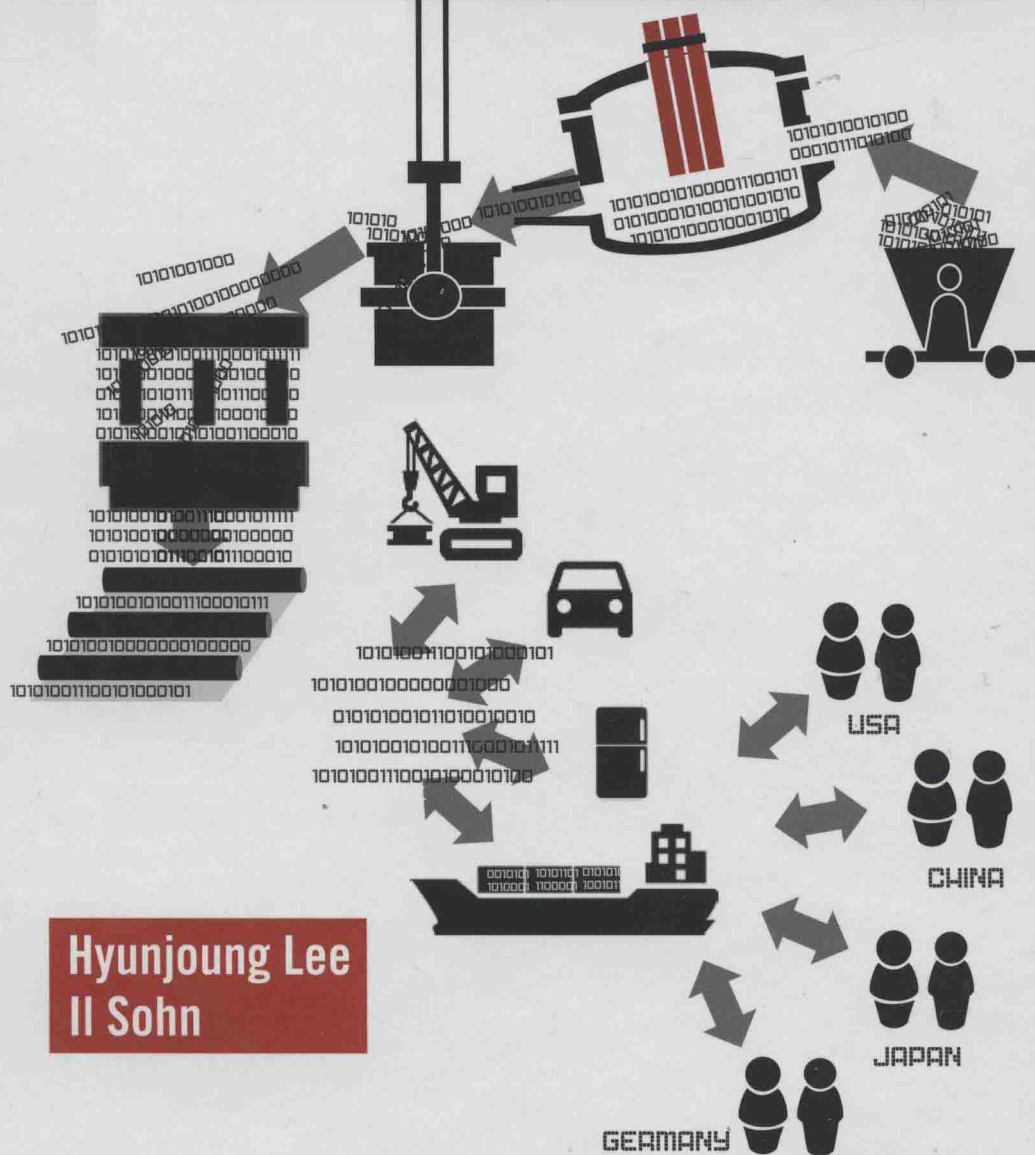


Fundamentals of Big Data Network Analysis for Research and Industry



**Hyunjaung Lee
Il Sohn**



WILEY

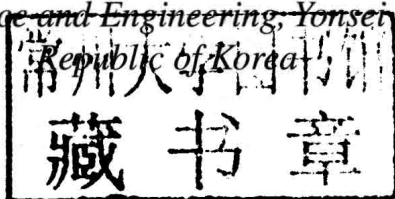
Fundamentals of Big Data Network Analysis for Research and Industry

Hyunjoung Lee

*Institute of Green Technology, Yonsei University,
Republic of Korea*

Il Sohn

*Material Science and Engineering, Yonsei University,
Republic of Korea*



WILEY

This edition first published 2016
© 2016 John Wiley & Sons, Ltd

Registered Office

John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

The advice and strategies contained herein may not be suitable for every situation. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising herefrom.

Library of Congress Cataloging-in-Publication data applied for

A catalogue record for this book is available from the British Library.

ISBN: 9781119015581

Set in 10/12pt Times by SPi Global, Pondicherry, India
Printed and bound in Singapore by Markono Print Media Pte Ltd

Fundamentals of Big Data Network Analysis for Research and Industry

Preface

The concept of the book was first initiated and sponsored by the Future Steel Technology Forum, where future generations of steel researchers gathered to aggregate their knowledge to address the strategic implications of steel technology and product placement across the global trade community. Under the auspices of the Korea Iron and Steel Association, the authors initiated analysis on the steel commodity trade data and the social network relationships among the countries and products of steel currently being traded across the global frontier. From that initiation, the authors were inspired to provide the general public, industry analysts, and students of data analysis on the methodology of big data analysis using examples of steel product trade relations.

This book is separated into six chapters. Chapter 1 defines big data and how it can be applied to business management for higher productivity and efficiency. Chapter 2 describes the various programs related to big data analysis identifying the pros and cons of the commercially available analysis programs. Chapter 3 deals with network analysis and the basic concepts of the nodes and links related to the structure of social network relations between data. As we reach Chapter 4, details of setting up the research methodology for network analysis, methods of data gathering, and cleansing of unwarranted and unnecessary data is illustrated. In Chapter 5, the centrality analysis, which include degree of centrality, betweenness centrality, closeness centrality, is described in detail and the cohesive subgroup is presented. With the conclusion in Chapter 6, the property of the network and equivalence between node pairs or data pairs is outlined with emphasis on the connectivity of nodes. The appendix in the back of the book provides detailed examples of the network analysis performed using the NetMiner program on steel research topics from keyword analysis of journals published in Wiley.

We have come a long way to reach the final destination to inclusively understand the preceding analyses of big data. The various analyses methods and procedures related to big data network analysis introduced here are the most frequently used methods. This book is designed to comprehensively understand the fundamentals of big data and the expected analysis methods to be conducted within a relatively short time for the beginner and intermediate users. A large part of this effort to complete the book has only been possible through the support and sacrifice of many close to the present authors.

We would like to extend our gratitude to Professor Dong Joon Min for his helpful comments and insights for understanding the limitations of steel data analysis,

Dr. Jae Wook Ryu for his consistent support of the authors, and Professor Doo-Hee Lee for inspiration and the drive for academic excellence.

This book is dedicated to our families, whose sacrifice and support has never been fully appreciated by the authors.

Hyunjoung Lee and Il Sohn

About the Authors

Hyunjoung Lee (Ph.D., Korea University, 2007)

Hyunjoung Lee has published 20 articles on issues related to marketing and social network analysis. She is working on several proposals to study various industrial marketing strategies, focusing on trading network structures and underpinning factors behind those trading network structures. She works at Yonsei University as a research professor and teaches marketing, methodology, and statistics to both graduate and undergraduate students in South Korea.

Il Sohn (Ph.D., Carnegie Mellon University, 2007)

Il Sohn has been a faculty member of the Materials Science and Engineering Department at Yonsei University, Korea, since 2009. He received his doctorate from Carnegie Mellon University and has worked in the steel-related industry and academia for more than a decade at U.S. Steel Corporation and Yonsei University. His experience ranges from fundamental research in continuous casting and steel production to the economic analysis and optimization for raw materials utilization in steelmaking. He is currently an associate professor and the associate director for the Research Institute for Iron and Steel Technology, serves on the board of review for *Metals and Materials Transactions B*, is an advisory board member for Steel Research International and the Korean Institute of Metals and Materials, is an associate editor for the *Journal of Sustainable Metallurgy*, and is the founding chair of the Future Steel Technology Forum of Korea. Professor Sohn has been acknowledged by both the academic and the industrial community, receiving numerous awards for his contributions to the profession, including the AISI Medal, Charles-Herty Award, the Hunt-Kelly Award, Marcus A. Grossman Award, and the Iron and Steel Commendation Award from the Ministry of Trade, Industry, and Energy of Korea.

Contents

Preface	ix
About the Authors	xi
List of Figures	xiii
List of Tables	xvii
1 Why Big Data?	1
1.1 Big Data	1
1.2 What Creates Big Data?	6
1.3 How Do We Use Big Data?	9
1.4 Essential Issues Related to Big Data	13
References	14
2 Basic Programs for Analyzing Networks	15
2.1 UCINET	15
2.2 NetMiner	20
2.3 R	22
2.4 Gephi	28
2.5 NodeXL	31
References	32
3 Understanding Network Analysis	35
3.1 Defining Social Network Analysis	35
3.2 Basic SNA Concepts	37
3.2.1 Basic Terminology	37
3.2.2 Representation of a Network	38
3.3 Social Network Data	40
3.3.1 One-Mode and Two-Mode Networks	40
3.3.2 Attributes and Weights	42
3.3.3 Network Data Form	42
References	44
4 Research Methods Using SNA	45
4.1 SNA Research Procedures	46
4.2 Identifying the Research Problem and Developing Hypotheses	47

4.2.1	Identifying the Research Problem	47
4.2.2	Developing Hypotheses	47
4.3	Research Design	49
4.3.1	Defining the Network Model	49
4.3.2	Establishing Network Boundaries	51
4.3.3	Measurement Evaluation	52
4.4	Acquisition of Network Data	54
4.4.1	Survey	54
4.4.2	Interview, Observation, and Experiment	55
4.4.3	Existing Data	56
4.5	Data Cleansing	58
4.5.1	Extraction of the Node and Link	59
4.5.2	Merging and Separation of Data	59
4.5.3	Directional Transformation in the Link	61
4.5.4	Transformation of the Weights in Links	64
4.5.5	Transformation of the Two-Mode Network to a One-Mode Network	66
	References	69
5	Position and Structure	71
5.1	Position	71
5.1.1	Degree Centrality	72
5.1.2	Closeness Centrality	82
5.1.3	Betweenness Centrality	84
5.1.4	Prestige Centrality	85
5.1.5	Broker	88
5.2	Cohesive Subgroup	91
5.2.1	Component	91
5.2.2	Community	92
5.2.3	Clique	93
5.2.4	k-Core	95
	References	96
6	Connectivity and Role	97
6.1	Connection Analysis	98
6.1.1	Connectivity	98
6.1.2	Reciprocity	99
6.1.3	Transitivity	102
6.1.4	Assortativity	104
6.1.5	Network Properties	104
6.2	Role	104
6.2.1	Structural Equivalence	105
6.2.2	Automorphic Equivalence	107
6.2.3	Role Equivalence	109

6.2.4 Regular Equivalence	111
6.2.5 Block Modeling	115
References	117
7 Data Structure in NetMiner	119
7.1 Sample Data	119
7.1.1 01.Org_Net_Tiny1	120
7.1.2 02.Org_Net_Tiny2	120
7.1.3 03.Org_Net_Tiny3	121
7.2 Main Concept	122
7.2.1 Data Structure	122
7.2.2 Creating Data	124
7.2.3 Inserting Data	125
7.2.4 Importing Data	129
7.3 Data Preprocessing	130
7.3.1 Change of Link	130
7.3.2 Extraction and Reordering of the Node and Link	133
7.3.3 Data Merge and Split	136
Reference	140
8 Network Analysis Using NetMiner	141
8.1 Centrality and Cohesive Subgroup	141
8.1.1 Centrality	141
8.1.2 Cohesive Subgroup	147
8.2 Connectivity and Equivalence	153
8.2.1 Connectivity	153
8.2.2 Equivalence	156
8.3 Visualization and Exploratory Analysis	161
8.3.1 Visualization	161
8.3.2 Transformation of the Two-Mode Network to a One-Mode Network	168
Appendix A Visualization	171
A.1 Spring Algorithm	171
A.2 Multidimensional Scaling Algorithm	173
A.3 Cluster Algorithm	173
A.4 Layered Algorithm	174
A.5 Circular Algorithm	174
A.6 Simple Algorithm	175
References	176
Appendix B Case Study: Knowledge Structure of Steel Research	179
Index	193

List of Figures

1.1	Hard-disk drive average cost per gigabytes (unit: US\$)	8
2.1	UCINET 6 interface	16
2.2	Results of density and degree centrality using UCINET. (a) Density and (b) degree centrality	19
2.3	Visualization using NetDraw	19
2.4	NetMiner4 work environment	20
2.5	Results of density and degree centrality analyses in NetMiner. (a) Density and (b) degree centrality	21
2.6	NetMiner data structure and data set. (a) Data structure and (b) data set	25
2.7	The R interface	25
2.8	The Gephi interface	28
2.9	Gephi data laboratory and preview screens	30
2.10	NodeXL interface	32
3.1	A network graph and matrix. (a) Graph (b) matrix	38
3.2	(a) Path and (b) degree	39
3.3	Cut-point and bridges of a network component	40
3.4	Structure of the network data	41
3.5	Transformation of a two-mode network into a one-mode network	41
4.1	Research procedure	46
4.2	PSY's tweets	57
4.3	Visualization of the extracted node and link. (a) Visual network of the extracted node. <i>Node attribute: total export amount >US\$5000million.</i> (b) Visual network of the extracted link. <i>Link attribute: export amount >US\$2500million</i>	61
4.4	Visualization of the two-mode and one-mode networks. (a) two-mode network (export: products–countries) and (b) one-mode network (export: countries–countries)	67
5.1	Visual representation of iron and steel trade	74
5.2	Visualization of the non-directional trade relationship	76
5.3	Visualization of the trade relationship with direction	81
5.4	Visualization of betweenness centrality	86
5.5	Type of broker	89
5.6	(Strong vs. weak) component	92
5.7	Results of component analysis. (a) Weak component and (b) strong component	92

5.8	Community	93
5.9	Results of community analysis	94
5.10	Clique, n-clique, n-clan, and k-plex, and k-core	95
5.11	Results of clique, n-clique, n-clan, k-plex, and k-core	95
6.1	Walk, trail, path	99
6.2	Results of link connectivity	99
6.3	Type of dyad relationship	102
6.4	Type of triad relationship	103
6.5	Triad isomorphism classes	103
6.6	Assortative relationship	104
6.7	Network properties	105
6.8	Structural equivalence	106
6.9	Dendrogram of structural equivalence. (a) Import relationship dendrogram and (b) export relationship dendrogram	109
6.10	Automorphic equivalence	109
6.11	Regular equivalence	111
6.12	Block modeling. (a) Visualization of network, (b) matrix (node by node), (c) block-node affiliation matrix (node by group), (d) block image matrix (group by group), and (e) visualization of block image matrix	116
6.13	Results of block modeling. (a) Block-node affiliation matrix (node by group), (b) block image matrix (group by group), and (c) visualization of block image matrix	116
7.1	Hierarchical structure of NetMiner data	123
7.2	Attribute of node and link	123
7.3	New project type	124
7.4	Workfile	125
7.5	Create the network and node set. (a) Create new 1-mode network, (b) create new sub nodeset, and (c) create new 2-mode network	126
7.6	Insert nodes and node's attributes. (a) Insert new node and (b) insert new node attribute	127
7.7	Insert links and link's attributes. (a) Insert new link and (b) insert new link attribute	128
7.8	Data import	129
7.9	Symmetrize	131
7.10	Transpose	131
7.11	Dichotomize	132
7.12	Reverse	132
7.13	Normalize	133
7.14	Recode. (a) Input variable, (b) dialog box for recode, (c) recoding rules, and (d) output of recoding	134
7.15	Self-loop	135
7.16	Extraction of node and link. (a) QuerySet and (b) new workfile	135
7.17	Neighbor node. (a) Output summary and ego network details	136
7.18	Merge. (a) Main process for merge, (b) one-mode networks before the merge, and (c) one-mode network after the merge	138

7.19	Split. (a) Main process for split and (b) one-mode networks after the split	139
8.1	Degree. (a) [R]Main, (b) [T]Degree, and [T]Node Type	142
8.2	[M]Spring map of degree	143
8.3	Degree centrality. (a) [R]Main, (b) [T]Degree centrality vector, (c) [M]Spring (node size: in degree centrality), and (d) [M]Concentric	143
8.4	Closeness centrality. (a) [R]Main, (b) [T]Closeness centrality vector, (c) [M]Spring (node size: in-closeness centrality), and (d) [M]Concentric	145
8.5	Betweenness centrality. (a) [R]Main, (b) [T]Betweenness centrality vector, (c) [M]Spring (node size: betweenness centrality), and (d) [M]Concentric	145
8.6	Prestige centrality. (a) [R]Main, (b) [T]Eigenvector centrality vector, (c) [T]Reflected/Derived/Constant, (d) [M]Spring (node size: Eigenvector centrality), and (e) [M]Concentric	146
8.7	Brokerage. (a) [R]Main, (b) [T]Brokerage, (c) [M]Spring (node size: total score of brokerage), and (d) [M]Concentric	147
8.8	Component. (a) Input data and main process, (b) [R]Main, (c) [T]Component partition vector, and (d) [M]Clustered	148
8.9	Modularity (community). (a) [R]Main and (b) [T]Community Partition	149
8.10	Betweenness (community). (a) [T]Community Cluster Matrix, (b) [T]Permutation Vector, (c) [C]Dendrogram, and (d) [M]Clustered	150
8.11	Clique. (a) [R]Main, (b) [T]Clique Affiliation Matrix, and (c) [M]Spring	151
8.12	n-Clique. (a) [R]Main, (b) [T]n-Clique Affiliation Matrix, and (c) [M]Spring (n-clique 2)	152
8.13	k-core. (a) [R]Main and (b) [T]k-Core Affiliation Matrix	152
8.14	Connectivity. (a) [T]Node Connectivity Matrix and (b) [M]Spring	153
8.15	Reciprocity and transitivity. (a) Dyad census and (b) triad census	154
8.16	Assortativity. (a) [R]Main–Degree, (b) [R]Main–Team, (c) [T]Assortativity–Degree, and (d) [T]Assortativity–Team	155
8.17	Network properties	155
8.18	Structural equivalence (profile). (a) [T]Profile Matrix and (b) [M]MDS	156
8.19	Structural equivalence (CONCOR). (a) [T]CONCOR Matrix and (b) [M]MDS	157
8.20	Role equivalence. (a) [T]Triad Role Matrix and (b) [T]Local Role Matrix	158
8.21	Regular equivalence. (a) [T]REGGE Matrix, (b) [T]CatRE Matrix, and [T]SimRank Equivalence Matrix (direction: in, dampening parameter: 0.8)	159
8.22	Regular equivalence. (a) [T]Block Image Matrix, (b) [T]Block Sum Matrix, (c) [T]Block Density Matrix, (d) [T]Block-Node Affiliation Matrix, (e) [T]# Nodes, (f) [T]Block Role Typology, and (g) [M]Clustered (G1: department manager, G2: finance, G3: HR, G4: management, G5: marketing, G6: sales)	160

8.23	Node and link styling. (a) Node and link attribute styling and (b) node and link styling (<i>mouse right button > (Multiple) Node (Link) Style</i>)	162
8.24	Pie chart and matrix diagram. (a) Pie chart (input vector: department) and (b) matrix diagram (input network: work interact, permutation vector: department)	165
8.25	Area bar and box plot. (a) Area bar (input two-mode network: purchase) and (b) box plot (dependent variable: age, independent variable: department)	166
8.26	Scatter plot, contour plot, and surface chart. (a) Scatter plot (X axis: age, Y axis: duration), (b) contour plot (X axis: Age, Y axis: duration, Z axis: job-ranking, fitting method: linear), and (c) surface chart (X axis: age, Y axis: duration, Z axis: job-ranking, fitting method: linear)	167
8.27	Network contour plot and network surface plot. (a) Network contour plot (input network: work interact, vector: duration, fitting method: linear) and (b) network surface plot (input network: work interact, vector: duration, fitting method: linear)	168
8.28	Transformation of the two-mode network to a one-mode network	168
A.1	Spring algorithm. (a) Kamada and Kawai, (b) stress majorization, (c) Eades, (d) Fruchterman and Reingold, (e) GEM, and (f) HDE	172
A.2	Multidimensional scaling algorithm. (a) Classical MDS, (b) nonmetric MDS, and (c) Kn-MDS	173
A.3	Cluster algorithm. (a) Clustered Eades and (b) clustered Cola	174
A.4	Layered algorithm	175
A.5	Circular algorithm. (a) Circumference, (b) concentric, and (c) radical	175
A.6	Simple algorithm. (a) Fixed and (b) random	176
B.1	two-mode network (article-keyword)	180
B.2	one-mode network (keyword-keyword)	181
B.3	The screen for importing. (a) Importing on network data and (b) workfile after data importing of NetMiner	181
B.4	Transforming a two-mode network into a one-mode network. (a) Network transform and (b) child workfile	182
B.5	Process control area of centrality. (a) Degree centrality, (b) closeness centrality, and (c) betweenness centrality	185
B.6	The results of degree centrality. (a) [R]Main Report, (b) [T]Degree Centrality Vector, (c) [M]Spring, and (d) [M]Concentric	188
B.7	Query composer	188
B.8	The results of cohesive subgroup. (a) [T]Community Partition and (b) [M]Clustered	189
B.9	Subgroup area mapping in steel research	191