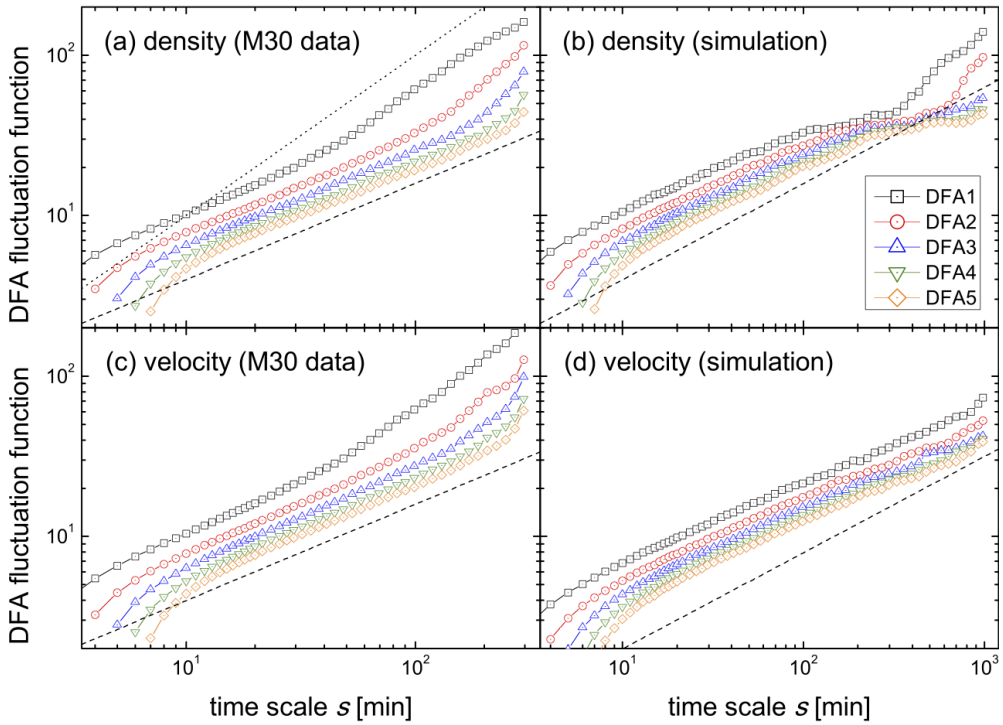

Figure 18.8.: (Color online) DFA fluctuation functions for (a,b) densities and (c,d) velocities from (a,c) M30 data and (b,d) simulations in region A of the traffic diagram. Data segments corresponding to the flow range 1800-2400 veh/h (for two lanes, averaged in intervals of 15 minutes) are taken into account. Different symbols correspond to five different DFAn detrending orders n (legend in (b)). Note that trends in the data affect results for M30 data and simulated density for large s. The straight dashed lines have slope $\alpha = 0.6$ and are shown for comparison; the dotted line in (a) has slope $\alpha = 1.0$ as seen in DFA1 for large s due to daily variations of traffic load.

fluctuation function $F^2_{\mathrm{DFA}n}(s)$, which often scales according to a power-law,

$$F^2_{\mathrm{DFA}n}(s) \sim s^{2\alpha}. \tag{18.3}$$

The exponent α can easily be extracted by linear fits of $\log(F_{\mathrm{DFA}n}(s))$ versus $\log(s)$. Uncorrelated fluctuations lead to $\alpha = 1/2$, while $\alpha > 1/2$ indicates positive long-term correlations (scaling persistence), and $\alpha < 1/2$ indicates anti-correlations. $1/f$ noise, which is typical of critical phenomena [5, 6], leads to $\alpha = 1$. Trends or non-stationarities in the signal, which may be due to time-dependent means of the considered quantity (e.g., daily variations of the traffic load), usually lead to a typical crossover in the fluctuation functions [18, 25]. In this case, the crossover time scale $s_\times$ depends on the detrending order n, and the slope above the crossover is the minimum of $n + 1$ and $p + 1.5$, where p is the order of the (polynomial) trend [18].

Figure 18.8 shows DFA fluctuation functions for density and velocity time series, comparing the results for different detrending orders and for measured data and simulations. One can see the trend-induced crossovers in the scaling behavior for measured density (Fig. 18.8(a)), measured velocity (Fig. 18.8(c)) and possibly also for simulated density (Fig. 18.8(b)), but not for simulated velocity (Fig. 18.8(d)). The 'trends' in the data are actually the variations of traffic load during the daily cycles. These non-stationarities are so strong that non-detrending methods like DFA0 or conventional fluctuation analysis (FA) merely yield a constant (trend-induced) slope of $\alpha = 1$ (not shown). In DFA1-DFA5, below the trend-induced crossover, a consistent slope of $\alpha \approx 0.6$ (dashed lines) can be observed for all detrending orders in Figs. 18.8(a,c). Note that the deviating (larger) slopes on small time scales, which also depend on the detrending order, are a well-known artifact of the detrending methodology [18] and can be overcome by more recent (but less established) detrending analysis techniques, see, e.g., [28, 29, 30].

In DFA3 of the measured data (Figs. 18.8(a,c)), scaling with $\alpha \approx 0.6$ (dashed lines) can be observed on time scales s up to ≈ 150 min. The simulated data, on the other hand, yields smaller slopes $\alpha \approx 0.5$ for all detrending orders (Figs. 18.8(b,d)). Our results show that (i) a detrending procedure is necessary to identify scaling properties of the measured data on time scales above ≈ 20 min., and (ii) there are weak intrinsic long-term correlations (scaling persistence) with α slightly (but significantly) above $1/2$ in the measured data, but not in the simulated data. We obtain similar results for the DFA of traffic flow Q. These observations are not evidently consistent with earlier findings of a slowly decaying auto-correlation function of density and flow and a quickly decaying auto-correlation function of velocity in the free-flow regime [12]. In agreement with the interpretation given in that report, we think that the slow decay of density and flow auto-correlation seen there was induced by the daily variation of

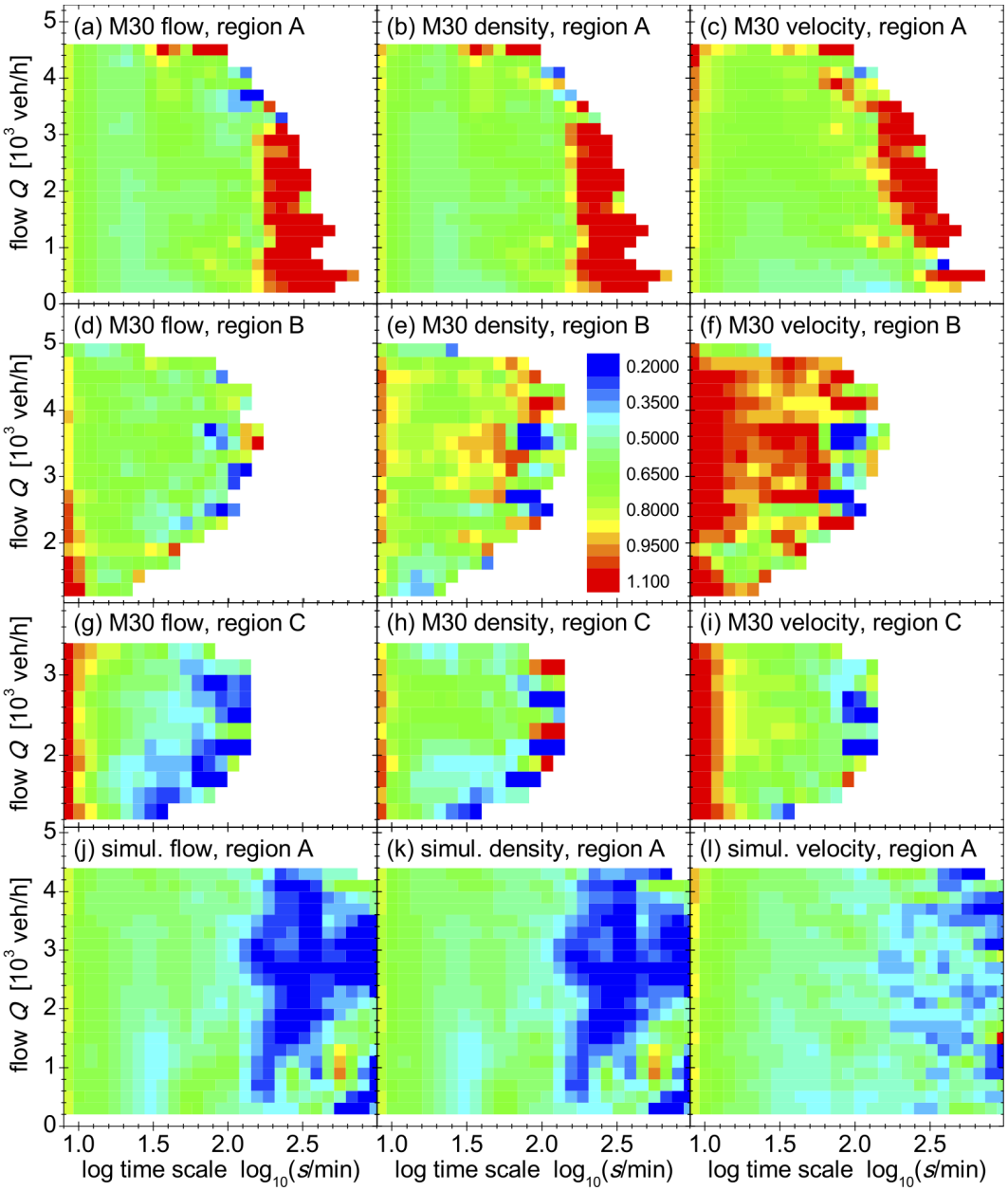

Figure 18.9.: (Color online) Color-coded DFA3 exponents α for M30 data in regions A (a-c), B (d-f), C (g-i), and simulations in region A (j-l), regarding flow (a,d,g,j), density (b,e,h,k), and velocity (c,f,i,l) versus average flow (vertical axis, 15-min. averages in overlapping ranges of ± 300 veh/h for each box) and time scale s (horizontal axis, logarithmic scale from $s = 10$ min. to $s = 1000$ min.). The color code for α is shown in (e).

traffic load. We suggest that a detrending scaling analysis technique beyond standard auto-correlation function is required to identify the weak long-term correlation scaling behavior of measured traffic data. Note that the detrending analysis can differentiate between intrinsic long-term auto-correlations and non-stationarity effects due to, e.g., variation of traffic load.

In order to compare results for different traffic observables and for regions A, B, and C, Fig. 18.9 shows the fitted effective DFA3 scaling exponents for all three regions (see Section 5 for definition) for measured and simulated flow, density and velocity data. We have fitted effective scaling exponents α in plots like those shown in Fig. 18.8, considering scales between $s/\sqrt{2}$ and $s\sqrt{2}$ for each $s = 10 \times 1.2^m$ with $m = 0, 1, \ldots, 26$. The trend-induced crossovers we saw in Figs. 18.8(a,c) thus appear as transitions from green color (α values between 0.5 and 0.75) to red color (α values close to 1.0) at scale $\log_{10} s \approx 2.2$ (corresponding to $s \approx 150$ min.) in Figs. 18.9(a,b). Note that the appearance of this trend-induced crossover is nearly independent of average flow $\bar{Q}$ for flow and density fluctuations scaling (Figs. 18.9(a,b)), while trends are somewhat less significant for velocity fluctuations scaling at low flows (Fig. 18.9(c)). At very large flows ($Q > 3500$ veh/h), the trend-induced crossover is not seen in Figs. 18.9(a-c), since the available data does not include sufficiently long uninterrupted sequences with such large flows, and therefore no fitting results are available for large s at $Q > 3500$ veh/h. This limitation becomes even

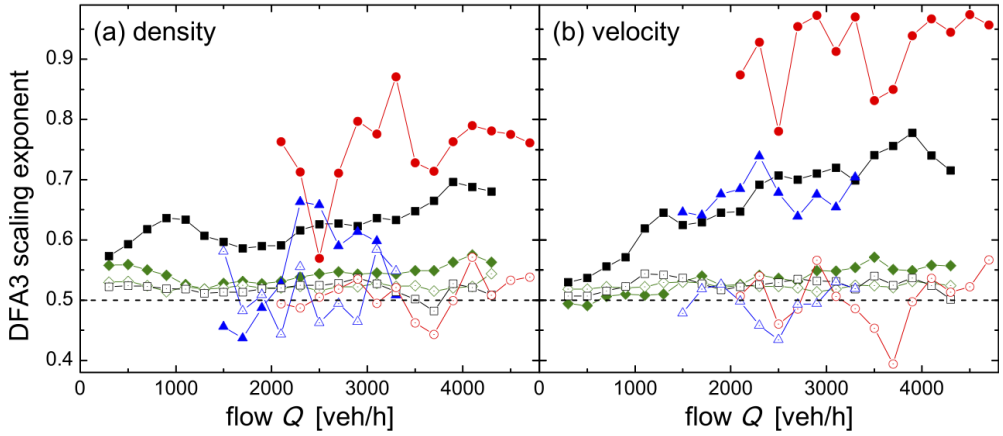

Figure 18.10.: (Color online) Plots of DFA3 scaling exponents α versus flow $\bar{Q}$ for (a) density time series and (b) velocity time series for region A (black squares), B (red circles), and C (blue triangles) and for simulations in region A (green diamonds). Open symbols mark the corresponding results for shuffled data, where $\alpha = 1/2$ (dashed line) is expected. The full scaling range from $s = 10$ min. to 160 min. has been used to fit the slopes in plots of $\log F_{\mathrm{DFA3}}(s)$ versus $\log s$. Significant long-term correlations are seen for the measured data, but not for the simulated data. The strength of the correlations (persistence) is significantly increased in region B and depends on flow for velocity.

more relevant for regions B and C, where uninterrupted sequences with approximately identical $\bar{Q}$ and $\bar{\rho}$ are hardly longer than 100 min. ($\log_{10} s = 2.0$, see Figs. 18.9(d-i)). Trend-induced scaling crossovers are thus irrelevant for regions B and C.

Nevertheless, we find characteristically different patterns of fluctuation scaling behavior in regions B and C compared with region A. In region B, pronounced short-term correlation emerge as can be seen by the yellow and red colors appearing on the left at $\log_{10} s = 1.0$ (corresponding to $s = 10$ min.). In addition, strong long-term correlations appear for velocity time series (extended red areas in Fig. 18.9(f)), but not for flow (Fig. 18.9(d)); density behavior (Fig. 18.9(e)) is intermediate and somewhat ambiguous. In region C, the strong long-term velocity correlations disappear again (Fig. 18.9(i)), so that the velocity scaling behavior becomes similar to region A except for remaining pronounced short-term correlations. The short-term correlations also remain for flow and density in region C (Figs. 18.9(g,h)). However, the slight positive long-term correlations (positive persistence) of flow in region A (with $\alpha \approx 0.6$, see Figs. 18.8(a) and 18.9(a)) is turned into a rather anti-correlated behavior (negative correlations, $\alpha < 1/2$) in region C (blue color in Fig. 18.9(g)). The observed qualitative differences in the long-term scaling properties for the three considered regions A, B, and C of the flow-density diagram retrospectively support this splitting. Note that a much finer splitting will not allow studying long-term scaling properties, since long uninterrupted data sequences are needed for this.

The results for the simulated data in region A (Figs. 18.9(j-l)) are similar to those for the measured data (compare with Figs. 18.9(a-c)), except for a slightly lower value of α in general and the absence of a trend-induced crossover to larger α values at large s. Contrarily to the measured data, we observe a crossover to even smaller (anti-correlated) α values at large time scales beyond $s \approx 150$ min. here. The reason for this opposite crossover may be that the simulations are performed at constant flow, so that variations in flow (and in density) must be compensated by opposite variations later during the simulation run.

Our final step in the scaling analysis is a quantitative comparison and significance test. Figure 18.10 summarizes the scaling analysis results for the original data (measured data and simulated data, filled symbols) and for shuffled data (open symbols) to quantify and confirm the degree and extent of long-term correlations (persistence scaling). Following the conclusions drawn from Fig. 18.9, the full range of time scales between $s = 10$ and 160 min. has been used in fitting the effective scaling exponent α. In the measured data, there are clear long-term correlations (persistences) with α significantly above 1/2. The strength of the correlations (persistence) is significantly increased in region B (red symbols, compared with regions A and C, black and blue symbols). It is quite independent of flow for density (Fig. 18.10(a)), but increasing with flow for velocity (Fig. 18.10(b)). The simulated data in region A (green symbols) does not show significant persistence, since the variations of the effective fitted scaling exponent α are similar as those for the shuffled data (open symbols) for both density and velocity scaling.

The quantitative results for the persistence properties in each of the regions (static classification) are given in Table 18.2 together with the corresponding results for the dynamic classification into the three states. A significance test shows that long-term scaling correlations are significant for measured flow in region A, for measured density in regions A and B, and for measured velocity in all three regions, but not for any of the simulated quantities. In addition, the scaling behavior of velocity is highly significantly different in region B from regions A or C, but differences between regions A and C are never significant.

data	M30 region A	M30 region B	M30 region C	simulation
original flow Q	$\mathbf{0.61 \pm 0.02}$	0.61 ± 0.03	0.58 ± 0.07	0.54 ± 0.01
shuffled flow Q	0.52 ± 0.01	0.56 ± 0.03	0.54 ± 0.02	0.52 ± 0.01
original density ρ	$\mathbf{0.63 \pm 0.03}$	$\mathbf{0.75 \pm 0.07}$	0.55 ± 0.08	0.52 ± 0.01
shuffled density ρ	0.52 ± 0.01	0.51 ± 0.03	0.51 ± 0.05	0.54 ± 0.01
original velocity v	$\mathbf{0.66 \pm 0.07}$	$\mathbf{0.92 \pm 0.06}$	$\mathbf{0.67 \pm 0.03}$	0.53 ± 0.02
shuffled velocity v	0.53 ± 0.01	0.50 ± 0.05	0.49 ± 0.03	0.52 ± 0.01
data	M30 state 1	M30 state 2	M30 state 3	simulation
original flow Q	$\mathbf{0.59 \pm 0.03}$	0.62 ± 0.09	0.55 ± 0.09	0.54 ± 0.02
shuffled flow Q	0.54 ± 0.03	0.52 ± 0.07	0.55 ± 0.09	0.52 ± 0.01
original density ρ	$\mathbf{0.60 \pm 0.03}$	0.68 ± 0.09	0.55 ± 0.10	0.54 ± 0.02
shuffled density ρ	0.53 ± 0.02	0.54 ± 0.04	0.54 ± 0.07	0.52 ± 0.01
original velocity v	$\mathbf{0.63 \pm 0.05}$	0.83 ± 0.21	0.66 ± 0.09	0.53 ± 0.02
shuffled velocity v	0.52 ± 0.01	0.56 ± 0.06	0.54 ± 0.07	0.53 ± 0.02

Table 18.2.: Effective scaling exponents α for original flow, density and velocity data (M30 data and simulated data) and corresponding randomly shuffled data regarding the different regions (A, B, C, top part), and regarding the different states (1, 2, 3, bottom part). Results have been averaged over all available flows; error bars are single standard deviations. The results for shuffled data are all consistent with the expected $\alpha \approx 1/2$. Bold face script marks results for measured data that exhibit highly significant long-term correlations ($p < 10^{-6}$ in t test comparison with shuffled data). Differences of measured and simulated data in region A and in state 1 are significant for all three observables. Differences of measured data in regions A and C and between any states are never significant. Exponents of measured velocity data in regions A and C are highly significantly different from region B; these differences are marginally significant for density.

Regarding the states, significant long-term scaling correlations are found only in state 1 (the free-flow state, see Section 5). The reason is that, although we have used 15-minute segments for the dynamic classification of the states, the corresponding cross correlation results are still too unreliable to yield sufficiently long uninterrupted time series for DFA in states 2 and 3. This observation was actually our main motivation for defining the (static) regions in addition to the (dynamic) states. The actual values of the fluctuation scaling exponents in states 2 and 3 show similar changes as in regions B and C. This is not surprising because of the close relation between states and regions, see Table 1. However, these changes are statistically significant only for the regions (Table 2).

18.7. Prognosis of Traffic Changes and Phase Transitions

Machine learning methods that examine longer time-scales are popular for practical applications, because the hope is to be able to either complete missing data (i.e. establish a reliable traffic flow indicator, but not necessarily a state) and/or forecast the future based on some properties of the past [31]. A thorough review can be found in [32] and a shorter one in [33]. These approaches have become more tractable due to good road measurement infrastructure, data archival capacity, and computer processing power. However, this research concentrates almost exclusively on methods of prediction: both parametric and non-parametric methods can be applied [34] meaning that either predictive equations are derived from data or that the data itself is tested for similarity against past observations [34].

Figure 18.11 explores the predictive value of density fluctuation and flow-density cross correlations. One can see that drastically increased density fluctuations and drastically reduced flow-density cross correlations can predict changes in average velocity in region B, and also – to a much weaker extent – in region A. However, these significant deviations of these quantities from their average values can apparently not predict the *direction* of the upcoming velocity changes. Combining several such indicators of upcoming changes will probably yield empirical predictors for changing traffic parameters (velocity, density or flow) or early indicators for changing traffic states. More work with the empirical data (and possibly also more detailed and calibrated traffic simulations) will be needed to develop such predictors in the future.

18.8. Discussion and Conclusions

The dynamic approach of our traffic state classification is different from most current model- and data-driven approaches, since we have begun by exploring the cross-correlation properties of the data and not initially looked for a static classification by any thresholds or limiting lines involving the absolute values of flow, density or velocity. This led to our definition of three traffic states (1, 2, 3) by cross-correlations, which turned out to be very stable

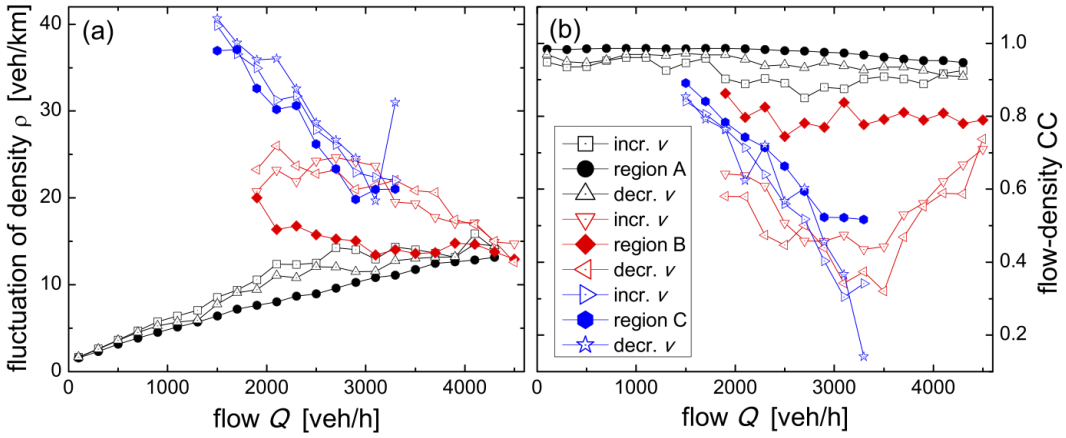

Figure 18.11.: (Color online) (a) Fluctuations of density and (b) cross-correlations of density and flow for M30 data in regions A (black), B (red), and C (blue). Different symbols correspond to data from 15-minute segments followed by increasing average velocity (increase by at least 4 km/h in the next 15 minutes; open symbols: squares, triangles down, triangles right), by similarly decreasing average velocity (open symbols: triangles up, triangles left, stars), or by approximately constant velocity (filled symbols: circles, diamonds, hexagonals). One can see that (a) drastically increased density fluctuations and (b) drastically reduced flow-density cross correlations are usually followed by changes in average velocity in region B, and – to a much weaker extent – in region A.

even for rather different time segment durations between 4 and 60 minutes (Table 18.1, Fig. 18.5). This dynamic classification is nevertheless related to a static classification into three regions A, B, C, since we find each of the three states most frequently in each of these three angular segments (Figs. 4 and 5).

Our dynamic state classification and our static region classification can both be compared with traffic state classifications in literature. As we noted already in Section 5, our region A fully corresponds to the free-low state in other approaches [2, 3, 4]. It combines states 1 (free traffic) and 2 (dense traffic) of the MARZ process classification [8]. We also find that most of the data segments belonging to the free-flow region A are characterized by negative cross correlations between flow and velocity and between density and velocity, while the cross-correlations between flow and density are very large (above 0.95, Table 18.1). Furthermore, fluctuation strength and cross correlations of the main traffic variables in region A indicate monotonous and smooth dependences on the flow Q without any transitions. In particular, σ_Q and σ_ρ increase with average Q (Figs. 18.6(a,b)), σ_v is approximately constant (Fig. 18.6(c)), $C_{Q,\rho}$ is very large (> 0.9) and only slightly decreasing (Fig. 18.7(a)), in agreement with [12], and $C_{Q,v}$ as well as $C_{\rho,v}$ are mostly negative and decreasing (becoming more negative) with increasing Q (Figs. 18.7(b,c)). The scaling behavior of the fluctuations is characterized by weak long-range correlations that slightly increase with Q (Fig. 18.10). Long-term persistences are significant for all three quantities in region A (static classification) and also in state 1 (dynamic classification, see Table 18.2). All of these empirical findings are reproduced in the microscopic traffic model simulations, except of the missing long-term scaling correlations in simulated data. Even non-stationarities, like significant decreases or increases in velocity, hardly affect fluctuations and cross correlations in region A (Fig. 18.11).

Many previous studies assume that a traffic phase transition occurs when traffic flow approaches its maximum value for the motorway [2, 4]. In our data analysis, we do not find clear signs of such a transition even up to very large flows of $Q \approx 5000$ veh/h (for two lanes, see Section 2). This is not really surprising since a traffic breakdown usually occurs quite fast, typically within a few minutes, see e. g. Figs. 15 and 19 in [35]. Since we consider 15-min. averages here, a sudden breakdown would lead to an intermediate point (with exact position depending on the fractions of free-flow and jammed parts – this data point is most probably classified as state type 'other', see Table 18.1 and red points in Fig. 18.5) and then points in the high-density regions B and C. In the large-flow flow range, our results for fluctuations and cross-correlations in regions A and B are fairly similar (Figs. 18.6, 18.7(c,d)). Only $C_{Q,\rho}$ and $C_{Q,v}$ are a bit closer to zero in region B than in region A (Figs. 18.7(a,b)), which may indicate an approaching breakdown of stable traffic flow self-regulation. However, in region B we see a significant increase in the effective long-term correlations scaling exponent for velocity (Figs. 18.9(c,f), 18.10(b)), and – to a weaker extent – also for density (Fig. 18.10(a)). This change corresponds to an increasing persistence of velocity (and density) fluctuations with increasing density at top flow values. The increase in persistence may be related with critical slowing down around a phase transition. In this sense, our region B may be associated with the critical phase around a phase transition. If, on the other hand, region B was regarded as a phase of its own, the similarities of fluctuations and cross-correlations at top flow values in regions A and B might suggest that the transition from region A to region B (towards more persistent fluctuations) corresponds to a second order phase transition at this point.

We find that most data points in region B belong to state 2 which is defined by positive cross correlations of flow and velocity and negative cross-correlations of density and velocity, irrespective of the considered time segment duration (Table 18.1). When comparing our (static) region B and (dynamic) state 2 with classifications in previous literature and practice, we see a close similarity with MARZ state 3 (viscous flow, 'zähfließender Verkehr', see Section 2). Clearly, this regime is different from free flow and describes congested traffic at relatively large flow and intermediate density. Nevertheless neither our region B nor our state 2 can be equated with 'wide moving jams', the second phase of Kerner's three-phase traffic theory [23, 3]. The reason is that 'wide moving jams' represent a linear regime with a clear dependence among average traffic variables down to very low flows, which we cannot confirm in our empirical data. At least parts of our region B and state 2 must thus be considered as belonging to Kerner's third phase, 'synchronized flow'. Hence, our empirical approach does not yield indications of systematic differences between Kerner's second and third phase; the same can be said regarding our region C and state 3. Although this conclusion can be seen as a support for Helbing's criticism of three-phase theory [24, 4], the reason for it could also lie in our restricted time windows (15 min.) and/or our disregard of spatial correlations, which disallows distinguishing the spatio-temporal patterns associated with synchronized flow.

The transition from region A to region B (or, as happens more frequently, from B to A) at flows below the maximal flow $Q \approx 5000$ veh/h is significantly different. Here, fluctuations σ_Q, σ_ρ, and σ_v increase drastically (Fig. 18.6), $C_{Q,\rho}$ drops drastically (Fig. 18.7(a)), and $C_{Q,v}$ changes its sign (Fig. 18.4(b), 18.7(b)). Only $C_{\rho,v}$ is unaffected by the transition from region A to region B (Fig. 18.4(c)). Such abrupt changes in characteristic quantities suggest a first-order phase transition, in agreement with the conclusions in previous work [2, 4]. Note that the effective long-term correlation scaling properties in regions A and B hardly depend on flow in the range where these two regions are in contact (Figs. 18.9(b,c,e,f), 18.10), so that they may be regarded as different dynamical phases. Region B (and also state 2, see Table 18.2) are also characterized by the observation that increased fluctuations and/or decreased cross-correlations can predict significant decreases or increases in velocity (Fig. 18.11). This effect does not occur in regions A or C.

For region C, we find that most data points belong to state 3 which is defined by positive cross correlations of flow and velocity and density and velocity (Table 18.1). Region C is again characterized by only weak long-term correlations, i.e., a weak long-term persistence of density and velocity (Figs. 18.9(h,i), 18.10), similar to region A and different from region B. Note, however, that there are strong short-term correlations in region C unlike region A (Figs. 18.9(a,c,g,i)). Therefore, and because of large absolute fluctuations (Fig. 18.6(a,b)), region C must be considered as another characteristic dynamical phase. It is quite similar to region B at large flows ($Q \approx 3000$ veh/h), but $C_{\rho,v}$ has a different (positive) sign compared with region B at lower flows (Figs. 18.4(c), 18.7(c)). Again, our findings may suggest that the transition between regions B and C is second order at large flows and first order at lower flow values. Based on our empirical results, it seems that only data points at velocities below 45 km/h (dashed line) and flows below $2500 - 3000$ veh/h should be regarded as typical traffic jam regime (blue points in Fig. 18.4(d)) in contrast to the MARZ classification, where a much larger area of the fundamental traffic diagram is considered as traffic jam regime (Figs. 18.1(a,f)).

In summary, our data analysis yielded stylized facts that support the classification of traffic data into three states with different dynamical properties (different cross correlations between the major variables). These state can be identified with free flow, viscous traffic, and traffic jam. They are closely related with three regions in the fundamental diagram defined by separation lines at $Q = \langle v - 10\text{km/h}\rangle\rho - 200$ veh/h and $Q = \langle v + 5\text{km/h}\rangle\rho/2 - 200$ veh/h, respectively, for two-lane roads. State classification in the free-flow, linear area (our region A and state 1) is relatively straightforward and in agreement with other approaches. We also found further two regions (B and C) and states (2 and 3) that are distinct from free traffic. The two differ mainly by the sign of density-velocity cross correlations and by the presence of pronounced long-range correlations (scaling of density and velocity fluctuations) in region B, while there are only short-term correlations in region C.

In further work, our approach should be tested on data from different countries and data from roads with different topology, e.g., comparing results for traffic flow in straight roads with those for merging points and studying single-lane data instead of data accumulated over several lanes. Even more information could be gained from single-vehicle data, which allowed the calculation of additional quantities to provide a clearer overall picture. In addition, a more evolved microscopic traffic model with realistic road topology, bottlenecks and calibrated flow conditions should be employed to study what is required to reproduce realistic long-term correlated fluctuations of traffic flow, density and velocity.

Acknowledgment

The authors thank the European Union project SOCIONICAL (FP7 ICT, grant no. 231288) for financial support. We also would like to acknowledge the Govern Area for Environment, Security and Mobility of Madrid City Council, having provided us with real traffic data from the M30 motorway. We thank Mathias Baur, Slavica Grošanić and Tobias Schendzielorz for their helpful suggestions that improved the paper.

References

[1] P. Ball, Why society is a complex matter (Springer, Berlin, 2012).

[2] A. Schadschneider, D. Chowdhury, K. Nishinari, Stochastic transport in complex systems: from molecules to vehicles (Elsevier, Amsterdam, 2010).

[3] B. S. Kerner, Introduction to modern traffic flow theory and control (Springer, Heidelberg, 2009).

[4] D. Helbing, Traffic and related self-driven many-particle systems, Rev. Mod. Phys. **73** (2001) 1067.

[5] P. Bak, C. Tang, K. Wiesenfeld, Self-organized criticality – an explanation of 1/f noise, Phys. Rev. Lett. **59** (1987) 381 .

[6] F. Schweitzer (ed.), Self-organization of complex structures: from individual to collective dynamics (Gordon & Breach, London, 1997).

[7] H. Kantz, T. Schreiber, Nonlinear time series analysis (Cambridge University Press, 2004).

[8] Merkblatt für die Ausstattung von Verkehrsrechnerzentralen und Unterzentralen, Edition 1999 (MARZ '99), (German) Federal Highway Research Institute (Bundesanstalt für Straßenwesen), (Bergisch Gladbach, Germany, 1999).

[9] Handbuch für die Bemessung von Straßenverkehrsanlagen (HBS) (Forschungsgesellschaft für Straßen- und Verkehrswesen, Köln, Germany, 2001).

[10] Highway Capacity Manual (HCM 2010) (Transportation Research Board, Washington, USA, 2010).

[11] D. Helbing, M. Treiber, A. Kesting, M. Schönhof, Theoretical vs. empirical classification and prediction of congested traffic states, Eur. Phys. J. B **69** (2009) 583.

[12] L. Neubert, L. Santen, A. Schadschneider, M. Schreckenberg, Single-vehicle data of highway traffic: a statistical analysis, Phys. Rev. E **60** (1999) 6480.

[13] S. Assenmacher, diwa - Direkte Information und Warnung für Autofahrer - Zwischenbericht: Untersuchung des derzeitigen Meldungsmanagements (Munich University of Technology, 2005).

[14] Y. Kim, Online traffic flow model applying dynamic flow-density relation, Ph.D. Thesis, (Munich University of Technology, 2002).

[15] C. Deng, F. Wang, H. Shi, G. Tan, Real-time freeway traffic state estimation based on cluster analysis and multiclass support vector machine, International Workshop on Intelligent Systems and Applications (ISA 2009).

[16] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E **49** (1994) 1685.

[17] A. Bunde, S. Havlin, J. W. Kantelhardt, T. Penzel, J. H. Peter, K. Voigt, Correlated and Uncorrelated Regions in Heart-Rate Fluctuations during Sleep, Phys. Rev. Lett. **85** (2000) 3736.

[18] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde, Detecting long-range correlations with detrended fluctuation analysis, Physica A **295** (2001) 441.

[19] D. Helbing, From microscopic to macroscopic traffic models, in: A perspective look at nonlinear media, Lecture Notes in Physics **503** (1998) 122.

[20] S. Krauß, Microscopic modeling of traffic flow: investigation of collision free vehicle dynamics, Ph.D. Thesis, (Universität zu Köln, 1998).

[21] M. Fellendorf, P. Vortisch, Microscopic traffic flow simulator VISSIM, in: J. Barceló, (Ed.) Fundamentals of traffic simulation, Internat. Series in Operations Research and Management Science (Springer, 2010), p. 63.

[22] B. D. Greenshields, A study of highway capacity, Proc. Highway Res. Record **14** (1935) 448.

[23] B. S. Kerner, Three-phase traffic theory and highway capacity, Physica A **333** (2004) 379.

[24] M. Schönhof, D. Helbing, Criticism of three-phase traffic theory, Transport. Res. B **43** (2009) 784.

[25] K. Hu, P. C. Ivanov, Z. Chen, P. Carpena, H. E. Stanley, Effect of trends on detrended fluctuation analysis, Phys. Rev. E **64** (2001) 011114.

[26] Z. Chen, P. C. Ivanov, K. Hu, H. E. Stanley, Effect of nonstationarities on detrended fluctuation analysis, Phys. Rev. E **65** (2001) 041107.

[27] B. Podobnik, H. E. Stanley, Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Nonstationary Time Series, Phys. Rev. Lett. **100** (2008) 084102.

[28] A. Bashan, R. Bartsch, J. W. Kantelhardt, S. Havlin, Comparison of detrending methods for fluctuation analysis, Physica A **387** (2008) 5080.

[29] J. Ludescher, M. I. Bogachev, J. W. Kantelhardt, A. Y. Schumann, A. Bunde, On spurious and corrupted multifractality: The effects of additive noise, short-term memory and periodic trends, Physica A **390** (2011) 2480.

[30] A. Y. Schumann, J. W. Kantelhardt, Multifractal moving average analysis and test of multifractal model with tuned correlations, Physica A **390** (2011) 2637.

[31] F. Maier, Abschnittsweise Regressionsanalyse zur Schätzung von Verkehrskenngrößen, Ph.D. Thesis, (Munich University of Technology, 2010).

[32] J. Guo, Adaptive estimation and prediction of univariate vehicular traffic condition series, Ph.D. Thesis, (North Carolina State University, 2005).

[33] D. Hussein, An object-oriented neural network approach to short-term traffic forecasting, Eur. J. Operational Res. **131** (2001) 253.

[34] Traffic flow forecasting using approximate nearest neighbor nonparametric regression, Research Report No. UVACTS-15-13-7 (University of Virginia, 2001).

[35] W. Knospe, L. Santen, A. Schadschneider, M. Schreckenberg, Single-vehicle data of highway traffic: Microscopic description of traffic phases Phys. Rev. E **65** (2002) 056133.

Acknowledgments

First of all, I want to express my deepest gratitude to PD Dr. Jan W. Kantelhardt, for his guidance and support during the last years, while I have been working in his group at Martin-Luther-University Halle-Wittenberg as part of the SOCIONICAL project and afterwards, during the phase of finalizing my Ph.D. thesis.

I am also very thankful to Professor Shlomo Havlin, who was guiding me during my visit in Bar-llan University in 2011.

I want to thank Dror Y. Kenett for the impulse he gave me during the DPG conference in 2010. During my visit in Bar-Ilan University in 2011 I have had the chance to collaborate with Dror again, working on the idea of dependency networks.

Furthermore, I highly appreciate the financial support of the Minerva foundation in Germany, and the sponsors of the European Union project SOCIONICAL (FP7 ICT, grant no. 231288). The SOCIONIACAL project was a very inspiring environment to shape ideas in an interdisciplinary context.

Without the SOCIONICAL project, the Minerva short-term research grant, and all the great support and guidance from my supervisors and colleagues, my Ph.D. studies would not have been possible.

I want to thank my dear colleagues, Sebastian Tismer, Anja Kunhold, Berit Schreck, Arne Boeker, and Christian Napierala, who all collaborated with me as part of the research group guided by PD Dr. Jan W. Kantelhardt. Many thanks also to Cristina Beltran Ruiz, Lev Muchnik, and Matthew Fullerton for a nice time during our collaboration in the SOCIONICAL project. Finally, I want to thank Eric Tessenow for the nice time we spent together in several places, in Halle (Saale), Tel Aviv, Jerusalem, Potsdam, Leeds, and several other locations, and most of all, for his friendship.

Thanks a lot to Jan W. Kantelhardt, Tom Wheeler, and Ingo Deutschmann for reading the manuscript of my dissertation and for providing helpful feedback.

Danksagung

Neben den bereits genannten fachlichen Aspekten haben auch die persönlichen Beziehungen zu mir sehr nahestehenden Menschen entscheidend auf diese Arbeit gewirkt.

Meiner Familie möchte ich an dieser Stelle besonders danken, für tatkräftige Unterstützung während beider, diese Arbeit begleitenden Umzüge, und für Verständnis sowie viel Geduld in den Zeiten meiner beruflich bedingten Abwesenheit.

Zu verschiedenen Zeiten ward Ihr für mich da, auf unterschiedliche Weise, entweder als Vorbild prägend, als aktive Helfer - in entscheidenden und kritischen Lebensphasen - unterstützend, oder inspirierend durch spannende Erzählungen und lebhafte Diskussionen. Liebe Doreen, liebe Mary Ann, liebe Lara Sophie, liebe Eltern, liebe Katja, liebe Großeltern, liebe Schwiegereltern, und auch Dir, lieber Silvio möchte ich sagen: "Vielen Dank für Eure Geduld mit mir, für Eure Zuwendung und Freundschaft, und für die *Heimat*, die ich stets bei und mit Euch fand - egal wo wir uns gerade aufhielten."

Dr. Werner Hauck, Dr. Dominique Böhme, Eric Tessenow, Christian Napierala und Peter Barthel danke ich an dieser Stelle für Ihre moralische Unterstützung und für die schönen Zusammenkünfte mit oft tiefgründigen, immer sehr offenen, wissenschaftlich, kulturell und gesellschaftlich geprägten Diskussionen.

Publications / Veröffentlichungen

[1] Mirko Kämpf, Eric Tessenow, Dror Y. Kenett, and Jan W. Kantelhardt. The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks. *PLoS ONE*, 10(12):e0141892, 2015.

[2] Jan W. Kantelhardt, Matthew Fullerton, Mirko Kämpf, Cristina Beltran-Ruiz, and Fritz Busch. Phases of Scaling and Cross-correlation Behavior in Traffic. *Physica A: Statistical Mechanics and its Applications*, 392(22):5742–5756, 2013.

[3] Berit Schreck, Mirko Kämpf, J. W. Kantelhardt, and Holger Motzkau. Comparing the Usage of Global and Local Wikipedias with Focus on Swedish Wikipedia. *ArXiv e-prints: arXiv:1308.1776v1*, August 2013.

[4] Mirko Kämpf and Jan W. Kantelhardt. Hadoop.TS: Large-scale time-series processing. *International Journal of Computer Applications*, 74(17):1–8, 2013.

[5] Mirko Kämpf, Sebastian Tismer, Jan W. Kantelhardt, and Lev Muchnik. Fluctuations in Wikipedia Access-rate and Edit-event Data. *Physica A: Statistical Mechanics and its Applications*, 391(23):6101–6111, 2012.

[6] Mirko Kämpf, Jan W. Kantelhardt, and Lev Muchnik. From Time Series to Co-evolving Functional Networks: Dynamics of the Complex System 'Wikipedia', *Proc. Europ. Conf. Complex Syst. (ECCS)*, Brussels, 2012.

[7] P. Gawroński, K. Kułakowski, M. Kämpf, and J. W. Kantelhardt. Evacuation in the Social Force Model is not Stationary. *Journal Acta Physica Polonica*, 121(2-B):77–81, 2011.

[8] T. Cerquitelli, S. Chiusano, M. Kämpf, L. Bellatreche, B. Catania, M. Golfarelli, G. Guerrini, K. Kaczmarski, Y. A. Ameur, W. Andrzejewski, and others. New Trends in Databases and Information Systems, *Springer International Publishing*, Contributions from ADBIS 2013, 1–13, 2014.

Affirmation / Eidesstattliche Erklärung

I herewith avow and confirm that this PhD thesis has been written only by the undersigned and without any assistance from third parties. Furthermore, I declare that no sources have been used in the preparation of this thesis other than those indicated in the thesis itself. This thesis, in same or similar form, has not been available to any audit authority yet.

Hiermit erkläre ich, dass ich die vorgelegte Dissertation selbstständig und ohne Hilfe Dritter angefertigt habe. Ich habe ferner keine anderen als die von mir angegebenen Quellen und Hilfsmittel zur Erstellung meiner Dissertation verwendet. Den verwendeten Quellen wörtlich oder inhaltlich entnommene Stellen sind als solche gekennzeichnet. Ich erkläre, die Angaben wahrheitsgemäß und in bestem Wissen gemacht zu haben und dass die wissenschaftliche Arbeit an keiner anderen wissenschaftlichen Einrichtung zur Erlangung eines akademischen Grades eingereicht wurde.

Mirko Kämpf

Frankleben, den 18. Dezember 2016

Lebenslauf

Persönliche Daten

Name:	Mirko Kämpf
Anschrift:	Müchelner Str. 23 06259 Frankleben
Tel.:	(01 76) 20 63 51 99
E-Mail:	mirko.kaempf@googlemail.com
Geburtsdatum:	16.04.1975
Geburtsort:	Stollberg/Erzgebirge
Nationalität:	deutsch

Ausbildung und wissenschaftlicher Werdegang

10/2011 - 11/2011	Minerva short-term research grant Physics Department and Minerva Center of Bar-Ilan University Thema: *Information Spread and Phase Transitions in Complex Systems*
10/2009 - 02/2013	Wissenschaftlicher Mitarbeiter am Institut für Theoretische Physik Martin-Luther-Universität Halle-Wittenberg Fachgruppe von PD Dr. Jan W. Kantelhardt Projekt: *SOCIONICAL*
04/2009 - 09/2009	Wissenschaftlicher Mitarbeiter am Institut für Physik Technische Universität Chemnitz Fachgruppe von Prof. Dr. Gerlich *Ionenspeicher und Massenspektroskopie*
10/2008 - 03/2009	Diplomarbeit: Experimentalphysik Thema: *Wellenlängenselektive Heitzung von C_{60} Molekülen*
10/2006 - 03/2009	Diplomstudium *Physik* an der Technischen Universität Chemnitz Abschluss: *Diplom Physiker*
10/2004 - 09/2006	Vordiplom *Physik* an der Technischen Universität Chemnitz
09/2003 - 06/2004	Studium *Physikalische Technologien* and der FH Zwickau
09/2002 - 08/2003	Fachabitur an der Fachoberschule in Aue
09/1997 - 07/1999	Berufsausbildung zum Kommunikationselektroniker (Fachrichtung: *Funktechnik*)
01/1995 - 03/2001	Wehrdienst bei der Bundeswehr
09/1991 - 07/1994	Berufsausbildung zum Tischler
09/1980 - 08/1991	Grundschule und POS in Stollberg/Erzgeb.

Weiterbildung

2002	FU Hagen, berufsbegleitendes Studium: *SQL Datenbanken*
2001	IHK Zwickau, *Nachweis von berufs- und arbeitspädagogischer Qualifikation (AdA)*
2001	FH Deggendorf, berufsbegleitendes Studium: *Softwaretechnologie*

Frankleben, den 18. Dezember 2016