

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

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.

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

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

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

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

© Cambridge University Press 2016

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

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

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

Contents

	<i>List of contributors</i>	page xiii
	<i>Preface</i>	xvii
Part I	Mathematical foundations	1
1	Tensor models: solution methods and applications	3
	Shiqian Ma, Bo Jiang, Xiuzhen Huang, and Shuzhong Zhang	
	1.1 Introduction	3
	1.2 Tensor models	5
	1.2.1 Sparse and low-rank tensor optimization models	5
	1.2.2 Tensor principal component analysis	6
	1.2.3 The tensor co-clustering problem	8
	1.3 Reformulation of tensor models	11
	1.3.1 Low- n -rank tensor optimization	11
	1.3.2 Equivalent formulation of tensor PCA	13
	1.4 Solution methods	16
	1.4.1 Directly resorting to some existing solver	16
	1.4.2 First-order methods	18
	1.4.3 The block optimization technique	22
	1.5 Applications	24
	1.5.1 Computational results on gene expression data	25
	1.6 Conclusions	30
	References	31
2	Sparsity-aware distributed learning	37
	Symeon Chouvardas, Yannis Kopsinis, and Sergios Theodoridis	
	2.1 Introduction	37
	2.2 Batch distributed sparsity promoting algorithms	39
	2.2.1 Problem formulation	39
	2.2.2 LASSO and its distributed learning formulation	40
	2.2.3 Sparsity-aware learning: the greedy point of view	42
	2.2.4 Other distributed sparse recovery algorithms	45

2.3	Online sparsity-aware distributed learning	46
2.3.1	Problem description	46
2.3.2	LMS based sparsity-promoting algorithm	47
2.3.3	The GreeDi LMS algorithm	49
2.3.4	Set-theoretic sparsity-aware distributed learning	51
2.4	Simulation examples	56
2.4.1	Performance evaluation of batch methods	57
2.4.2	Performance evaluation of online methods	58
	References	61
3	Optimization algorithms for big data with application in wireless networks	66
	Mingyi Hong, Wei-Cheng Liao, Ruoyu Sun, and Zhi-Quan Luo	
3.1	Introduction	66
3.1.1	Motivation	66
3.1.2	The organization of the chapter	67
3.2	First-order algorithms for big data	67
3.2.1	The block coordinate descent algorithm	67
3.2.2	The ADMM algorithm	69
3.2.3	The BSUM method	70
3.3	Application to network provisioning problem	72
3.3.1	The setting	72
3.3.2	Network with an uncapacitated backhaul	75
3.3.3	Network with a capacitated backhaul	82
3.4	Numerical results	88
3.4.1	Scenario 1: Performance comparison with heuristic algorithms	89
3.4.2	Scenario 2: The efficiency of N-MaxMin WMMSE algorithm	91
3.4.3	Scenario 3: Multi-commodity routing problem with parallel implementation	92
3.4.4	Scenario 4: Performance evaluation for Algorithm 1 with zones of nodes	94
3.5	Appendix	94
	References	97
4	A unified distributed algorithm for non-cooperative games	101
	Jong-Shi Pang and Meisam Razaviyayn	
4.1	Introduction	101
4.2	The nonsmooth, nonconvex game	104
4.3	The unified algorithm	106
4.3.1	Special cases	108
4.4	Convergence analysis: contraction approach	110
4.4.1	Probabilistic player choices	115

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

4.5	Convergence analysis: potential approach	116
4.5.1	Generalized potential games	121
	References	122
	Appendix	125
Part II	Big data over cyber networks	135
5	Big data analytics systems	137
	Ganesh Ananthanarayanan and Ishai Menache	
5.1	Introduction	137
5.2	Scheduling	139
5.2.1	Fairness	139
5.2.2	Placement constraints	142
5.2.3	Additional system-wide objectives	145
5.2.4	Stragglers	146
5.3	Storage	148
5.3.1	Distributed file system	148
5.3.2	In-memory storage	151
5.4	Concluding remarks	156
	References	158
6	Distributed big data storage in optical wireless networks	161
	Chen Gong, Zhengyuan Xu, and Xiaodong Wang	
6.1	Introduction	161
6.2	Big data distributed storage in a wireless network	163
6.2.1	Wireless distributed storage network framework	163
6.2.2	Optical wireless framework	165
6.2.3	Rateless coded distributed data storage	167
6.2.4	Network coded system with full downloading	167
6.2.5	Network coded system with partial downloading	168
6.3	Reconstructability condition for partial downloading	169
6.3.1	μ -Reconstructability for MSR point	169
6.3.2	μ -Reconstructability for a practical MBR point coding scheme	170
6.4	Channel and power allocation for partial downloading	173
6.4.1	Wireless resource allocation framework	174
6.4.2	Optimal channel and power allocation for the relaxed problem	174
6.5	Open research topics	176
6.5.1	General research topics for wireless distributed storage networks	176
6.5.2	Research topics for data storage in optical wireless networks	177
6.5.3	Research topics for data storage in named data networks	177
	References	178

7	Big data aware wireless communication: challenges and opportunities	180
	Suzhi Bi, Rui Zhang, Zhi Ding, and Shuguang Cui	
	7.1 Introduction	180
	7.2 Scalable wireless network architecture for big data	182
	7.2.1 Hybrid processing structure	182
	7.2.2 Web caching in wireless infrastructure	185
	7.2.3 Data aware processing units	187
	7.3 Wireless system design in big data era	188
	7.3.1 Analog vs. digital backhaul	188
	7.3.2 Joint base station and cloud processing with digital backhaul	191
	7.3.3 Section summary	198
	7.4 Big data aware wireless networking	198
	7.4.1 Wireless big data analytics	200
	7.4.2 Data-driven mobile cloud computing	205
	7.4.3 Software-defined networking design	207
	7.4.4 Section summary	209
	7.5 Conclusions	210
	Acknowledgement	211
	References	211
8	Big data processing for smart grid security	217
	Lanchao Liu, Zhu Han, H. Vincent Poor, and Shuguang Cui	
	8.1 Preliminaries and motivations	217
	8.2 Sparse optimization for false data injection detection	219
	8.2.1 State estimation and false data injection attacks	219
	8.2.2 Nuclear norm minimization	223
	8.2.3 Low-rank matrix factorization	225
	8.2.4 Numerical results	227
	8.3 Distributed approach for security-constrained optimal power flow	232
	8.3.1 Security-constrained optimal power flow	232
	8.3.2 ADMM method	235
	8.3.3 Distributed and parallel approach for SCOPF	236
	8.3.4 Numerical results	238
	8.4 Concluding remarks	241
	Acknowledgement	241
	References	242
Part III	Big data over social networks	245
9	Big data: a new perspective on cities	247
	Riccardo Gallotti, Thomas Louail, Rémi Louf, and Marc Barthelemy	
	9.1 Big data and urban systems	247
	9.2 Infrastructure networks	249

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

9.2.1	Road networks	249
9.2.2	Subway networks	255
9.3	Mobility networks	257
9.3.1	A renewed interest	257
9.3.2	Individual mobility networks	258
9.3.3	From big data to the spatial structure of cities	261
9.4	Scaling in cities	268
9.5	Discussion: towards a new science of cities	272
	Acknowledgments	273
	References	273
10	High-dimensional network analytics: mapping topic networks in Twitter data during the Arab Spring	278
	Kathleen M. Carley, Wei Wei, and Kenneth Joseph	
10.1	Introduction	278
10.2	Arab Spring	280
10.3	General background	280
10.4	Data	281
10.5	The social pulse: geo-temporal trends in Twitter topics and users	284
10.5.1	Methodology	284
10.5.2	Topic overview	285
10.5.3	Over time analysis	287
10.5.4	Characterization of user–topic similarity network	288
10.5.5	Social interaction overview: the reply network	290
10.5.6	Characterization of group structure	291
10.5.7	Key actors	294
10.6	Discussion	295
10.7	Conclusion	297
	Acknowledgements	298
	References	299
11	Social influence analysis in the big data era: a review	301
	Jianping Cao, Dongliang Duan, Liuqing Yang, Qingpeng Zhang, Senzhang Wang, and Feiyue Wang	
11.1	Introduction	301
11.2	Social influence measurement	304
11.2.1	Network-based measures	304
11.2.2	Behavior-based measures	309
11.2.3	Interaction-based measures	312
11.2.4	Topic-based measures	313
11.2.5	Other measures	316
11.3	Influence propagation and maximization	317
11.3.1	Opinion leader identification	317
11.3.2	Influence maximization	319

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

x

Contents

11.3.3	Diffusion network inference	324
11.3.4	Challenges of IP&M	327
11.4	Challenges in big data	327
11.5	Summary	328
	Acknowledgement	329
	References	329
Part IV	Big data over biological networks	335
12	Inference of gene regulatory networks: validation and uncertainty	337
	Xiaoning Qian, Byung-Jun Yoon, and Edward R. Dougherty	
12.1	Introduction	337
12.2	Background	338
12.2.1	Markov chains	339
12.2.2	Logical regulatory networks	340
12.2.3	Control policy for maximal steady-state alteration	341
12.2.4	Inference algorithms	341
12.3	Network distance functions	343
12.3.1	Semi-metrics	343
12.3.2	Rule-based distance	343
12.3.3	Topology-based distance	344
12.3.4	Transition-probability-based distance	344
12.3.5	Steady-state distance	345
12.3.6	Control-based distance	345
12.4	Inference performance	346
12.4.1	Measuring inference performance using distance functions	346
12.4.2	Analytic example	347
12.4.3	Synthetic examples	348
12.5	Consistency	352
12.6	Approximation	352
12.7	Validation from experimental data	353
12.7.1	Metastatic melanoma network inference	354
12.8	Uncertainty quantification	354
12.8.1	Mean objective cost of uncertainty	356
12.8.2	Intervention in yeast cell cycle network with uncertainty	358
	References	360
13	Inference of gene networks associated with the host response to infectious disease	365
	Zhe Gan, Xin Yuan, Ricardo Henao, Ephraim L. Tsalik, and Lawrence Carin	
13.1	Background	365
13.2	Factor models in gene expression analysis	366

13.3	Factor models	367
13.3.1	Shrinkage prior	368
13.3.2	Multiplicative gamma process	369
13.4	Discriminative models	370
13.4.1	Bayesian log-loss	370
13.4.2	Bayesian hinge-loss	372
13.5	Discriminative factor model	372
13.5.1	Multi-task learning	374
13.6	Inference	374
13.7	Experiments	376
13.7.1	Performance measures	377
13.7.2	Experimental setup	377
13.7.3	Classification results	378
13.7.4	Interpretation	382
13.8	Closing remarks	384
13.9	Inference details	385
	Acknowledgements	387
	References	388
14	Gene-set-based inference of biological network topologies from big molecular profiling data	391
	Lipi Acharya and Dongxiao Zhu	
14.1	Introduction	391
14.2	Big data to network components	393
14.3	Gene sets related to network components	394
14.4	Reconstructing biological network topologies using gene sets	395
14.4.1	A general setting	395
14.4.2	Gene set Gibbs sampling	397
14.4.3	Gene set simulated annealing	397
14.5	Discussion and future work	403
	References	406
15	Large-scale correlation mining for biomolecular network discovery	409
	Alfred Hero and Bala Rajaratnam	
15.1	Introduction	409
15.2	Illustrative example	414
15.2.1	Pairwise correlation	416
15.2.2	From pairwise correlation to networks of correlations	417
15.3	Principles of correlation mining for big data	419
15.3.1	Correlation mining for correlation flips between two populations	424
15.3.2	Large-scale implementation of correlation mining	426

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

xii

Contents

15.4	Perspectives and future challenges	427
15.4.1	State-of-the-art in correlation mining	427
15.4.2	Future challenges in correlation mining biomolecular networks	429
15.5	Conclusion	431
	Acknowledgements	431
	References	432
	<i>Index</i>	437

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

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

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

xiv

List of contributors

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

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

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 University, China

Feiyue Wang

National University of Defense Technology, China

and

Chinese Academy of Science, China

Xiaoning Qian

Texas A&M University, USA

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

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

Cambridge University Press

978-1-107-09900-5 - Big Data Over Networks

Edited by Shuguang Cui, Alfred O. Hero III, Zhi-quan Luo and José M. F. Moura

Frontmatter

[More information](#)

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