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 Katharina A. Zweig
June 2016

http://www.springer.com/978-3-7091-0740-9

Network Analysis Literacy
A Practical Approach to the Analysis of Networks
Zweig, K.A.
2016, XXIII, 535 p. 126 illus., 14 illus. in color.,
Hardcover
ISBN: 978-3-7091-0740-9