Bogumił Kamiński, Paweł Prałat, and François Théberge

Mining Complex Networks

Contents

Preface

Introduction

Data science is a multi-disciplinary field that uses scientific and computational tools to extract valuable knowledge from, typically, large data sets. Once the data is processed and cleaned, it is analyzed and presented in a form that is appropriate to support decision making processes. As collecting data has become much easier and cheaper these days than in the past, data science and machine learning tools have become widely used in companies of all sizes. Indeed, data-driven businesses were worth $1.2 trillion collectively in 2020, an increase from $333 billion in the year 2015, and it seems that this trend is going to persist in the future.

This book concentrates on mining networks, a subfield within data science. Virtually every human-technology interaction, or sensor network, generates observations that are in some relation with each other. As a result, many data science problems can be viewed as a study of some properties of complex networks in which nodes represent the entities that are being studied and edges represent relations between these entities. In these networks (for example, the Instagram on-line social network, the 4th most downloaded mobile app of the 2010s), nodes not only contain some useful information (such as the user's profile, photos, tags) but are also internally connected to other nodes (relations based on follower requests, similar users' behaviour, age, geographic location). Such networks are often large-scale, decentralized, and evolve dynamically over time. Mining complex networks in order to understand the principles governing the organization and the behaviour of such networks is crucial for a broad range of fields of study, including information and social sciences, economics, biology, and neuroscience. Here are a few selected typical applications of mining networks:

1. community detection (which users on some social media platform are close friends),
2. link prediction (who is likely to connect to whom on such platforms),
3. predicting node attributes (what advertisement should be shown to a given user of a particular platform to match their interests),
4. detecting influential nodes (which users on a particular platform would be the best ambassadors of a specific product).

After reading this book, one should be able to answer such questions, and much more, using state-of-the-art methods and computational techniques.

Target Audience

The book was written based on the lecture notes for a graduate course entitled *Graph Mining (DS 8014)* which was offered to students enrolled in the *Data Science and Analytics* Master's program at *Ryerson University* (Toronto, Canada). This textbook is aimed to be suitable for an upper-year undergraduate course or a graduate course. Students in programs such as data science, mathematics, computer science, business, engineering, physics, statistics, and social science will benefit from courses that are based on this textbook. Having said that, this book can be successfully used by all enthusiasts of data science at various levels of sophistication who would like to expand their knowledge or consider changing their career path. The Core Material (Part I) can be successfully used for a 12-week long course (for example, in Canadian system) but we additionally provide the Additional Material (Part II) that can be added for a 15-week long course (for example, in US or European systems).

Need for Another Book

This textbook is not the first (and certainly not the last) book related to network science. There are a number of excellent books, including those that we list in Section 1.15 that conceptually overlap with our book. Let us then present a few reasons why we decided to write this book.

Most books present a mixture of various topics in modelling and mining networks. Modelling complex networks is an important research direction and a few random graph models are included in our book but are mainly used as tools to benchmark and guide algorithms or to create synthetic networks for testing the behaviour of the tools in various scenarios. We focus on aspects related to mining complex networks, and carefully select the most important tools to create a nice and coherent blend that is appropriate for a one term course.

The three authors actively collaborate together, publishing research papers on various topics related to mining networks, including community detection algorithms, mining hypergraphs, unsupervised evaluation of graph embeddings, synthetic random graph models, anomaly detection algorithms, and link prediction algorithms. Our respective individual skills and experiences

nicely complement each other, providing three different perspectives: pure mathematics (Paweł), mining large networks (François), and applying machine learning tools in business (Bogumił). This cumulative experience enables us to carefully select problems and tools that are suitable for a one-term course on mining networks. The content of this textbook represents the most important and useful aspects of the daily life of a data scientist, and with its use, data scientists can make a meaningful impact in business.

Most existing related books concentrate on theory. On the other hand, in our book the theoretical foundations are combined with practical experiments where students are expected to code and analyze graph datasets by themselves. This book is accompanied by Jupyter notebooks[1] (in Python and Julia) which not only contain all of the experiments presented in the book but which also include additional material. We will continue updating them, making sure they work with currently available environments. In particular, we use the `igraph`[2] library for Python which distinguishes us from other books that also use Python for their experiments. The `igraph` network analysis tool was chosen due to its superior performance in dealing with large graphs, and the richness of its library of graph analytics. For example, many centrality measures and graph clustering algorithms are available directly within `igraph`. Moreover, the library is written in C and can be used as such, and there are packages for R and Python, two of the most popular languages for data science. Moreover, we made publicly available videos that walk the reader through our notebooks which should be useful for readers that read the book by themselves and not as a part of a course offered at some university. Finally, we also made slides publicly available for the instructors to use, which should help them to adopt the textbook for their needs and their audience.

A distinguishing feature of mining networks, as opposed to traditional data mining, is that very often one needs to implement custom algorithms to perform an analysis for a given problem at hand. In traditional data mining, there are standard tools such as deep-learning networks, XGBoost, etc., to which we typically just pass appropriately prepared data. In mining networks, despite the fact that there exist standard tools and techniques, they usually require slight modifications to fit the studied problem. Because of this, apart from applying standard algorithms that are pre-implemented in the libraries such as `igraph`, one often needs to complement them with carefully tailored code that is computationally intensive. The reader will be able to notice this characteristic in virtually every chapter of this book. In such cases, one needs tools that allow one to implement such custom code efficiently while ensuring the code's speed (as usually complex networks are large). Traditionally, in such situations data scientists faced the so-called *two language problem*. In order to write the code efficiently Python was used, as it is a nice language for prototyping. However, these implementations were usually not scalable. Therefore,

[1] see `jupyter.org`; also available in Anaconda (`www.anaconda.com`) and other sources
[2] `igraph.org/python`

the next step was to re-write the prototype in some low level language such as C++.

In order to solve the two language problem, in this book we provide implementations of the examples not only using the Python language but also using the Julia language. Julia, like Python, is a high-level language (actually, in many cases the code is quite similar) but at the same time it is compiled (as opposed to Python which is interpreted), which allows the execution speed of the programs to be comparable to languages such as C++. These features of the Julia language have resulted in its popularity increasing recently, not only for mining complex networks but for all kinds of data science tasks that require performance and scalability.

About the Authors

Bogumił Kamiński is the Chairman of the Scientific Council for the Discipline of Economics and Finance at SGH Warsaw School of Economics. He is also an Adjunct Professor at the Data Science Laboratory at Ryerson University. Bogumił is an expert in applications of mathematical modelling for solving complex real life problems. He is also a substantial open-source contributor to the development of the Julia language and its package ecosystem.

Paweł Prałat is a Professor of Mathematics at Ryerson University, whose main research interests are in random graph theory, especially in modelling and mining complex networks. He is the Director of Fields-CQAM Lab on Computational Methods in Industrial Mathematics at The Fields Institute for Research in Mathematical Sciences, and has pursued collaborations with various industry partners as well as the Government of Canada. He has written over 170 papers and 3 books with 130 plus collaborators.

François Théberge holds a B.Sc. degree in applied mathematics and computer science from the University of Ottawa, a M.Sc. in telecommunications from INRS and a PhD. in electrical engineering from McGill University. He has been employed by the Government of Canada since 1996 during which he was involved in the creation of a data science team as well as the research group now known as the Tutte Institute for Mathematics and Computing. He also holds an adjunct professorial position in the Department of Mathematics and Statistics at the University of Ottawa. His current interests include relational-data mining and deep learning.

Accompanied Material

Additional complementary material can be found here

`https://www.ryerson.ca/mining-complex-networks/`