Ludvig Bohlin

Toward higher-order network models
Ludvig Bohlin

Toward higher-order network models
Abstract

Complex systems play an essential role in our daily lives. These systems consist of many connected components that interact with each other. Consider, for example, society with billions of collaborating individuals, the stock market with numerous buyers and sellers that trade equities, or communication infrastructures with billions of phones, computers and satellites.

The key to understanding complex systems is to understand the interaction patterns between their components – their networks. To create the network, we need data from the system and a model that organizes the given data in a network representation. Today’s increasing availability of data and improved computational capacity for analyzing networks have created great opportunities for the network approach to further prosper. However, increasingly rich data also gives rise to new challenges that question the effectiveness of the conventional approach to modeling data as a network. In this thesis, we explore those challenges and provide methods for simplifying and highlighting important interaction patterns in network models that make use of richer data.

Using data from real-world complex systems, we first show that conventional network modeling can provide valuable insights about the function of the underlying system. To explore the impact of using richer data in the network representation, we then expand the analysis for higher-order models of networks and show why we need to go beyond conventional models when
there is data that allows us to do so. In addition, we also present a new framework for higher-order network modeling and analysis. We find that network models that capture richer data can provide more accurate representations of many real-world complex systems.
**Sammanfattning**

**Komplexa system** spelar en viktig roll i våra dagliga liv. Dessa system består av en mängd sammanlänkade komponenter som påverkar varandra. Tä exempelvis samhället med miljarder individer som interagerar, aktiemarknaden med köpare och säljare som handlar med olika typer av värdepapper, eller kommunikationsinfrastrukturer med miljarder telefoner, datorer och satelliter.


Med data från verkliga komplexa system inleder vi med att visa att konventionell nätverksmodellering kan ge värdefulla insikter om det underliggande systemets funktion. För att undersöka effekterna av att använda rikare data analyserar vi nätverksmodeller med högre ordning och demonstrerar varför vi behöver
utöka de konventionella modellerna när det finns data som gör detta möjligt. Dessutom presenterar vi en ny metod som använder högre ordningens nätverksmodellering för att analysera nätverk. Vi visar att nätverksmodeller som använder rikare data kan utgöra bättre representationer av många komplexa system.
## Contents

**Papers** 9

**Preface** 11

**Introduction** 13

I  **The rise of network modeling** 17

*How and why network science came to be*

1  *From simple to complex data* 19

2  *Modeling complex data as networks* 27

II  **Conventional network modeling** 35

*How complex systems are studied using networks*

1  *First-order models* 37

2  *The gap between first-order models and data* 45
III Higher-order network modeling 51

How using more data can benefit network analysis

1 Higher-order models 53

2 Variable-order models 59

IV Conclusions 69

Contributions 73

Words of thanks 75

Bibliography 77
Papers

This thesis is based on the following papers, reprinted with kind permission from the publishers:


Other publication by the author not included in the thesis:

Preface

“Just five more minutes? Please!? I was told to stop browsing the web. It was one of those rare, and unreasonably exciting, occasions in the late 90s when we visited my aunt and I got the chance to use this new thing called the world wide web. Back then, digital information was transferred along the telephone network through a computer modem. This meant that the telephone line was occupied during web usage, and that a rather high fee was charged for said usage. Data was expensive. My search for information on hockey and football players therefore had to stop. Little did I know back then that I was experiencing some of the first steps in the digital revolution.

The digital revolution is now in full swing, and much has happened since the time when I searched for information with the help of a screeching 56K computer modem. If I hadn’t studied natural science in gymnasium, I would never had started as an engineering physicist. And if the web hadn’t started to become practical and sophisticated in the late 1990s, the field of network science, the scope of this thesis, would never have emerged.

Network science itself has taken many steps as it continues to grow. Ever heard of companies like Facebook, LinkedIn, Google or Twitter? Do you know what they have in common? Network science, in one form or another. The emergence of network science has been intricately intertwined with the development of

In the late 90s I was using a search engine named AltaVista, which was purchased in 2003 by the perhaps more well-known company Yahoo. AltaVista was a predecessor of Google - a company that originally built its business on a ranking algorithm that applied methods from network science.
the web and the data-driven solutions related to it. Unlike in the late 90s, today the cost of data is not an issue. Extensive amounts of data can be collected from a wide variety of systems, containing far more than the information about hockey and football players that I once found, and I guess still find, so exciting.

Today we are often data-rich but insight-poor. Many times we have more than enough data; we just don’t know how to use it to gain understanding. The aim of this thesis is to shed some light on methods that can bring insights into data, and I promise that the methods won’t keep your phone occupied.

So please give me at least five more minutes to explain? Please!?

Ludvig Bohlin
Spring 2018
Umeå, Sweden
Introduction

Over the past two decades, research in the field of network science has focused on modeling data from complex systems in conventional network models. In recent years, increasing data availability has opened up the possibility of expanding the conventional model. However, figuring out how to use increasingly rich data to create and analyze network models is still a domain in its infancy. The resulting lack of insight is significant because knowing how to simplify and highlight important interaction patterns in network models from richer data will open the door to more effective analysis of complex systems. To address this problem, we will explore the effects of going beyond conventional network models and provide some concrete methods on how to create and analyze network models from richer data.

Thanks to progress made within data processing and computation, today we can collect extensive data on a wide variety of complex systems. For example, by using flu diagnoses reported by health care providers, we can collect data on how diseases spread geographically, and by using mobile phone tracking systems, we can collect data on how people move on their way to work. If we could simplify and highlight important interaction patterns in these systems to better predict the connection between outbreaks and people movements, and rapidly select appropriate interventions in response, the societal benefits would be enormous [6].
An effective method for analyzing data from complex systems is to model them as networks. Numerous studies have shown that the network approach can help us to provide insights into the structure, dynamics, and function of complex systems [7, 8]. The increasing availability of data on complex systems creates a great opportunity for this approach to further prosper.

But increasingly rich data also gives rise to new challenges that question the effectiveness of the conventional network approach. Studies have shown that, when conventional approaches aggregate different types of data into a single network, they destroy higher-order information such as multilayer interactions and multistep pathways [9, 10]. Consider, for example, social relationships where the way we interact with relatives, friends, and colleagues may depend on location, time, or means of interaction. If all contact events are aggregated into a conventional network, important information is inevitably lost. Recent evidence also shows that higher-order information is necessary for capturing important phenomena in the dynamics and function of complex systems [11, 12]. This observation raises significant questions: When are conventional network models sufficient and when are they not? With access to more data, how can we create more accurate representations using higher-order network models? How can we provide methods for simplifying and highlighting important structures in higher-order network models?

We need to better understand these questions on how to use richer data such as multilayer interactions and multistep pathways to create more accurate models of complex systems. These models will allow for more effective analysis of real-world systems that can help us gain insight into the structure and dynamics of such systems.

To explore these questions, we use data from real-world complex systems and examine the impact of representing the data using conventional network models, then show why we need to use higher-order models when there is data that allows us to do so. To address the problem of how to create more accurate repre-
sentations using more data, we present a model approach for representing various forms of higher-order network models in a unified way. To analyze these models, we also provide a framework for simplifying and highlighting important structures in higher-order networks.

The goal of the first part of this thesis is to give a gentle introduction and go through the relevant background to explain and contextualize the research that constitutes the second part. The first chapter describes the background on how and why network science came to be, using the perspective of data. The second chapter explains more about complex systems and how they are studied, using the conventional network model. The third chapter goes beyond the conventional approach and explains how using more data in higher-order models can be beneficial. The final chapter briefly concludes the main findings of the thesis.
Chapter I

The rise of network modeling
"Every two days we create as much information as we did from the dawn of civilization up until 2003." This quote, attributed to Eric Schmidt, former CEO of Google, has often been used to illustrate the age of information that we are currently experiencing. Whether this statement is true or not has been debated, but what we do know is that collecting and storing data has been more difficult throughout history than it is today.

The modern information age is tightly connected to the evolution of network science — the scope of this thesis. As new information technologies were developed, previous limitations on how data could be collected and stored disappeared. All of a sudden, the acquired data had a complexity that no one had ever seen before. Standard approaches could make sense out of traditional data, but how could even more information be extracted from complex data?

For an illustration of the data explosion from the search engine perspective: In 1998, Google served over 10,000 queries a day. In the second half of 1999, it was serving 3,000,000 queries a day [13]. Today, it is estimated that about 60,000 search queries are served every second, which translates to over 5 billion searches per day [14].
A short journey to big data

The explosion of data is largely due to the rise of new technology that is capable of capturing information from the world we live in. But data in itself isn’t a new invention, and people have been gathering data for a long time. Historically the process of gathering data was often costly in time and effort, and therefore the goal of the collection had to be decided in advance. Today the conditions are fundamentally different.

Computers, and particularly spreadsheets and databases, have provided us methods to store and organize large-scale data in an easily accessible way. Suddenly, gathering data is no longer costly, and it is possible to collect it without predefined goals. How did we get to this point? To gain a perspective on data throughout history, we will give a brief historical background. A description of all the historic background is beyond the scope of this thesis; instead, we will mention a few important factors that catalyzed the emergence of the new discipline of network science.

Mankind’s main information storage has always been the human mind. One of the first indicators of humans storing data externally dates back more than 30,000 years, when tally sticks were used to record data [15]. It is believed that these sticks were used for recording and documenting numbers, quantities, or even messages. The first record room, or archive, as a predecessor to libraries, is said to have been built more than 4000 years ago, and around 2300 years ago, the Library of Alexandria became the world’s largest data storage center, representing our first attempts at mass data storage [16].

In the 15th century, Johannes Gutenberg invented the printing press and created one of the world’s first printed books [18]. In 1928, German-Austrian engineer Fritz Pfleumer created a method of storing data magnetically, which formed the basis of modern digital data storage technology [19]. In 1991, the world
wide web was made publicly available [20], enabling anyone to go online and upload their own data, or view and analyze data uploaded by other people. A few years later, in 1996, the price of digital storage fell to the point where it was more cost-effective than paper [21], as a result of, for example, flash drives and DVDs. Another major step was taken in 2007, when Apple released the smartphone [22]. The use of personal smartphones has since then exploded, and in 2015 mobile internet use overtook desktop computer use [23].

The journey to big data was enabled by technology developed in the digital revolution, which was itself enabled by building on the developments in the technological revolution. As computational power and storage improved and developed, so too did the possibilities of collecting more data. The term Big Data was coined to denote data sets so large or complex that traditional data processing applications were inadequate.

Big Data has been a buzzword in both business and academia during recent years, and the concept has had a big impact on how society has been directed [26]. But data in itself has no real value. In order for data to have a value, we need to understand it and extract information that can say something about the underlying system it came from.

**Complex data from real-world complex systems**

As new technologies for collecting and handling data were developed, previous limitations on data collection disappeared. Suddenly, data could be extracted from many systems of interest to scientists, systems that were previously considered too complex.

Many interesting systems are composed of individual parts linked together in some way. Some examples include human and technological systems, such as the world wide web. One
approach, when studying those systems, is to look at certain parts and examine their properties separately. This approach works well, but problems can occur when the goal of the study is to understand the emergent properties created by the collective behavior that occurs when all the system's parts interact. Such behavior is generally not simply the sum of the individual parts, which means that examining certain parts is not enough for understanding the whole system [27]. To overcome this problem, data collection has to happen at another level, making the collection more complicated to control due to the number and complexity of parameters to consider. These challenges motivate the approach when a complex system is studied as a whole.

A complex system is commonly defined as a system that consists of interacting components whose collective behavior cannot be explained by the behavior of the individual units alone [29]. The components may act according to rules that may change over time and that may not be easily understood. Despite its frequent use in many disciplines, complex systems lack a precise definition. We will not attempt a formal definition here, except to say that it is usually a system with a large number of components (complexity of size), intricate relationships among components (complexity of interconnection), and many degrees of freedom in the possible actions of components (complexity of interaction) [30]. Complex systems are, by definition, systems that we cannot fully control or predict, and they are neither perfectly regular nor completely random. A complex system's nontrivial structures are indeed difficult to deal with analytically. The standard way of dealing with this problem is to use models and approximations, and therefore data is needed.

It is only possible to collect a small fraction of the data contained in many complex systems. Even in cases where substantial amounts of data can be collected, the complex nature still makes it challenging to use that data to predict and control the system. Complex data from complex systems provides challenges that traditional analysis methods cannot handle [31]. To analyze and better understand complex systems, new methods

The difference between a complex and a somewhat complicated system can be illustrated with an example. Compare a smartphone, a complicated system, with a flock of birds, a complex system. Superficially the birds are all similar and the flock has far fewer members than the smartphone has parts. Therefore, it could be tempting to think that the smartphone is more complex than the flock of birds. However, the flock's collective behavior cannot be explained from the behavior of the individual birds alone. The flock as a whole responds to changes in the environment and when flying, the rules of the flock are fluid since the head of formation often are changed. The smartphone on the other hand is not a complex system since all its parts have strictly defined roles and prescribed interactions [28].
The emergence of network science to study complex data

Network science emerged as a research field thanks to its ability to analyze large data sets from real-world complex systems [32]. The study of networks has a long history, and after the classical study of the Königsberg bridges in 1735 [27], networks in one form or another appeared in various scientific fields, from Kirchhoff’s electrical circuits [33] to Kekulé’s diagrams of chemical structure [34]. In 1878, the mathematical term for a network, graph, was introduced [35], and since then mathematicians have studied the properties of graphs1 [37]. In parallel, and mainly after the 1950s, social scientists used the concepts of networks to understand the impact of social ties on human relations [38]. The first paper in the field of networks is often ascribed to the work of Barabási and Albert in 1999, in which they study the world wide web and its growth mechanisms [39]. Though networks had been reinvented and applied in different fields since they first appeared, modern network science was established with this publication.

1 An important distinction commonly used for the difference between a graph and a network: a network consists of a graph plus some data [36].

While the emergence of network science may appear to have been a rather sudden event, the field was responding to a wider social awareness of the role and importance of networks, and the emergence was a result of the interplay of many factors. One such factor was the availability of computers and communication networks, which allowed us to gather and analyze data on a scale far larger than previously possible [32]. In its earliest form, the analysis of networks focused on small data sets and the properties of individual entities instead of systems. This change of scale enabled network science to go beyond systems where properties were observed using the human eye, and instead explore systems where direct visual analysis observation is hopeless. The development of methods for quantifying large networks was, to a large extent, an attempt to find something
to take the place of the eye in the network analysis of the 20th century [7]. Another factor that helped catalyze the emergence of a new field was its interdisciplinary nature, incorporating interest from many disciplines such as physics, computer science, biology, sociology and psychology.

To expand on the factors that lead to the growing interest of network science during the first decade of the 21st century, two developments are worth exploring further: universal features and maps [8].

The universal features aspect of networks was a key discovery grounded in the fact that, despite the obvious diversity of complex systems, the structure and the evolution of the networks behind each system are driven by a common set of fundamental laws and principles. This discovery meant that network science could use a generic set of mathematical tools to explore a variety of systems — something that before had been done mainly by visual inspection [7].

The development of maps to visualize complex relationships and interaction also played a key role in the evolution of network science [8]. Maps are needed to be able to describe the detailed behavior of a system consisting of hundreds to billions of interacting entities. For the same reason that keeping track of large amounts of data was cumbersome, there were no appropriate methods to map the networks that the data represented. Thanks to effective data-sharing methods and cheap digital storage, the information revolution fundamentally changed the ability to collect, assemble, share, and analyze data relevant to real-world networks. During the first decade of the 21st century, these technological advances resulted in an explosion of map making that had never been seen before. Some examples include the initial maps of the Internet [41, 42], the maps of protein-protein interactions in human cells [43, 44], and the maps of friendships and professional ties [45, 46] created by social network companies [47]. The sudden availability of these maps in the first decade of the 21st century helped catalyze the emergence of
network science.

In summary, while many disciplines have made important contributions to network science, the emergence of a new field was partly made possible by data availability. Given the large amount of new data, maps of networks were developed in different disciplines. These diverse maps helped network scientists to identify the universal properties of various network characteristics. This universality offered the foundation of a new discipline well-suited for analyzing complex data: network science.
Modeling complex data as networks

Key characteristics of the network approach

Behind many complex systems there is a network that defines the interactions between the components. In order to understand the systems, we therefore need to understand the networks. Network-based modeling has quickly emerged as the norm for representing the patterns of interactions in complex systems for both analysis and modeling. A concise summary of why we study networks is provided by Mark Newman, known for his fundamental contributions to the fields of complex networks and complex systems [7]:

"The ultimate goal in studying networks is to better understand the behavior of the systems networks represent."
From data to a network representation

Networks emerge when connections exist between objects. In an informal sense, a network can be simply defined as a representation of those connections. Networks thus can be found everywhere and, in principle, it is possible to build a network for almost any natural or artificial system. While this poses challenges for understanding when different network-modeling approaches are appropriate, it also highlights the potential for using networks in many different disciplines [8]. To better understand complex systems, researchers across the sciences therefore model those systems as networks.

Networks are simplified representations of the complex pattern of interactions in the systems they represent, and they provide an alternative way of looking at data [7]. A network is a structure that consists of a set of nodes, corresponding to specific objects of the system, connected by links, corresponding to interactions between the objects [27]. The simple representation can be expanded to include more specific information of the system. In general, both the nodes and the links can have features related to either the intrinsic properties of the objects and relations they are representing or to the network structure itself. For instance, links between nodes can be either directed or undirected. When, for example, nodes represent persons with e-mail addresses, a directed link can be established from one person to another if an e-mail is sent between them. If the person receiving the e-mail replies, the link is then considered undirected. Links can also be weighted, representing, for example, how many emails are sent between two persons during a time period. An example of a network consisting of 6 nodes and 8 undirected and unweighted links can be seen in Figure 1.

A network representation can be considered a re-organization of data that makes the data suitable for analysis. When representing data in an alternative format, the goals are usually to reveal the main patterns of the data in order to reduce storage space,
facilitate interpretation and visualization, prepare for generalization, regression, or prediction, or enable access to data analysis tools in another domain. To be useful, the re-organization of data as a network therefore must be a simplified version of the pattern of interactions in the complex system it represents. The network-based re-organization has become the standard model for representing rich interactions among the components of complex systems for analysis [48].

**Properties of networks**

Networks from real-world systems are complex, and often consist of thousands or even billions of nodes and links [7]. To answer questions like *Can the large-scale properties in complex systems be explained by an underlying network structure? Is the structure and behavior of complex systems governed by universal laws?* and *How do we design, organize, build, and manage complex networks for functionality?*, a simple inspection of the network representation is not enough. Instead, the network science community has developed measures to quantify and explain different properties and characteristics of networks.

Network measures can roughly be divided into two types: those aiming to measure static properties, and those aimed at characterizing the dynamic properties of the network. The static properties of the structure, or topology, of a network refer to the kind of patterns of connections that exist in the network. An example of four different connection patterns for a small network can be seen in Figure 2. The dynamic properties refer to the processes that can take place in the structure, and the evolution of the network structure. Since static and dynamic properties tend to be dependent on each other, empirically observed static features such as short path-length, high clustering, and nonrandom degree distributions have been followed by network models with the aim of reproducing the property in an attempt to uncover its dynamic origin [39, 49, 50].

![Figure 2: Different examples of network structures. Four ways of representing relationships between six nodes.](image-url)
Constructing a network model is a way to understand the process behind the formation of the network, and thereby to explain the function of the network. Basically, a network model is a sequence of instructions to decide which nodes are connected to which other nodes. Since many aspects of network science have been influenced by the physics approach, a number of measures and models have been developed from the tools, techniques, and mindset of physics. In the physics approach, domain-specific details of a problem are most often detached to isolate and investigate the network’s most fundamental features. The standard procedure when applying this physics approach to networks has been to combine the use of graph theory with the tools and techniques of statistical mechanics \cite{28, 32, 51, 52}. In this context, several descriptive models that attempt to characterize the structure and evolutionary dynamics of real-world network structures have been developed. These models are often created as minimalistic models containing only the essential ingredients needed to obtain real-world network structures, often with random graphs \cite{53} as the underlying null hypothesis for comparison.

While it is common to only refer to the theoretical models as network models, it is valuable to also consider the empirical models, which are simply the network representation of the observed relations in the empirical data. In other words, the system is converted into a network model. Common to empirical and other contexts is that the network models can be used as playgrounds for different processes on the networks, and for analysis of the networks.

**Current state, applications, and impact**

If we want to understand real-world complex systems, we must first understand their basic interaction patterns. To accomplish this, we collect data about those systems and create network models to gain insights about their structure, dynamics, and function \cite{29}. Although network science is a relatively young field, its inherent mix of data-driven, empirical, interdisciplinary
and computational nature has proven to be a successful approach.

Network science provides a foundation that connects and enables interaction between several different disciplines. The challenge of extracting information by characterizing connections in data from complex systems is common to many disciplines. Consequently, network science has an inflow of tools, techniques, and mindsets from different disciplines\(^1\), which in turn has led to a cross-disciplinary fertilization of methods and ideas in the field.

The empirical nature and focus on data and applicability are key characteristics of network science. Many times extensive data is needed to construct networks from systems of practical interest, and therefore network analysis often encounters computational challenges. For this reason, algorithms, database management and data mining methods are actively developed, resulting in a strong computational character. To address these computational challenges, a series of software tools has been developed, enabling a wider community to apply the methods and analyze networks [55–58].

**Criticism of network science**

The applicability of networks to real-world problem has been criticized [59] due to the fact that, by reducing a complex system to a simple network, many times key features that differentiate one real-world system from another are eliminated [60–62]. The problem related to the construction of the network is also affected by the field’s interdisciplinary nature, and the differences in assumptions and methods employed by researchers from different fields. Since the results can be heavily influenced by the underlying network perspective, it is critical to acknowledge the possibility that different researchers with different approaches can arrive at opposite conclusions\(^2\) about the same system [30].

\(^1\) For example, the measure of the centrality of a node in a network first emerged in 1970 in the social network literature [54], and today the measure is used for identifying high traffic nodes on the Internet.

\(^2\) As an example, Albert et al. (2000) [63], used network models to conclude that the Internet is vulnerable to attacks on the most highly connected routers. However, Doyle et al. (2005) [62] later showed the network to be quite robust to attacks on highly connected routers, but vulnerable to attacks of software protocols. These protocols had been abstracted away from the models in the work of Albert et al.
It is therefore important to be careful with assumptions about the underlying problem formulation and solution when applying results from network science to, for example, decision problems.

Network science has also been criticized for its considerable focus on using statistical characterizations in the analyses. This criticism can be divided into three parts. First, networks that share a particular statistical feature can often be quite different, and many statistical descriptions do not uniquely characterize the system of interest. A rigorous method of evaluating when different methods are applicable is still missing, for the most part [64], and the scale-free feature is a good example [65, 66]. Second, it can be hard to infer the underlying process that actually caused an observed feature, since many processes can generate similar networks. Network science has therefore been criticized for bringing forth descriptive rather than explanatory models [67]. Third, network models, like small-world and scale-free models, are often applied when the underlying or implicit assumptions are fulfilled, not whenever statistical indications are found.

The criticism of network science both raises warnings and opens for opportunities. Since network science is still a relatively young field, alternative mechanisms that produce structural patterns are yet to be discovered. An important direction of future work in network theory will therefore be the development and validation of novel mechanisms for understanding the structure of networks generated from real-world systems. Similarly, theoretical results concerning the behavior of dynamical processes running on top of network, may need to be reassessed in light of the genuine structural diversity of real-world networks.

**Societal impact of network science**

There is no doubt that the field of network science has had a significant impact on many modern aspects of society. One of
the most noticeable impacts of network science has been on the business side. Global corporations benefited to great degree from the tools from network science, and many of the most successful companies of this century — Google, Facebook, Twitter, and LinkedIn — base their technology and business on networks.

Network science has also offered new opportunities for health and medicine research. Many processes in the human cell, from food processing to sensing changes in the environment, rely on connections that can be represented in cellular, metabolic and molecular networks. The network modeling approach provides a way to understand how cells function through maps of interactions between genes, proteins, metabolites and other cellular components. The breakdown of such networks is responsible for human diseases, and therefore research in both biology and medicine uses network science to, for example, identify drug targets in bacteria and humans and to develop paths toward the development of future drugs [68].

The network-based framework has also brought fundamental changes to epidemic modeling, offering a new level of predictability. Thanks to fundamental advances in understanding the role of transportation networks in the spread of viruses, the H1N1 pandemic was the first pandemic whose course and time evolution were accurately predicted months before the pandemic reached its peak 2009 [69]. Today epidemic prediction is being used to foresee the spread of influenza and to contain viruses, and it is one of the most active applications of network science [70]. Remarkably, network science also provides tools for predicting the conditions necessary for the emergence of viruses spreading through mobile phones3.

In neuroscience, network methods have also been shown to be helpful. One of the least understood systems from the perspective of network science is the complex human brain, partly because we have no map to describe the hundreds of billions of interlinked neurons. The only fully mapped brain available for research is that of Caenorhabditis elegans, a roundworm ex-

3 The first major mobile epidemic outbreak started in the fall of 2010 in China, infecting more than 300,000 phones each day, and closely following the predicted scenario [71].
tensively used in medical and biological studies, whose brain consist of only 302 neurons [72]. To be able to construct detailed maps of mammalian brains would lead to a revolution in brain science [73], opening the door to increased knowledge about numerous neurological and brain diseases.

Network thinking has also been increasingly present in the work against terrorism [74]. For example, it has been used to disrupt the financial network of terrorist organizations and to map opposing networks, helping to uncover the roles of their members and their capabilities [4]. Many times the work in this area is classified, but some documented case studies have been made public. An example is the use of social networks to find those responsible for the 2004 Madrid train bombings through the examination of the mobile call network [76].

Last but not least, network science itself has also had an impact on the scientific community. To examine this impact, the citation patterns of the two most cited papers in the area of complex systems [39, 49] have been compared to the citations of the five most cited papers in complexity [77–81]. These comparisons show that, in the area of complex systems, the most rapid rise of citations is seen in network science papers [8].

It is worth noting that network science methods can also be misused. An example is the network mapping performed by the US intelligence agency, the National Security Agency (NSA). NSA monitored the communications of hundreds of millions of individuals worldwide, rebuilding its social network with the questionable aim of stopping terrorist attacks [75].
Chapter II

Conventional network modeling
First-order models

The conventional approach for analyzing networks

Modeling complex data as a network is not a straightforward process. The model will depend on what type of data exists, and what properties of the real-world system we want to study. When the system has been modeled as a network, it serves as the foundation for subsequent network analysis.

In network analysis, the complex structure is represented with a network of nodes and links, and the dynamics are modeled with random flow on the network. This conventional modeling approach implicitly assumes a single type of link and that where the flow moves in the network depends only on where it is now — a first-order model.

Modeling real-world data as networks

Network science often relies on empirical datasets. Since the collection of empirical data from real-world systems requires
an experimental procedure, the approach poses challenges and limitations when it comes to collecting data. Challenges and limitations also occur at the point where the empirical data is to be modeled as a network. How should we use the data to create a network that appropriately captures the interactions among the components of a complex system?

Data from complex systems can come in many forms. The most common format is perhaps pairwise data, which is a collection of one-to-one relationships between entities\(^1\). This pairwise data can be both weighted and directed. Temporal pairwise data is another extension of the pairwise data, where time stamps represent when the pairwise relationship is activated. Other types of data include group data, where every observation can include an arbitrary number of elements instead of just two, and sequential data, where observations involve multiple entities, and where the order of elements is important.

For some data, it is not clear what are actually meaningful relationships and interactions, and sometimes it is not even clear what constitutes the nodes. In some cases the node assumption is well-justified, as in the choice of individual publications in citation network studies and the use of individual humans as nodes in studies of friendship networks. In other cases, studies of interactions between aggregates such as groups or organizations can be more complicated due to the shifting nature of the interacting parts and the fact that subunits of a larger part may interact with others parts or subunits [85]. The number of nodes and links can therefore vary widely for different network representations.

In general, networks created from real-world systems are sparse [86], meaning that the number of links in the network is much smaller than the maximum possible number of links. Another feature of real-world networks is that they seldom display a typical degree that is representative of most nodes\(^2\). Instead, the situation is very different, with skewed degree distributions seen in many networks [87, 89]. Another important characteristic of real-world

\(^1\) A common practice when using data from complex systems to create networks is to directly take the sum of pairwise connections in the sequential data as the link weights in the network, for example, the sum of traffic between locations in an interval [82, 83], or for human mobility patterns [84].

\(^2\) Real-world networks are commonly said to have a degree distribution that is scale-free. That is, the fraction \(P(k)\) of nodes in the network having \(k\) connections to other nodes goes for large values of \(k\) as \(P(k) \sim k^{-\gamma}\), where \(\gamma\) is a parameter whose value is typically in the range \(2 < \gamma < 3\) [39, 87]. However, this is often difficult to show empirically due to the finiteness of the data sets [88]. Therefore, it often make more sense to speak about networks with a highly skewed degree distribution.
networks is the presence of communities [32], and the critical aspect that they are often dynamic or temporal [10].

The question of how to accurately model empirical data derived from complex systems as networks is a prerequisite for the subsequent network analysis. Despite this, the modeling seldom receives much attention compared to the network analysis itself. Given the diversity in empirical data, and the fact that different network models can capture different types of information, it is important to consider what connections the network should model. Links can, for example, be created by connecting objects that are spatially or temporally close, by connecting objects that share a common existence, by connecting objects if there is a communication channel between them, by connecting objects based on correlations, or by connecting objects if they refer to each other.

Constructing a representative network model often relies on the existence of good data, and it is important to ask whether the chosen modeling actually preserves the information in the empirical data. Real-world data can provide diverse representations for the same information, but not all representations contain the same information. In the end, network representations are just proxies for a complex system that can be analyzed by various network analysis methods. Appropriate use of network analysis depends on choosing the correct network representation for the available data and the problem at hand.

**Modeling dynamics on networks**

The static topology itself is not enough to fully characterize a network. To understand the function of complex systems, we also have to understand the dynamical processes that emerge from the interconnections between the components in the system [39, 90, 91]. The components depend on the system under study and can be, for example, people or airports. The intercon-
nections often come from the flow of some entity between the components and can represent, for example, messages circulating among people or passengers traveling through airports.

While both network structures and the related dynamical processes can vary between systems, the challenge is to find a method that can cope with the complexity in all dimensions. The approach of implicitly decoupling the network structure from its dynamics [92] provides a way to analyze various complex systems within a single framework. Given a network structure, this approach makes it possible to define a corresponding Markov process \(^3\) by interpreting the network as the state space of a random walker, and assigning the state-transition probabilities according to the link weights \(^4\). With this approach, researchers model dynamical processes such as, for example, people navigating the web [94], rumors moving around among citizens [95], and passengers traveling through airports [96], with random walkers on the network structure.

The random walk on the network corresponds to a first-order Markov approach, and it is the conventional approach for modeling dynamics on the networks. The random walker moves between nodes \(i \in \{1, 2, \ldots, N\}\), and in \(t\) steps, the walker generates a sequence of random variables \(X_1, X_2, \ldots, X_t\). The transition probabilities when moving between nodes in the network

\[
P(X_t|X_{t-1}, X_{t-2}, \ldots) = P(X_t|X_{t-1})
\]

only depend on the previously visited node’s outlinks. If the link weight between nodes \(i\) and \(j\) is \(w_{ij}\), and the total outlink weight of node \(i\) is \(w_i = \sum_j w_{ij}\), the first-order transition probabilities are

\[
P(i \rightarrow j) = p_{ij} = \frac{w_{ij}}{w_i},
\]

which gives the stationary visit rates

\[
\pi_i = \sum_j \pi_j p_{ji}.
\]

To ensure ergodic\(^5\) stationary visit rates, from each node we can

\[^3\] A Markov process is a simple stochastic process in which the distribution of future states depends only on the present state and not on how it arrived at the present state [93].

\[^4\] Note that this requires positive link-weights.

\[^5\] A dynamical system is called ergodic if every state is reachable on a finite number of steps from any other state. As the length of the random walk tends to infinity, the fraction of times that a random walker spends on a node converges to a number different from zero — the stationary probability. When teleporting is used to ensure ergodicity, the resulting stationary probability is also known as PageRank [97].
First-order models of real-world complex systems

In papers I and II, Stock Portfolio Structure of Individual Investors Infers Future Trading Behavior and Mapping Bilateral Information Interests Using the Activity of Wikipedia Editors, we study two real-world complex systems using the conventional network approach. We show that using the network approach can provide valuable insights into the function of the underlying system.

Financial networks from correlation data

In paper I, Stock Portfolio Structure of Individual Investors Infers Future Trading Behavior, we apply networks to financial data from a stock market. To create the network, we consider individual investors as nodes, and construct links between investors according to their correlations in stock portfolios. The aim is to examine two main questions: (1) How do investors in the stock market structure their portfolios? and (2) Can we learn about trading behavior by looking at the investors' portfolio structure?

In recent decades, economists have realized that using only agent-based models is not enough to fully grasp large-scale financial phenomena [99]. A specific example that has attracted attention, especially since the global financial crisis⁶, is the network aspects of financial markets and how network modeling can be used to analyze systematic risks [101, 102]. Traditionally, models of financial markets have assumed that traders make rational decisions about investments. New data and new science on human psychology have both challenged and changed this picture of the rational trader [103, 104]. However, the scarcity

---

⁶ The global financial crisis in 2008 is one example of the complexity of financial markets. Researchers are not even sure if data alone is adequate to model the financial system, and it has been suggested that human behavior makes financial systems chaotic [100].
of data, often due to confidentiality constraints, has limited the possible studies related to financial markets.

In the paper, we investigate the relationship between what stocks investors hold and what stocks they buy. Such studies of individual investor behavior have not been undertaken before due to data limitations. Similar to other financial markets, the Swedish stock market represents a complex system with interconnected buyers and sellers. The data we use from this complex system are provided from the central register of shareholdings in Sweden and constitutes shareholdings of individual investors on a quarterly basis. The data enable us to represent investor holdings in portfolio vectors, and thus compute similarities between investors based on the stocks in their portfolios. With this data approach, we model the market as a network by considering investors as nodes, connected by links representing the similarity of their stock portfolios.

By analyzing the network, we find investor groups that not only identify different investment strategies, but also represent individual investors trading in a similar way. These findings suggest that the stock portfolios of investors hold meaningful information, which could be used to gain a better understanding of the complex stock market system.

**Information networks from aggregated data**

In paper II, *Mapping Bilateral Information Interests Using the Activity of Wikipedia Editors*, we examine an information network that can transmit, disseminate and help discover information. To create the network, we consider countries as nodes and construct links between countries based on the aggregated co-editing of articles. Such information networks have become a reality, with new technology that has provided new ways of accessing and sharing information. Thanks to the digital revolution, we can now communicate with people on the other side of the globe and...
interact in many other ways. Independent of geographical proximity, we can communicate with whomever we want. But does this globalization actually bring people’s opinions and interests come closer together?

To study this question, we analyze extensive amounts of online data to reveal what information matters to which regions. In the paper, the base for the analysis is Wikipedia, the largest online collaborative encyclopedia. Thanks to its structure and popularity, Wikipedia provides a unique opportunity to analyze global data – data that was not available a decade ago.

Editors in Wikipedia are people from various backgrounds and education who contribute by creating and sharing content online. From the IP addresses of editors, we extract their geographical location with a resolution of country level. If the editors from two countries happen to co-occur more than expected in editing certain articles, we create a link between the countries and assign a weight related to the strength of their co-occurrence. This method creates a network with countries as nodes, linked according to how many articles the people from the different countries co-edit. This information network is then analyzed with methods from network science.

The results show that people care about local and regional information related to sports, media, celebrities or local places. Moreover, people from countries with similar language or historical backgrounds care about similar information. For example, in Europe, countries are divided into eight clusters. Scandinavian countries are in one cluster with shared interests, while Portuguese and Spaniards have more interests in common with Brazilians than with other Europeans.

In summary, the tendency to edit Wikipedia articles is affected by language, geographic and historic backgrounds. Although the means of communication have become global, thanks to the information revolution, people’s interests still remain local. This locality in interests sets a limits for how information propagates.
Therefore, the network extracted in the paper could potentially be used to study information spreading in a global complex system in a more realistic way.
The gap between first-order models and data

Challenges of conventional network modeling

The progress made in data collection and processing has led to an increasing availability of data from complex systems and improved computational capacity for analyzing networks. Assuming that a network approach is the correct framework for analyzing data, the challenge is to create a network model from the data, such that the relevant patterns in the data are correctly captured.

If the network is not an adequate representation of the underlying data, the result of the subsequent network analysis will have no significance in the system. As result, increasingly rich data such as multilayer interactions and multistep pathways also gives rise to new and unique challenges associated with the conventional network approach.
Connecting structure and dynamics

For most systems, the network structure is the underlying topology where dynamical processes can take place. The network of roads, for instance, defines the possible pathways for vehicles to move around. Another example is the global computer network of Internet, which defines the possible routes for information packages. For these networks, the connection between the network’s structure and dynamics is apparent.

In a first-order Markov approach, with a decoupling of the network structure from its dynamics, there is a direct correspondence between the state space of the Markov process and the dynamics on the network structure [12]. The approach therefore allows for exploring the interplay between structure and dynamics from two perspectives: either we can explore how the network structure influences the dynamical process, or we can use the dynamical process to explore the network structure. Whatever perspective is taken, the ultimate goal is to better understand the connection between the structure and the dynamics in the system. An example of applying the different perspectives is the examination of the modular organization of a network is to be examined. Real-world networks tend to have a modular organization with communities that often corresponds to the functional and behavioral units of a system [32]. To find those units, community-detection methods are used.

Community detection is a powerful approach to uncovering important structures in large real-world networks\(^\text{1}\). There are different methods of community detection that take either the perspective of analyzing the network structure, as, for example, modularity [106], or the perspective of the dynamical process, as, for example, Infomap [107]. Along with, for example, ranking, spreading, and traversing community detection constitutes some of the high-level analyses used to understand the connection between structure and dynamics in networks.

\(^1\) An example of a community-detection problem is identifying clusters of customers with similar interests in order to provide better product guides. This can be done by analyzing networks of purchase relationships between customers and products offered by online retailers [105].
Shoehorning data into conventional network models

New technology has enabled us to collect data on almost any complex system. An important question is how to accurately represent the data as networks. This important representation step — which determines the quality of subsequent analysis — is often overlooked [108].

To be useful, a network model must be a simplified version of the pattern of interactions in the complex system it represents. At the same time, it is critical for the network to truly represent the inherent phenomena in the complex system in order to avoid incorrect analysis results or conclusions. This duality poses challenges in choosing an appropriate approach, and many times interaction data about a complex system is shoehorned into an unweighted and undirected network to enable analysis with conventional network methods. This approach limits what regularities can be detected, and the process might lead to distortion, addition and deletion of edges. As a result, the observed relationships might not be totally equivalent to the relationship that is of interest [64].

Independent of the data, a common practice when creating networks is to create a one-to-one mapping from entities in the data to nodes in the network, then count the number of interactions between entity pairs in the data and take that as the edge weights in the network. This approach corresponds to creating a simple network. Simple networks capture the pairwise relationship in the data, and can be extended to become weighted and directed networks. Since the minimalistic simple network is trivial to build and has many related analysis methods, it is perhaps the most frequently used network representation. However, while we can achieve diverse network representations of the same data, not all representations contain the same information [109].

What model to choose when creating networks depends on
how well the data from the complex system can be reproduced, given the network representation, and if the representation itself is sufficient for further analysis. A failure to represent the dependencies in the original data in the network will lead to inaccurate results when applying a wide range of network analysis tools that are based on the simulation of movements on the network, such as community detection, ranking and information spreading [11]. These methods are rooted within the conventional framework, and when the assumptions used in this framework do not serve as reasonable approximations of the system of interest, alternative representations and techniques may be necessary. We therefore need to consider what factors should be considered when choosing a network representation, and what are the consequences when this choice is poorly made [85].

**Limitations of first-order models**

Many ranking and community detection methods, as well as epidemic models, build directly on the first-order Markov process. While a first-order model is sufficient to capture flow dynamics in some systems, recent studies have shown that higher-order flows are required to capture meaningful dynamics in many complex systems [84, 110, 111]. Importantly, first-order models lack memory in their dynamics, an assumption that is often not realistic in practice [12].

Take an air traffic network as an example, see Figure 3. In the conventional approach without memory, nodes represent airports and links represent flight legs, and random walkers move on the links between the nodes to represent passenger flow. This dynamical process corresponds to a first-order Markov model of network flows: a passenger arriving in an airport will randomly continue to an airport proportional to the air traffic volume to that airport. That means, for example, that two passengers who arrive in Chicago, one from San Francisco and one from New York, will have the same probability, 44 percent, of flying to New
York next. In reality, however, passengers are more likely to return to where they come from \[11\]. In a second-order model with memory of the previously visited airport, 91 percent of all passengers from New York actually return to New York after arriving in Chicago. As a result, describing network flows with a first-order Markov model suffers from memory loss and washes out significant dynamical patterns \[110–112\]. Similarly, aggregating flow pathways from multiple sources, such as different airlines or seasons in the air traffic example, into a single network can distort both the topology of the network and the dynamics on the network \[9, 113–115\]. The conventional first-order approach of modeling dynamics on networks therefore oversimplifies the real dynamics and sets a limit of what can actually be detected in the system.

Figure 3: Memory effects in networks of air traffic. Passenger flows from left to right, to and from Chicago (a) The conventional network approach without memory washes out dependencies and flow mixes between destinations; (b) In the higher-order approach, memory information of real travel pathways is captured in the model. Where people go depends on where they come from. People tend to return to the city they came from.
Chapter III

Higher-order network modeling
Higher-order models

Conventional network models implicitly assume a single type of link and that the dynamics simulated on the network are first-order. The conventional model therefore ignores the multifaceted nature of interactions and that movements between nodes may not only depend on the current node, but also on previous nodes.

For some complex data, this approach is sufficient, but for data with higher-order dependencies, the first-order assumption can quickly become limiting. Failure to represent such higher-order information in the network can lead to inaccurate or even incorrect results when applying network analysis methods. To fill the gap between the network model and the data, we can account for higher-order information using higher-order models.
The meaning of higher-order models

In higher-order network models, the movements between nodes in the network depend on more information than just the previously visited node. Such higher-order dependencies exist in many systems and are crucial to understanding these systems [11]. Some examples include web browsing [116, 117], where the probability of the user following a specific hyperlink depends on the previously visited web pages, and conversations [118], where succeeding communication depends on previous conversation partners. Models of higher-order data dependencies are therefore often required to capture critical phenomena in the dynamics and function of complex systems [10, 84, 110–112, 119], see Figure 4.

Higher-order dependencies in data have been studied in various fields [120–122], but the focus of these studies has been on the stochastic process rather than on how to represent higher-order data from complex systems in networks. To address this
issue, recent research has argued for higher-order network models that capture both temporal and topological characteristics of sequential data [11, 12, 109, 123, 124]. Two examples of such higher-order network models are memory [11] and multilayer [125] networks. In memory networks, higher-order data dependencies can represent multistep pathways such as flight itineraries [11, 108, 109, 126], and in multilayer networks, they can represent temporal or multi-source data such as multiseason air traffic [9, 115, 127, 128]. Whereas memory networks can capture the next movement depending on where they come from, multilayer networks can capture the next movement depending on the current layer.

**Memory and multilayer networks**

In memory networks, the movement on the network depends on multiple previously visited nodes. Specifically, for a random walker that steps between physical nodes and generates a sequence of random variables \(X_1, \ldots, X_t\), the transition probabilities for a higher-order model of order \(m\),

\[
P(X_t|X_{t-1}, X_{t-2}, \ldots) = P(X_t|X_{t-1}, \ldots, X_{t-m}),
\]

depending on the \(m\) previously visited physical nodes.

In multilayer networks with the same physical nodes possibly present in multiple layers, the movement depends on both the previously visited node \(i\) and layer \(\alpha \in \{1, 2, \ldots, L\}\). In many cases, empirical interaction data only contain links within layers. Without data with links between layers, it is useful to couple layers by allowing a random walker to move between layers at a relax rate \(r\). With probability \(1 - r\), a random walker follows a link of the node in the currently visited layer, and with probability \(r\) the random walker relaxes the layer constraint and follows any link of the node in any layer. In both cases, the random walker follows a link proportional to its weight.
With total outlink weight \( w_{\alpha i} = \sum_{\beta j} w_{\alpha i\beta j} \) from physical node \( i \) in layer \( \alpha \), and total outlink weight \( w_{i} = \sum_{\alpha,\beta} w_{\alpha i\beta j} \) from node \( i \) across all layers, which both correspond to total intralayer outlink weights as long as there are only empirical links within layers, the transition probabilities for multilayer networks with modeled interlayer transitions are

\[
P (\alpha_i \rightarrow \beta_j) (r) = P_{\alpha_i\beta_j} (r) = (1 - r) \delta_{\alpha\beta} \frac{w_{\alpha i\beta j}}{w_{\alpha i}} + r \frac{w_{i\beta j}}{w_{i}}, \quad (5)
\]

where \( \delta_{\alpha\beta} \) is the Kronecker delta, which is 1 if \( \alpha = \beta \) and 0 otherwise. In this way, the transition probabilities capture completely separated layers for relax rate \( r = 0 \) and completely aggregated layers for relax rate \( r = 1 \). The system under study and the problem at hand should determine the relax rate, but for relax rates around 0.25, results are robust in many real multilayer networks [115].

**Fixed-order models**

Addressing higher-order dependencies still remains at the stage of assuming a fixed-order model of dependency when creating the network [11, 123, 124, 129]. The fixed-order models are exclusively of second order, and although there have been efforts to incorporate network models of higher order for ranking and clustering [130], the approaches need to modify existing algorithms and are application-specific. As a result, the methods are not generalizable to broader applications, although the goal is to provide a network representation that is independent of the end-analysis methods.

While fixed-order models can provide a path to modeling higher-order dependencies in networks, there are some challenges to consider. In particular, the increasing complexity of fixed higher-order models comes with three significant draw-
backs; First, the models grow exponentially with Markov order and quickly become computationally inefficient. A fixed \( m \)-th-order model can be directly converted to a first-order model \([11]\), but the state space grows exponentially with the order. Second, the exponentially larger models require exponentially more data for statistically sound fits. Third, there are, in practice, too few models for good fits to the data; an \( m \)-th-order model may underfit and an \((m + 1)\)-th-order model may overfit the data. There has been research on using Markov order estimation to determine the optimal order in higher-order data, such as using different information criteria \([131, 132]\), cross-validation \([116]\), and surrogate data \([133]\), but these approaches produce a single global order for the model rather than variable orders, and no discussion was given to network modeling.

The drawbacks related to fixed higher-order models will inevitably transfer to the analysis methods that build on them. So while first-order models can lead to incorrect results, an appropriate fixed higher-order model requires an approach with an effective model selection that can control for the increasing complexity in order to create more accurate network models that also allow for efficient analysis.

The effects of a higher-order model for journal data

In paper III, *Robustness of Journal Rankings by Network Flows with Different Amounts of Memory*, we examine the impact of using different amounts of memory in the network model, and show why we need to go beyond first-order models when there is information that allows us to do so. The paper examines the ranking of scientific journals, and the analysis focuses on the application area of bibliometrics and the question: How robust are different rankings to the selection of included journals?

As the number of scientific journals has multiplied, journal rankings have become increasingly important for scientific decisions.
From submissions and subscriptions to grants and hirings, researchers, policy makers and funding agencies make important decisions influenced by journal rankings such as the ISI journal impact factor. Therefore, researchers put great effort into finding the best work by other researchers and into achieving maximum visibility of their own work by publishing in prominent journals. But the importance of journals may change over time, because their importance is ultimately set by the quality of the work they publish. As a consequence, it is in everyone’s interest to continually assess the importance and rankings of journals.

Typically, rankings are derived from the citation network between a selection of journals and unavoidably depend on this selection. However, little is known about how robust rankings are to the selection of included journals. In the paper, we compare the robustness of three journal rankings based on network flows induced on citation networks. They model pathways of researchers navigating scholarly literature, stepping between journals and remembering their previous steps to different degrees: zero-step memory as impact factor, one-step memory as Eigenfactor, and two-step memory, corresponding to zero-, first-, and second-order models of citation flow between journals.

We conclude that higher-order models perform better and are more robust to the selection of journals. Whereas our analysis indicates that higher-order models perform better, the performance gain for the higher-order network model comes at the cost of requiring more data extending over a longer time period.
Variable-order models

More compact models of the data for efficient analysis

More data is often better, but not always. When adding more data the curse of dimensionality strikes, and sometimes more data is not even needed to provide proper results. In the cases when more data with higher-order information is needed to describe a system, fixed higher-order representations can provide useful insights. But not all systems are of a fixed-order. This problem can be addressed with a very simple idea: the order is allowed to be of variable length in variable-order models.

Variable-order models for representing data

Variable-order models allow the movements between nodes in the network to depend on information that vary depending on the state of the movement [134–136]. The variable-order model was originally introduced in the field of information theory for data compression purposes [137], and since then it has been
implemented in various research areas such as bioinformatics, financial series [138, 139] and human navigation [117].

Variable-order models combine states with similar transition probabilities into single states using so called *state lumping*. The lumped states replaces the original states and assembles their links such that the transition probabilities remain the same. This gives a more compact description of the data and make it possible to fit models in between fixed-order models. These properties make variable-order models highly efficient for compressing and predicting higher-order data such as sequential data, and they have successfully been applied in bioinformatics and natural language modeling [140].

Different classes of variable-order models use variable-length Markov chains to impose different constraints on the state lumping, and the unconstrained and therefore most general class is called sparse Markov chains [141]. There is a suite of inference methods for different classes of variable-length Markov chains, based on Kullback-Leibler distance [142], cross validation [109] and information-theoretic or Bayesian model selection [140, 141]. However, with few exceptions [109], they are, in the network terminology, developed for small sets of nodes, such as the four letters of DNA or the 128 characters of ASCII text.

Variable-order models are more compact in size than fixed-order models since they allow for variable representations of the dependencies in the original data. As an example, instead of representing the data with a fixed high order, they can use variable orders by adding supplementary nodes and links to a first-order network only where necessary. For example, a variable-order model of maximum second-order will have fewer nodes and about half the number of edges compared with a fixed second-order network, because it uses the first order whenever possible and embeds second-order dependencies only when necessary.
Sparse memory networks as higher-order networks

Variable-order models are designed for sequences with the small alphabets of DNA or ASCII text and not for networks with thousands of nodes. To address this problem, researchers have recently introduced so-called higher-order networks [109]. However, these higher-order networks are general-purpose representations for existing network methods. This leaves open the question of how to best integrate variable-order approaches for efficiently modeling higher-order networks.

The challenge with variable-order models is to find both a good general way of modeling the data and a way of estimating the minimal state space to use. The first task is cumbersome, as direct application of, for example, community-detection algorithms for fixed-order models confuse relevant and irrelevant states [12, 109]. The second task can be seen as a model selection problem. However, due to the often large number of potential variable-order submodels of a fixed high-order model, it is a challenge to find a good global model selection technique.

To overcome the memory-associated limitations of conventional network models, one approach is to represent higher-order data with two set of nodes: one set for the physical components, called physical nodes, and one set for the flows between the components, called state nodes, see Figure 5. Thanks to this modeling, it is possible to discriminate between physical nodes for describing the concrete objects that flow can visit, and the state nodes for representing the dynamics. This makes it possible to use flow nodes to include memory of any order in the network models. Various higher-order network flow representations, including memory and multilayer networks, can be represented with sparse memory networks.

Compared to fixed-order models, sparse memory networks have several advantages. At the same time, they overcome the problems related to variable-order models. First of all, sparse
memory networks are often more reflective of the underlying real-world system, which seldom has a fixed order. Thanks to this, a variable-order model provides a more accurate representation of the data. Second, a variable-order model scales more efficiently to higher orders, because auxiliary higher-order nodes and edges are added to a first-order network only where necessary. This makes the variable-order models more efficient for subsequent analysis. Third, the sparse memory network model is consistent with the conventional network representation. This allows for existing network analysis methods and algorithms to be modified to run on the same model.

Figure 5: Sparse memory network representation. (a) Example of network with three flow pathways. The first pathway goes from the blue physical node to the orange physical node, ending in the red physical node. The second flow pathway starts in the green physical node, goes to the orange physical node, and back to the green node. The third pathway goes from the white physical node to the orange physical node, ending in the red physical node; (b) The network represented with two sets of nodes: one set for the physical nodes, and one set of state nodes for the flows. In this way the same flow pathways with memory are represented with a single link between two state nodes. For example, the first pathway corresponds to a step between the blue state node in the orange physical node, and the orange state node in the red physical node; (c) The lumped network where the redundant blue and white state nodes have been merged.
Analysis of real-world systems using sparse memory networks

In papers IV and V, *Mapping Higher-Order Network Flows in Memory and Multilayer Networks with Infomap* and *Maps of Sparse Markov Chains Efficiently Reveal Community Structure in Network Flows with Memory*, we look at how to reveal communities in higher-order network flows with a general approach. We present a new framework, an extension of the Infomap methodology, that is able to capture higher-order dependencies. The aim is to shed light on two main questions: (1) How can we develop more accurate representations using higher-order network models? and (2) How can we provide methods for simplifying and highlighting important structures in higher-order network models?

Mapping of higher-order networks

In paper IV, *Mapping Higher-Order Network Flows in Memory and Multilayer Networks with Infomap*, we show that various higher-order network flow representations, including memory and multilayer networks, can be represented with sparse memory networks. The modeling of higher-order data with sparse memory networks can be seen in Figure 6.

In the paper, we demonstrate how the mapping approach for sparse memory networks can be used for modeling and analysis of higher-order network representations. With the broad spectrum of higher-order network representations required to best describe diverse complex systems, the apparent need for specialized community-detection methods raises the question that we explore in the paper: How can we reveal communities in higher-order network flows with a general approach?

The paper also derives the map equation applied to sparse memory networks and describes its search algorithm, Infomap, which can exploit the flexibility of sparse memory networks.
Figure 6: Modeling higher-order network flows with sparse memory networks. (a) Multistep pathway data from two sources illustrated on a network with five physical nodes; (b) The pathway data modeled with a second-order Markov model on a memory network, where state nodes capture where flows come from; (c) The memory network represented as a multilayer network where layers, one for each physical node, capture where flows come from; (d) The pathway data modeled on a two-layer network, one layer for each data source; and (e) Both memory and multilayer networks mapped on a sparse memory network with no redundant state nodes. The black link highlights the same step in all representations.
In summary, the paper presents an approach that uses an efficient variable-order network model for representing data and a method for efficiently analyzing the higher-order network models to simplify and highlight important structures.

**A real-world system analyzed with sparse variable-order networks**

Paper V, *Maps of Sparse Markov Chains Efficiently Reveal Community Structure in Network Flows with Memory*, builds on the framework presented in paper IV. This work presents the most ambitious project of this thesis, and although a paper was written already in late 2015, we decided to take the analysis further to make the project even more significant.

In the paper, we show that describing higher-order network flows with so-called sparse memory networks and identifying modules with long flow persistence times, such as groups of airports that contain frequently traveled routes, provides a general solution that reveals modular patterns in higher-order flows through complex systems. The paper uses the sparse memory network approach and applies it to large-scale real-world citation data.

Citation data and the flows of citations have become increasingly important in science, as the citations form the basis for ranking and classification of scientific literature [143]. Both researchers and research councils rely on good classifications of scientific journals for efficient search and analysis of the ever-growing scientific literature. Today, the most commonly used journal classification scheme Web of Science, provided by Thomson Reuters, uses a heuristic classification that is based on manually developed categories of journals. Although this classification may be useful, it has problems due to its peculiar design, and the ad hoc division of journals has been widely criticized. To find a classification of scientific journals that overcomes the shortcomings and better captures the actual flow of citations, we use the approach
with sparse memory networks. This approach provides several advantages. First, it allow for journals to be classified in more than one category. The result is that multidisciplinary journals, which are not bound to a specific scientific field, can be classified under their different disciplines. Second, unlike heuristic classifications, it provides a transparent method for dividing journals. Third, the approach is both automated and computationally efficient.

To illustrate the results, Figure 7 represents citation flows between research fields as a visual map. Compared with currently used journal classifications, the research fields form better units of analysis because the map more effectively captures how ideas flow through science. This demonstrates that the use of higher-order models with more data can benefit the understanding of real-world systems. The sparse memory network approach effi-
ciently simplifies and highlights important structures in higher-order data from a real-world system. The paper thereby show that the higher-order approach can capture richer data and provide a more accurate representation of a complex system.
Chapter IV

Conclusions
Conclusions

To better understand complex systems, we need to understand the networks behind them. In this thesis, we have shown that we can create more accurate network representations if we move toward higher-order network models. Simplifying and highlighting important interaction patterns in higher-order network models allows for more effective analysis of many real-world systems.

First, we explored the conventional network approach to analyze data from real-world complex systems. For example, we created a financial network from data of shareholdings to examine trading behavior, and we created an information network from data on Wikipedia edits to examine information interests on a global scale. We found that conventional network modeling could provide valuable insights about the function of the underlying complex systems.

We then expanded the analysis for higher-order models of networks to explore the impact of using richer data in the network representation. Using data on bibliographic citations, we analyzed journal relevance using both conventional and higher-order models. This analysis demonstrated the impact of using richer data in the network representation and showed why we need to go beyond conventional models when there is data that allows us to do so. In addition, we also presented a new frame-
work for higher-order network modeling and analysis that can more efficiently capture structural information in networks with richer data. Within that framework, we illustrated how simplifying and highlighting important patterns in higher-order network models allows for more effective analysis of complex systems. Specifically, we used a variable-order approach to create network maps that efficiently revealed community structure in a higher-order citation network.

While networks are at the heart of some of today’s most revolutionary technologies, empowering everything from search engines to social platforms, network science is still a relatively young discipline, and although much has been done in network science, there is still a lot to be developed and explored. Efficient higher-order network models and increasingly rich data will accelerate the development and open doors to unexplored territories in different disciplines of science. Questions remain regarding how to generalize various network analysis methods to higher-order models, as, for example, visualizations, to better understand and highlight important patterns in complex systems.

Finally, I would like to thank you for reading this far, and encourage you to contact me if you want to discuss something in the thesis or just anything under the sun.

Respectfully submitted,
Ludvig Bohlin
Contributions

Short description of the contributions by the author to each of the papers included in the thesis

Paper I  The author helped design and conceive the experiments. He managed the dataset collection, performed the numerical simulations, and compiled the results. The author also actively participated in the writing of the paper.

Paper II  The author was involved in the development and testing of the method and the interpreting of the results. He also actively participated in the writing of the paper.

Paper III  The author managed the project and planned and carried out the analysis and simulations. He managed the dataset collection and tested and compared the results. The author also actively participated in the writing of the paper.

Paper IV  The author helped plan the project and was involved in the design of the experiments. He helped analyze the data and actively participated in the writing of the paper.

Paper V  The author helped design and conceive the experiments. He managed the dataset collection, performed the numerical simulations and compiled the results.
The author also actively participated in the writing of the paper.
Words of thanks

I owe a lot to my parents, especially my mother and my father — Greg Norman

Misleading. That is probably the correct word to use, considering the fact that only my name is printed on the thesis. In reality, several persons have been involved in the progress of shaping it.

First of all, I would like to express my debt to my personal heroes and mentors from whom I have learned so much: mom, dad and grandma Estrid – thank you! The thesis would not have been possible without the guidance from another hero: Martin Rosvall. Your ideas, mentality, mentoring, and support helped me not only to take the steps necessary to keep my work on track, but also to develop as a person. For that I will always be grateful.

I would also like to thank all the wonderful people I met over the years at IceLab and the Physics department, of which there are too many to provide a complete list. I especially thank Ludvig Lizana, Markus Nyberg, Daniel Edler, Andrea Lancichinetti, Magnus Lindh, Xavier Thibert-Plante, Peter Fransson, Alcides Viamontes Esquivel, Atieh Mirshahvalad, Alexis Rojas-Briceno, Joaquin Calatayud, Chris Blöcker, Magnus Neuman, Fariba
Karimi, Christian Persson and Jonas Wickman for fun times, exciting table hockey games and interesting conversations. The same goes for the new generation of people in IceLab whom I will sadly leave. However, I believe strongly that in the future I will join IceLab for a coffee from time to time.

I am thankful for all the help and support from my colleagues in the Physics department, especially Katarina Hassler, Ann-Charlott Dahlberg, Lena Burström and Jörgen Eriksson. I would also like to give a special thanks to my partners in crime, Petter Lundberg and Mattias Lindh.

Many readers, too numerous to name here, have given feedback on the thesis, and to them all I extend my sincere thanks and acknowledgments.

My work has depended on coffee and the generosity of free software authors. My computer has run editors, programming languages, scripts, etc, provided for free by several generous developers. Thanks, everyone!

My Friday lunch partner, alias Terry Bull, deserves a special thank you.

And of course, I would like to thank my family for their love and their ability to get me to think about what is really important in life. Last but not least, I thank my best friend and partner, Emelie, for your encouragement, inputs, and endless support during good times and bad.
Bibliography


[92] Renaud Lambiotte, Roberta Sinatra, J-C Delvenne, Tim S Evans, Mauricio Barahona, and Vito Latora. Flow graphs:


[115] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall. Identifying modular flows on multi-


