Big Data over Networks

Utilizing both key mathematical tools and state-of-the-art research results, this text explores the principles underpinning large-scale information processing over networks and examines the crucial interaction between big data and its associated communication, social, and biological networks.

Written by experts in the diverse fields of machine learning, optimization, statistics, signal processing, networking, communications, sociology, and biology, this book employs two complementary approaches: first, analyzing how the underlying network constrains the upper layer of collaborative big data processing, and second, examining how big data processing may boost performance in various networks. Unifying the broad scope of the book is the rigorous mathematical treatment of the subjects, which is enriched by in-depth discussion of future directions and numerous open-ended problems that conclude each chapter.

Readers will be able to master the fundamental principles for dealing with big data over large systems, making it essential reading for graduate students, scientific researchers, and industry practitioners alike.

Shuguang Cui is Professor at Texas A&M University. He is a Fellow of the IEEE and was selected as a Highly Cited Researcher by Thomson Reuters, 2014.

Alfred O. Hero III is R. Jamison and Betty Williams Professor of Engineering at the University of Michigan, Ann Arbor, with appointments in the departments of Electrical Engineering and Computer Science, Biomedical Engineering and Statistics. He is a Fellow of the IEEE.

Zhi-Quan Luo is Professor at the University of Minnesota. He has served as the Editor-in-Chief of *IEEE Transactions on Signal Processing* and is a Fellow of the IEEE, SIAM, and the Royal Society of Canada.

José M. F. Moura is Philip L. and Marsha Dowd University Professor at CMU with appointments in the Departments of Electrical and Computer Engineering and, by courtesy, of Biomedical Engineering. He is a Fellow of the IEEE and the AAAS, a corresponding member of the Academy of Sciences of Portugal, and a member of the US NAE.

Big Data over Networks

Edited by

SHUGUANG CUI
Texas A&M University

ALFRED O. HERO III
University of Michigan, Ann Arbor

ZHI-QUAN LUO
University of Minnesota

JOSÉ M. F. MOURA
Carnegie Mellon University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107099005

First published 2016

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalog record for this publication is available from the British Library

ISBN 978-1-107-09900-5 Hardback

Contents

Contributors

Shiqian Ma
The Chinese University of Hong Kong, China

Bo Jiang
Shanghai University of Finance and Economics, China

Xiuzhen Huang
Arkansas State University, USA

Shuzhong Zhang
University of Minnesota, USA

Symeon Chouvardas
University of Athens, Greece

Yannis Kopsinis
University of Athens, Greece

Sergios Theodoridis
University of Athens, Greece

Mingyi Hong
Iowa State University, USA

Wei-Cheng Liao
University of Minnesota, USA

Ruoyu Sun
University of Minnesota, USA

Zhi-Quan (Tom) Luo
University of Minnesota, USA

Jong-Shi Pang
University of Southern California, USA

Meisam Razaviyayn
Stanford University, USA

Ganesh Ananthanarayanan
Microsoft Research, USA

Ishai Menache
Microsoft Research, USA

Chen Gong
University of Science and Technology of China, China

Zhengyuan Xu
University of Science and Technology of China, China

Xiaodong Wang
Columbia University, USA

Suzhi Bi
National University of Singapore, Singapore

Rui Zhang
National University of Singapore, Singapore

Zhi Ding
University of California at Davis, USA

Shuguang Cui
Texas A&M University, USA

Lanchao Liu
University of Houston, USA

Zhu Han
University of Houston, USA

H. Vincent Poor
Princeton University, USA

Riccardo Gallotti
Institut de Physique Théorique, CEA, France

Thomas Louail
Institut de Physique Théorique, CEA, France

Rémi Louf
Institut de Physique Théorique, CEA, France

Marc Barthelemy
Institut de Physique Théorique, CEA, France
and
Centre d'Analyse et de Mathématiques Sociales, EHESS, France

Kathleen M. Carley
Carnegie Mellon University, USA

Wei Wei
Carnegie Mellon University, USA

Kenneth Joseph
Carnegie Mellon University, USA

Jianping Cao
National University of Defense Technology, China

Dongliang Duan
University of Wyoming, USA

Liuqing Yang
Colorado State University, USA

Qingpeng Zhang
City University of Hong Kong, China

Senzhang Wang
Beihang Univerisity, China

Feiyue Wang
National University of Defense Technology, China
and
Chinese Academy of Science, China

Xiaoning Qian
Texas A&M University, USA

Byung-Jun Yoon
Hamad bin Khalifa University, Qatar

Edward R Dougherty
Texas A&M University, USA

Zhe Gan
Duke University, USA

Xin Yuan
Duke University, USA

Ricardo Henao
Duke University, USA

Ephraim L. Tsalik
Durham Veterans Affairs Medical Center, USA
and
Duke University Medical Center, USA

Lawrence Carin
Duke University, USA

Lipi Acharya
Dow AgroSciences LLC, USA

Dongxiao Zhu
Wayne State University, USA

Alfred Hero
University of Michigan, USA

Bala Rajaratnam
Stanford University, USA

Preface

In each day of our modern world, quintillions of bytes of data are generated. This rate keeps increasing, far outpacing the rate at which we can upgrade the computing systems. In fact, a 2011 study by IDC estimates that the amount of data available in the world doubles every two years. This rate of growth closely follows Moore's Law. If we cannot fully understand the issues involved and invent new data processing methods, society will soon be flooded with data, *big data*. Big data sets can exist within a single entity; but most such sets are distributed and can only be aggregated through some type of network. A bigger challenge is that the big data dynamics and the underlying network dynamics are almost always highly correlated. Therefore, understanding the interplay between big data and the associated networks is a critical step in our effort to tackle big data. However, to date, we do not have a systematic theory with which to study the problem thoroughly. Even worse, in some cases, we do not even know how to formulate and approach those problems. We are in need of a comprehensive book to survey and cover both the critical mathematical tools and the state of the art in related research fields.

This book focuses on large-scale information processing over networks, where the meaning of the term information processing can refer to data processing, data storage, or information retrieval, and the term networks may refer to cyber networks, social networks, or biological networks. We take three complementary angles to study the interaction between the data and the underlying network connections. First, we address ways that the underlying network can constrain the upper-layer collaborative big data processing; second, we show how certain big data processing perspectives can help boost the performance in various networks; third, we address the fundamental limits that govern statistical and computational bottlenecks in the analysis of big data. The book consists of chapters contributed by experts from diverse fields spanning machine learning, optimization, statistics, signal processing, networking, communications, sociology, and biology. The core unifying theme of the book is the rigorous mathematical treatment of various subjects, enriched by in-depth discussions of future directions at the end of each chapter. It is expected that this book will not only help researchers in related fields learn the basic tools and understand the state of the art, but also help practitioners formulate approaches to study other large systems that generate and process big data.

The book starts by introducing the recent development of several important mathematical tools for large-scale computation and learning. This provides a useful mathematical framework for studying the big data problem over different types of

networks. Afterwards, the book turns to big data problems in several specific application domains: big data over cyber networks, big data over social networks, and big data over biological networks. One of our main goals is to encourage readers to ponder the large number of interesting interdisciplinary research efforts across different fields that illustrate the problem of big data over networks.

Specifically, in Part I of the book, we focus on some mathematical tools for large-scale data modeling and processing. We start with Chapter 1 on tensor models, which provides a powerful modeling framework for data of high dimensions, with special coverage on the tensor principal component analysis, the tensor low-rank and sparse decomposition models, and the tensor co-clustering problems. Chapter 2 is on the sparsity-aware distributed learning method, where both batch and online algorithms will be discussed. In the batch learning context, the distributed LASSO algorithm and distributed greedy technique will be presented. Furthermore, an LMS-based sparsity promoting algorithm, revolving around the l_1 norm, as well as a greedy distributed LMS will be discussed. In addition, a set-theoretic sparsity promoting distributed technique will be examined. Chapter 3 is on the introduction of optimization algorithms for big data, with the focus on modern first-order large-scale optimization techniques. A few popular first-order methods for large-scale optimization will be surveyed first, including the Block Coordinate Descent (BCD) method, the Block Successive Upper-Bound Minimization (BSUM) method, and the Alternating Direction Method of Multipliers (ADMM). Then the optimal management of a cloud-based densely deployed next-generation wireless network will be studied as a design example. Chapter 4 is on the distributed algorithms for game theoretical approaches, presenting a unified framework for the design and analysis of distributed algorithms for computing first-order stationary solutions of noncooperative games with non-differentiable player objective functions. These games are closely associated with multi-agent optimization wherein a large number of selfish players compete non-cooperatively to optimize their individual objectives under various constraints.

In Part II of this book, we focus on big data over cyber networks, which include computer networks, communication networks, and the cyber part of a smart grid. In Chapters 5 and 6, we discuss the architecture-level issues for big data analytics and storage; in Chapters 7 and 8, we discuss the big data challenges and opportunities in communication networks and smart grid design. Particularly, in Chapter 5, two fundamental aspects are surveyed on the architecture design for a big data analytics system: scheduling and storage. Their key principles, and how these principles are realized in widely-deployed systems, will be described. In Chapter 6, a big data distributed storage system is discussed, employing existing regenerate codes where the storage nodes are scattered in an optical wireless network. A partial downloading scheme is proposed, which allows downloading a portion of the symbols from any storage nodes. A cross-layer wireless resource allocation framework is then formulated and discussed for data reconstruction in such a distributed storage systems employing partial downloading. Channel and power allocation schemes are also investigated for partial downloading in wireless distributed storage systems. In Chapter 7, the interaction between big data and communication networks is the focus. The challenges and opportunities in the design of scalable wireless systems to embrace the big data era are discussed. The state-of-the-art

techniques in wireless big-data processing are reviewed and the potential implementations of key technologies in the future wireless systems are studied. It is argued that proper wireless system designs could harness, and in fact take advantages of the mobile big data traffic. In Chapter 8, the cyber layer of a smart grid is studied from the point of view of data analytics for security. A distributed and user centric system will be introduced, which incorporates end-consumers into its decision processes to provide a cost-effective and reliable energy supply. The applications of big data processing techniques for smart grid security are investigated from two perspectives: how to exploit the inherent structure of the data, and how to deal with the huge size of the data sets.

In Part III of this book, we shift to big data over social networks, where the social network could refer to either the physical interaction or the virtual interaction (e.g., via Facebook or Twitter) among people. In Chapter 9, a city environment is used to illustrate what could be learned from data about this social network with physical interactions. By analyzing data over small, intermediate, and large time scales, different aspects about the city could be learned. At this period of human history that experiences a rapid urban expansion, such a scientific approach appears more important than ever in order to understand the impact of current urban planning on the future evolution of cities. In Chapter 10, we study a social network with virtual interactions, where large sets of Twitter discourses are analyzed. A high-dimensional network approach is presented for assessing such discourses and identifying not just what is being discussed, but the locality, the change, the associated groups, and the structure in these discourses. This approach is applied to data captured with respect to the Arab Spring. The results provide insight into the co-evolution of topics and groups across the region during a period of dramatic social change. In Chapter 11, a study of the social influence from big data is presented. First, different ways of making measurements of social influence are discussed, followed by the descriptions of the algorithms and models that can be used to quantify the propagation of social influences. Then research on optimization of social influence propagation is summarized. Finally, the method of diffusion network inference is presented, and several challenges and open problems are discussed.

In Part IV of this book, we turn to the data analytics over biological networks. In Chapter 12, a general paradigm is discussed for inference validation based on defining a distance between networks and judging validity according to the distance between the original network and the inferred network. Rather than assuming that a single network is inferred, one can take the perspective that the inference procedure leads to an uncertainty class of networks, which contains the ground truth network. Accordingly, a measure of uncertainty is defined in terms of the cost that uncertainty imposes on the objective, for which the model network is to be employed. The example discussed in the chapter involves interventions in the yeast cell cycle network. In Chapter 13, the importance of Bayesian modeling for gene expression analysis is highlighted. Discriminative factor models, which are the particular theme of this chapter, are presented within a principled framework to jointly build factor models and multiple classifiers. As an alternative to Bayesian classifiers based on the traditional probit link, logistic regression and support vector classification are integrated into the modeling scheme, using novel variable augmentation techniques. The factor models are equipped with global-local

shrinkage priors, recently proposed within the machine learning community, with which the number of factors are inferred automatically from the data. Extensions to multi-task learning are also presented. Inference is developed using both MCMC and variational Bayes algorithms, while online learning is further investigated to scale the model to large datasets. In Chapter 14, the focus is on the inference of biological network topologies from big molecular profiling data that is divided into three stages: identification of network components from molecular data, derivation of gene sets related to a network component, and gene-set-based inference of the underlying network topology in the given network component. In Chapter 15, some of the main advances and challenges are discussed in correlation mining in the context of large-scale biomolecular networks with a focus on medicine. The chapter emphasizes that there are fundamental statistical limits to reliable extraction of information from the sample correlation. These limits are associated with phase transitions in the false detection rate when looking for edges and hubs in a correlation or partial correlation network. A new regime of sample complexity is introduced that is ideally suited for big data problems in biology: the purely-high-dimensional regime, in which the number of samples n is fixed while the number p of biomarker variables is very large. A new correlation mining application also discusses discovery of correlation sign flips between edges in a pair of correlation or partial correlation networks. The pair of networks could respectively correspond to a disease (or treatment) group and a control group.

All the above chapters include comprehensive lists of references related to their contents. We hope that by reading this book, the readers could gain a foothold on the exciting activities in the many networking fields that are coupled with big data. We also hope that this book will serve to inspire the readers to develop mathematical approaches to other big data problems not covered here.

Finally, we thank all the chapter authors and the editors from Cambridge University Press, who together made this book possible.

Co-Editors:
Shuguang, Alfred, Zhi-Quan, and José
Spring, 2015