Lecture Notes in Social Networks

Katharina A. Zweig

Network Analysis Literacy

A Practical Approach to the Analysis of Networks

Springer

Lecture Notes in Social Networks

More information about this series at http://www.springer.com/series/8768

Katharina A. Zweig

Network Analysis Literacy

A Practical Approach to the Analysis of Networks

Springer

Katharina A. Zweig
TU Kaiserslautern
FB Computer Science
Graph Theory and Analysis of Complex
Networks
Kaiserslautern
Germany

ISSN 2190-5428 ISSN 2190-5436 (electronic)
Lecture Notes in Social Networks
ISBN 978-3-7091-0740-9 ISBN 978-3-7091-0741-6 (eBook)
DOI 10.1007/978-3-7091-0741-6

Library of Congress Control Number: 2016948283

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Austria
The registered company address is: Prinz-Eugen-Strasse 8-10, 1040 Wien, Austria

To Winfried Zweig
Thanks for all the good discussions
on network analysis.

Foreword

This is a delightful book. It 'is so easy to read, and you can almost accidentally learn quite a bit of network science without even noticing it. Written in a playful manner, it tends to enliven the brain rather than put it to sleep—quite a change from the usual pedantic tone. It's a quirky book that does not try to be systematic. For example, it does not cover "community detection" (that's cluster analysis to you social scientists). As a result, the book has a great deal of personality.

But what I really like about the book is the subtext. What it's actually about, in my opinion, is how to think, and here, that means how to think with models. Most academics are very gullible when it comes to concepts outside their disciplines. Within their area, any new idea or phrasing is treated with withering skepticism, but outside their area, they adopt ideas with the speed of teenagers adopting slang or fashion. Thus, a management scholar hears about small worlds and clustering coefficients and immediately shoehorns them into their next study. A physicist learns about betweenness centrality, and suddenly, there are 500 papers that reference the idea. If the first paper associates betweenness with influential spreaders in the spread of a disease, all of the following papers do the same. If you internalize this book, you won't make that mistake. You will realize that although there is a sense in which network measures are tools like hammers, there is much more to them. Hammers work pretty much the way they work in any setting, but using a network measure implicitly entails fitting a model of how things work. And if the model doesn't fit, the measure doesn't either.

Curiously, although I associate model-based thinking with the physical sciences, my experience is that both physical and social scientists are equally likely to have this mindless, "pluginski" attitude about network concepts. Therefore, I think this book would be useful for both audiences. But since the content of the book is mostly drawn from what Katharina calls the "network science" field (as opposed to

the “social network analysis” field), I’m guessing it will appeal mostly to budding physical scientists. Too bad, because if there was ever an introduction to network science that was especially suitable for social scientists, this is it.

I look forward to seeing this in print.

Steve Borgatti
Lexington KY, USA

Preface

> *On the other hand, this rapid expansion [of complex network science] creates the risk that existing methods may be misapplied or misinterpreted, leading to inappropriate conclusions and generally poor results*. (Carter Butts: "Revisiting the Foundations of Network Analysis," Science, 325, 414–416, 2009)

After finishing a study of biochemistry and in the middle of a bioinformatics study, I started my work as a doctoral student in the field of "Algorithm design and computational complexity" in 2003. I was immediately attracted to a budding new field, *complex network science*, which had just taken off some years ago.

I was lucky enough to meet Ulrik Brandes early in this endeavor, and he invited me to participate in the now classic textbook edited by him and Thomas Erlebach: "Network analysis—Methodological Foundations." With other doctoral students, I was assigned to the chapters on *centrality indices*. In the beginning, I was overwhelmed by the dozens of different indices that had been proposed so far and the seemingly never-ending flow of newly proposed centrality indices. The argumentation almost always went along the following lines: "So far, these indices have been proposed. In the new data set X, none of these measures matches with the intuition. Thus, we propose the new measure Y that matches our intuition of centrality in this network." I was lost which index to take in any specific situation.

Finding a first online version of Stephen P. Borgatti's paper on "Centrality and network flow" was a revelation: Borgatti basically says that a centrality index is a predictor of which node is used most heavily in a given network flow or network process. While others like Freeman had also hinted at a relation between processes on a network and a measure to quantify the network's structure, Borgatti was the first to make a tight connection between a process of interest and the measure to quantify the indirect effects induced by this network process. He also stated quite clearly that a mismatch between a complex network, the network process of interest, and the centrality index will lead to uninterpretable results: "the off-the-shelf formulas for centrality measures are fully applicable only for the

specific flow processes they are designed for, and (…) when they are applied to other flow processes they get the 'wrong' answer."[1]

This book is based on the idea that network processes and network analytic measures are even more intertwined, beyond the set of centrality indices. In the last ten years I have generalized this idea to all kinds of distance—and walk-based measures. The main hypothesis of this book is as follows:

> **Note 1.** To interpret the values of a distance-based measure, the way of calculating the distance must be matched to the process of interest. To interpret any walk-based measure, the set of walks used by the measure needs to be closely adapted to the process.

This includes the whole process of data observation, preprocessing, representation as a network, stating a network process of interest, and choosing a network analytic method to analyze it. It is the book that I would have loved to have at the beginning of my doctoral research.

Intended Audience

There seem to be three types of groups pursuing network analytic projects:

1. Groups consisting of scientists with a heap of data that want to analyze their data by network analytic methods—henceforth called *data experts*.
2. Groups consisting of scientists that primarily devise network analytic methods and then search for data that can be analyzed by their newly devised method—henceforth called *method experts*.
3. A quite small set of interdisciplinary groups, consisting of data **and** method experts.

As a biochemist, I was clearly in the first group as a *data expert* that was overwhelmed by the choice of methods; later, as an algorithm designer and method expert, it became clear that applying the best and most beautiful method to data and a research question it does not match with, is not helpful either.

This book will stress that people from both groups need to be *literate* in the other group's regime: If a data expert creates a beautiful data set which can be represented and analyzed as a network, it is important not to miss any vital pattern just because a particularly suitable method is not known to him or her. Similarly, for any method expert, it is vital to understand the data to which a chosen method is applied. In particular, it is not enough to just reference to the data expert's publication and to roughly know what the vertices and edges represent, but it is necessary to understand in detail how the data were produced, to know the odds of observing false-positive and false-negative relationships, and to know whether the resulting network is complete or not. However, many data experts do not include this

[1] Stephen P. Borgatti: "Centrality and Network Flow", Social Networks 27, 55–71, 2005.

information in their publications, for example, because the community from which the data originate is well aware of the applied procedures.

This book tries to build the bridge between the two groups and to show the different perspectives they have on their subjects and projects.

The Ideal Reader

Yes, I have some requirements toward you, my dear reader. Obviously, you are a data expert who thinks that network analysis could be helpful to reveal the most exciting mysteries in your field—and so do I. With this book, I will equip you with the necessary questions you need to ask your method expert to understand whether your research question matches with his or her method.

Or you are a method expert, maybe a mathematician or a computer scientist, and your advisor just gave you this piece of data and asked you to design a method to analyze it—then, this is the book that will help you to understand which questions to ask your data provider.

It is just the book I wanted to have when I was about one year into my doctoral studies, still overwhelmed by the amazing flexibility of network analysis and underwhelmed by the number of good guidelines to use it: guidelines on how to actually represent a complex system by a network or how to choose the best method to analyze it, and how all of this is connected to my research question. I was baffled by the daring approach of physicists that simplified complex systems beyond recognition to a set of nodes and edges—and at the same time, I was intrigued by the potential of this new approach. However, I was still biochemist enough to wonder whether there is actually a line where simplification needs to stop, in order to find contextually meaningful results. If you are at this point in your career, I wrote this book for you.

If you are not quite there yet, you might want to read the very good collection of papers, edited by Mark Newman, Albert-László Barabási, and Duncan J. Watts, called "The structure and dynamics of networks."[2] For those in a hurry, the following papers are the minimally required prerequisites to get a feeling for the field:

1. Start with the two papers that opened the field of (social) network analysis to a much broader community and transformed it into complex network analysis: The first was published 1998 by Duncan Watts and Steven H. Strogatz under the title "Collective dynamics of small-world networks," Nature 393, pp. 440–442. The paper introduced the small-world model. The second influential paper was published in 1999 by Albert-László Barabási and Réka Albert under the title "Emergence of scaling in random networks" in Science 286, pp. 509–512. It introduced the notion of scale-free networks and a model to produce them, the preferential attachment model. Both papers are shortly summarized in Chap. 6.

[2]Published by Princeton University Press, Princeton and Oxford, 2006

2. The first paper by Barabási was quickly followed by a disturbing one which showed that scale-free networks built with the preferential attachment model are robust against random failures of nodes but very sensitive to attacks on their most connected nodes: Albert, Jeong, and Barabási published these findings in 2000 under the title "Error and attack tolerance of complex networks" in Nature 406, pp. 378–382.
3. For the course of this book, the work on so-called *network motifs* by Uri Alon's group is very much important.[3] Other disciplines had already started earlier to explicitly compare a structural value found in a network with the expected one, for example ecology.[4] For the field of complex network analysis, the articles by Alon et al. were the first widely visible ones that proposed to assign a significance value to observed results by comparing the observation with the expectation.
4. The articles above are written by physicists. Read now the view of the sociologists, as stated by Borgatti et al.'s paper on "Network Analysis in the Social Sciences," published in Science, 323, pp. 892–895, in 2009.

By reading these papers, you might notice that the publications in the field of complex network analysis come from very different publication venues. This is caused by the very interdisciplinary origin of the field. For example, for computer scientists, it is common to publish their original research in a conference proceeding, and some of their conferences are as reliable and respected as journals. For physicists, a conference is a place to meet and exchange ideas, but most often, they report recent work that was already published elsewhere. Please read the chapter in the appendix discussing different publication styles, where to find which information and how to differentiate peer-reviewed from unreviewed publications.

Now, you are well-prepared for an instruction on how to read this book!

How to Read This Book

As a reader, I mostly skip these sections on "How to read this book," so I make it extra-short: This is a book to be read from left to right and from top to bottom, or to dip in as you please. The exercises are intended to deepen the understanding of the methods introduced in the text. Moreover, they teach what questions to ask whenever you make acquaintance with a new measure. Almost all exercises can be solved on two levels: a verbal, explanatory solution with the help of an example and

[3] I suggest to read: Ron Milo et al.: "Network Motifs: Simple Building Blocks of Complex Networks," Science 298, pp. 824–827, 2002 and Ron Milo et al.: "Superfamilies of Evolved and Designed Networks," Science 303, pp. 1538–1542, 2004.

[4] Nicholas J. Gotelli and Gary R. Graves: "Null-Models in Ecology," Smithsonian Institution Press, 1996.

by proof. For courses with mathematicians, physicists, and computer scientists, I normally require a proof for these exercises.

The book is divided in three parts: Part I gives you an overview of the field and names the necessary definitions. Part II is devoted to the most important methods, starting with some classic, network analytic measures, a basic discussion on how to represent data as complex networks, various random graph models and their use in network analysis, and centrality indices. Most importantly, it also contains a chapter on how to analyze a measure that you encounter somewhere. Both parts are just the preparation for the core of this book, Part III, which describes various aspects of *network analysis literacy*: when data cannot be represented as a network, when a method's results are difficult to interpret, and finally, why network analysis is a field that even sometimes requires an ethical perspective.

So, where would I recommend you to start? When you are an absolute beginner in network analysis, start with—surprise—Chap. 1. If you are already confused by network analysis, because there are so many different approaches to it, start with Chap. 2. Both groups only need to skim the definitions—just come back later whenever you need them (Chap. 3). If you are an intermediate network analyst, i.e., you have conducted at least 3 network analytic projects, start with Chap. 8 and then read Chaps. 5–7. When you are an expert, just read the literacy chapters, starting from Chap. 10.

You will find that in this book, I often switch between the male and the female pronoun as long as I refer to some group of people in general ("the user → she" or "the user → he"). You find this annoying? Well, so do I! But as long as you and I notice it and still find it surprising or annoying or pleasing or anything else but normal, I feel the need to stick to it. Of course, the pronoun 'she' refers to both, male and female, persons. ☺

And thanks go to…

I would like to thank my colleagues Ulrik Brandes, Johannes Glückler, Kai Fischbach, and Alexander Mehler for our long discussions on network analysis. I would also like to thank the countless reviewers and foremost my own students Emöke-Ágnes Horvát, Wolfgang Schlauch, Mohammed Abufouda, and Sude Tavassoli for their influence on my work and the successful collaboration; last but not least, I would like to thank my collaborators from biology, especially Kevin Bähner and Thorsten Stoeck.

I hope that the book will foster a discussion on a more principled way of when to use which network analytic methods. The set of guidelines enabling this choice is what allows *network analysis*. However, this book is only the beginning of this and far from complete. Let me know your opinion, send in good and bad examples of network analysis, propose your own set of guidelines, and share all of this with me at zweig@cs.uni-kl.de. I will discuss a selection of those proposals on my blog http://netz-werker.blogspot.de/.

The book is dedicated to my husband who has shared all of my ups and downs in network analysis and supported me to write this book. Thanks for all the discussions on this topic that others not involved in network analysis might not have found as worthwhile as you.

Kaiserslautern
June 2016

Katharina A. Zweig

Contents

Part I
Introduction

What is network analysis about?

Chapter 1
A First Encounter

Abstract The first chapter of the book gives a short overview of what network analysis does and why it is considered to be a vital part of complex system science: the network analytic framework allows to represent the interaction structure of a complex system as a complex network. The network's structure can then be analyzed by the application of several structural measures. However, there are two different branches in network analysis that either use the resulting values to find so-called *universal features* of complex systems or to allow a *contextual, semantic analysis*. The latter focuses on the connection between structure and function of a network with respect to the complex system of interest and some specific research question. There is a caveat, though: while, in principle, structural measures can be applied to all kinds of networks, if one is only searching for universal features, their results are not always interpretable with respect to a predefined research question. The term "network analysis literacy" is introduced to describe the knowledge of when to apply which measure to yield an interpretable result with respect to the complex system of interest.

1.1 Introduction to Network Analysis

Networks impress by their visual and intuitive quality: everyone of us is entangled in various friendship networks and business relationships, and the prospect of understanding the seemingly complex and erratic net of our personal relationships is an exciting one. Similarly, looking at scientific data in a new way, finding simple patterns that chip away individual noise to extract the main functional groups of entities in the complex system at hand, is surely one of the most gratifying moments in every scientist's life. Network analysis seems to be one of the most promising frameworks within which these two aspects, our personal life and our academic interest, can be combined, analyzed, and maybe even be understood. This prospect and the many exciting articles in journals such as Science, Nature, and PNAS, together with the interdisciplinary applicability of network analysis to various data sets and questions, has led to a tremendous interest in the methods provided by network analysis: Fig. 1.1 shows the dramatic increase of the number of articles with the keywords "network

K.A. Zweig, *Network Analysis Literacy*, Lecture Notes in Social Networks,
DOI 10.1007/978-3-7091-0741-6_1

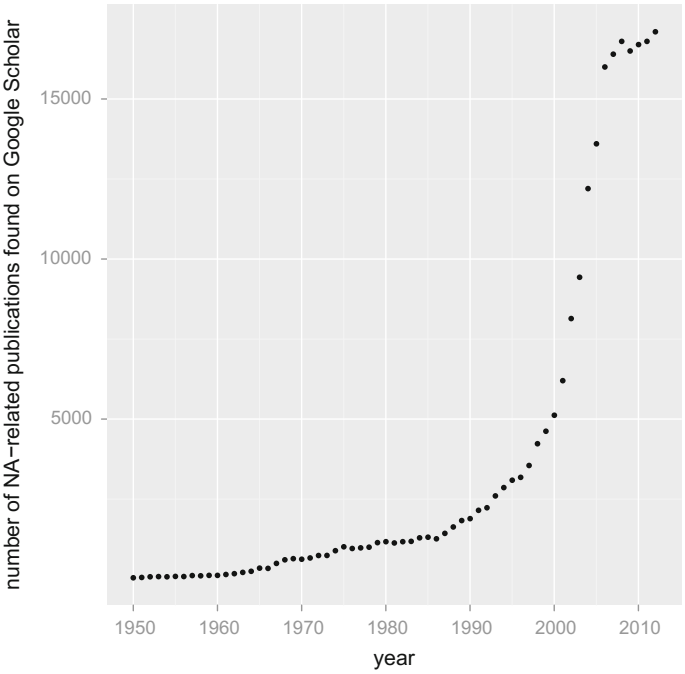

Fig. 1.1 Number of articles published in the given year containing the exact phrases "network analysis" and "complex network" as given by Google Scholar on the 12th of October, 2013

analysis" or "complex networks" as found by Google Scholar.[1] Starting from about 100 articles (as found on Google scholar) in the 1950s, the last years saw more than 15,000 articles with these terms.[2] This chapter gives a broad first encounter with network data by showing the first steps in analyzing a new set of network data.

The following chapters present a classic and widely used part of the toolkit in network analysis; but more importantly, they elaborate on the questions that are necessary to be answered in order to decide whether a given method is meaningful to the research question. One main caveat in network analysis is that once data is transformed into a graph representation, one can in principle apply any of the hundreds of network analytic methods to it—but not every method will compute meaningful and interpretable results with respect to the given data and the question

[1] For each year from 1950 to 2012, a Google Scholar search with both terms, connected by an "OR" was conducted. The number of results displayed was taken as the data point for the given year. The number of results is unlikely to hit the number of published articles in any way but gives at least an indication of the strongly increased interest in the topic.

[2] Note that double counting is as likely as an underestimation of the number of articles: articles with this topic may, for example, have been overlooked because they were published in a non-public journal which Google Scholar might not have access to. Again, the number given by Google scholar is only an indication of how many articles really have been published.

to answer. This book is thus not so much about introducing measures, many more can be found in the books by Wasserman and Faust [24], Newman [18], Borgatti et al. [6], or the book edited by Brandes and Erlenbach [7]. This book rather focuses on this last part, which I call *network analysis literacy*: it aims to empower its readers to know when to use which method so that they can quickly delve into the exciting analysis of networks themselves. Be warned that the technique by which this is accomplished follows the Socratic method which, in general, poses more questions than gives definite answers.

So, what is the first step in network analysis? One basic phase is the transformation of relational data into a complex network representation as described in the next section.

1.2 Data

The first question you might have is: what kind of data can actually be meaningfully represented as networks? A first answer is: almost any kind of data. The basic requirement is that there is a **distinct set of entities**, e.g., humans, organizations, proteins, computers, or books. The second requirement is that there is a known **relationship** between these entities. The information of whether any two entities are in the given relationship or not needs to be known for a large part of the entities since otherwise any kind of analysis will be quite shaky. Some obvious relationships between persons are: friendship, kinship, or employee-employer-relationships. Another interesting type of relationship is membership: it is a relationship between two different kinds of entities, namely persons and institutions, but that can also be represented by a network.

Relationships between non-human entities are equally abundant, and in many cases the resulting structures are also termed *networks* in our day-to-day language: examples are metabolic networks, protein-protein-interaction networks, neural networks, street networks, or computer networks. All of the above examples might be considered 'natural networks' but are there more abstract relationships that can also be represented as complex networks?

Interestingly, mathematicians have a very general understanding about what is a relation and what is not: in a mathematical sense two books can be defined to be "related" because their cover was created by the same designer. This "relatedness" does not mean that they are necessarily related in any colloquial sense: their content can of course be very different! Nonetheless, in a mathematical sense, the relation is meaningfully defined and can be easily checked by an external observer. In mathematics, a *relation* is simply defined as a subset of pairs of entities: $R \subseteq O \times O$, where $O \times O$ denotes the set of all possible pairs from some set of entities or objects O.

> **Note 2.** Mathematically, a *relation* R on a given set of entities or objects is just an arbitrary choice of pairs of these entities (objects), denoted by $R \subseteq O \times O$. In principle, any relation can be represented as a graph.

This is on the one hand much more general than the day-to-day notion of a relationship but on the other hand much less intuitive: a *relation* does not need to stand in any correlation to a real-world *relationship*; it can even represent a relationship that would not be seen as meaningful in the real world: for example, all humans with the same first name can be represented by a relation or all humans which share the same last digit of their ID-card number. Mathematical relations can also (meaningfully) be derived from other relations: One can build a second network based on the connection structure of another network by, for example, connecting two persons with each other if they share at least 8 friends in a friendship network. In this second network, there might be two persons that are connected because they share enough friends but which are not befriended themselves, and vice versa.

So, in this book, a *relationship* is something that can be observed in the real world, a *relation* is the mathematical structure which possibly represents a relationship. However, not all relations are associated with any relationship and the same relation, i.e., the same subset of pairs of a given set of elements, can represent different relationships. How is now a relationship turned into a complex network? This is discussed in the following.

1.2.1 From Relationship to Graph

In the moment a set of entities and a relationship of interest has been defined, there is a range of decisions to be made, to turn the concept of that relationship into a procedure that decides, for each pair of entities, whether they are in the associated mathematical relation or not. In most cases, when data is turned into a network representation, several decisions have to be made: if the relationship of interest contains a direction, is it necessary to include this information in the mathematical relation? Are there different levels of intensity of a given relationship and is it necessary to differentiate between them, by assigning weights to the pairs in the relation? Each of these decisions changes the set of available structural measures and the interpretation of the measure if it is applied to the network. Chapter 5 will explain in detail how data can be turned into networks. In any case, mathematically, a graph is the combination of a set of elements and a relation defined on these elements.

> **Note 3.** What is the difference between a (complex) *network* and a *graph*? The quick answer is that a graph is the *abstract representation* of a relation between entities while a network combines the graph with additional information about the entities and their **relationship** represented by the graph.

In most cases, a *complex network* represents only one set of entities (sometimes two) and **one relationship** between the entities, with some limited options on the attributes assigned to the (mathematical) relation and usually no attributes assigned to the elements. On the graph level, the elements are called *nodes* or *vertices*, and the pairs of nodes contained in the relation are called *edges*. The graph can indicate whether the relationship is directed by either containing a symmetric or asymmetric relation. In the first case, whenever (a, b) is contained in the relation, so is (b, a): for example, the graph can store information of whether Tim is father of Tom or vice versa (or none of that). It can also represent weights of the relationship, by assigning a weight to each element in the relation. Again, the graph itself does not store information about which entity is represented by which node, it is oblivious of any identity of the node. Thus, the *graph* is the more abstract representation which mainly concentrates on the connection structure.

The *network* makes the connection between the graph and the complex system whose interactions it represents. Especially, the network assigns entities to nodes. Furthermore, it is the set of all descriptions and observations of the system in which the entities and their relationship is valid. It can contain observations on the entities like the age and gender of a human actor or the year of publication of a film. It can also contain more than one type of relationship between the actors, or additional observations about the relationships between entities, like the duration of all calls between mobile phone users.

In summary, a *complex network* is a graph, in which the set of elements is associated with a set of entities, and in which the relation between these elements represents a relationship between the corresponding entities.

The distinction between a network and its graph is often not very important, and thus *network* and *graph* are used quite interchangeably in most articles and also in this book.

> **Note 4.** The promise of network analysis is that the abstraction of a complex system as represented by a complex network and its underlying graph still allows to infer something about the complex system of interest. That is actually a strong assumption and later chapters (e.g., Chaps. 10 and 14) elaborate preconditions to enable this transfer.

So, what are the first steps after the data is represented as a network? The following section shows some typical approaches to get a first impression of a new data set on the example of a movie-co-rating network.

Table 1.1 The movie-movie-similarity network contains 494 films and represents 9796 relationships between them

Statistics	Value
n	494
m	9796
$\rho(G)$	0.08

1.2.2 First Probes into the Data

The movie-co-rating network is a network deduced from a so-called *bipartite graph*: As indicated above, some data sets describe a relationship between two different entities, for example, how customer of a video rental store rate the films they rented. Such a data set documents a relationship between customers and the films they rated, but there is no direct relationship between any two customers or between any two films. This kind of data is represented by a bipartite graph, that is, one that can be split in two parts such that all known relationships are only between entities from different parts. Based on such data, one can compute a similarity measure between the films that quantifies whether the films have been more often liked by the same persons than expected or not; this technique is called a *one-mode projection* of a bipartite graph and described in Sect. 13.5.

Such a one-mode projection is the basis for the following demonstrations. It has been created such that the relation is undirected and it is assumed that a pair of films connected by an edge are similar in content. The data comes in a format[3] that is readable by various software applications, e.g., Gephi[4] which is well suited for visually exploring a medium-sized graph [22]. Similarly well suited are yEd [27], visone [13], or Cytoscape [26].

The very first useful information about the data is how many films it contains and how many relationships between them exists. In most visualizations of graphs, this information is given immediately when the graph is opened and displayed. In general, the number of entities is denoted by n and the number of relationships is denoted by m. With around 500 nodes and 10,000 edges, the network is of medium size. From these two basic statistics the so-called *density* of relationships can be computed as another, first inspection into the graph. It is defined as the number of existing relationships divided by the number of possible relationships: in principle, every pair of entities could be related to each other, thus, the number of possible relationships is given by $n(n-1)/2$. The density in the given data with $n = 494$ and $m = 9796$ can thus be computed to be 0.08. This density can also be interpreted as the *probability* that a randomly chosen pair of movies is related; this probability is obviously very small. Table 1.1 summarizes the basic statistics.

[3] Sections 3.6 and 3.7 discusses various graph data formats and how they can be transformed into each other.

[4] Freely downloadable from http://gephi.org/.