Quantitative Applications in the Social Sciences

154

Social Network Analysis

THIRD EDITION

David Knoke
Song Yang

SOCIAL NETWORK ANALYSIS

Third Edition

Quantitative Applications in the Social Sciences

A SAGE PUBLICATIONS SERIES

1. **Analysis of Variance, 2nd Edition** *Iversen/Norpoth*
2. **Operations Research Methods** *Nagel/Neef*
3. **Causal Modeling, 2nd Edition** *Asher*
4. **Tests of Significance** *Henkel*
5. **Cohort Analysis, 2nd Edition** *Glenn*
6. **Canonical Analysis and Factor Comparison** *Levine*
7. **Analysis of Nominal Data, 2nd Edition** *Reynolds*
8. **Analysis of Ordinal Data** *Hildebrand/Laing/Rosenthal*
9. **Time Series Analysis, 2nd Edition** *Ostrom*
10. **Ecological Inference** *Langbein/Lichtman*
11. **Multidimensional Scaling** *Kruskal/Wish*
12. **Analysis of Covariance** *Wildt/Ahtola*
13. **Introduction to Factor Analysis** *Kim/Mueller*
14. **Factor Analysis** *Kim/Mueller*
15. **Multiple Indicators** *Sullivan/Feldman*
16. **Exploratory Data Analysis** *Hartwig/Dearing*
17. **Reliability and Validity Assessment** *Carmines/Zeller*
18. **Analyzing Panel Data** *Markus*
19. **Discriminant Analysis** *Klecka*
20. **Log-Linear Models** *Knoke/Burke*
21. **Interrupted Time Series Analysis** *McDowall/McCleary/Meidinger/Hay*
22. **Applied Regression, 2nd Edition** *Lewis-Beck/Lewis-Beck*
23. **Research Designs** *Spector*
24. **Unidimensional Scaling** *McIver/Carmines*
25. **Magnitude Scaling** *Lodge*
26. **Multiattribute Evaluation** *Edwards/Newman*
27. **Dynamic Modeling** *Huckfeldt/Kohfeld/Likens*
28. **Network Analysis** *Knoke/Kuklinski*
29. **Interpreting and Using Regression** *Achen*
30. **Test Item Bias** *Osterlind*
31. **Mobility Tables** *Hout*
32. **Measures of Association** *Liebetrau*
33. **Confirmatory Factor Analysis** *Long*
34. **Covariance Structure Models** *Long*
35. **Introduction to Survey Sampling** *Kalton*
36. **Achievement Testing** *Bejar*
37. **Nonrecursive Causal Models** *Berry*
38. **Matrix Algebra** *Namboodiri*
39. **Introduction to Applied Demography** *Rives/Serow*
40. **Micro computer Methodsfor Social Scientists, 2nd Edition** *Schrodt*
41. **Game Theory** *Zagare*
42. **Using Published Data** *Jacob*
43. **Bayesian Statistical Inference** *Iversen*
44. **Cluster Analysis** *Aldenderfer/Blashfield*
45. **Linear Probability, Logit, and Probit Models** *Aldrich/Nelson*
46. **Event History and Survival Analysis, 2nd Edition** *Allison*
47. **Canonical Correlation Analysis** *Thompson*
48. **Models for Innovation Diffusion** *Mahajan/Peterson*
49. **Basic Content Analysis, 2nd Edition** *Weber*
50. **Multiple Regression in Practice** *Berry/Feldman*
51. **Stochastic Parameter Regression Models** *Newbold/Bos*
52. **Using Microcomputers in Research** *Madron/Tate/Brookshire*
53. **Secondary Analysis of Survey Data** *Kiecolt/Nathan*
54. **Multivariate Analysis of Variance** *Bray/Maxwell*
55. **The Logic of Causal Order** *Davis*
56. **Introduction to Linear Goal Programming** *Ignizio*
57. **Understanding Regression Analysis, 2nd Edition** *Schroeder/Sjoquist/Stephan*
58. **Randomized Response and Related Methods, 2nd Edition** *Fox/Tracy*
59. **Meta-Analysis** *Wolf*
60. **Linear Programming** *Feiring*
61. **Multiple Comparisons** *Klockars/Sax*
62. **Information Theory** *Krippendorff*
63. **Survey Questions** *Converse/Presser*
64. **Latent Class Analysis** *McCutcheon*
65. **Three-Way Scaling and Clustering** *Arabie/Carroll/DeSarbo*
66. **Q Methodology, 2nd Edition** *McKeown/Thomas*
67. **Analyzing Decision Making** *Louviere*
68. **Rasch Models for Measurement** *Andrich*
69. **Principal Components Analysis** *Dunteman*
70. **Pooled Time Series Analysis** *Sayrs*
71. **Analyzing Complex Survey Data, 2nd Edition** *Lee/Forthofer*
72. **Interaction Effects in Multiple Regression, 2nd Edition** *Jaccard/Turrisi*
73. **Under standing Significance Testing** *Mohr*
74. **Experimental Design and Analysis** *Brown/Melamed*
75. **Metric Scaling** *Weller/Romney*
76. **Longitudinal Research, 2nd Edition** *Menard*
77. **Expert Systems** *Benfer/Brent/Furbee*
78. **Data Theory and Dimensional Analysis** *Jacoby*
79. **Regression Diagnostics, 2nd Edition** *Fox*
80. **Computer-Assisted Interviewing** *Saris*
81. **Contextual Analysis** *Iversen*
82. **Summated Rating Scale Construction** *Spector*
83. **Central Tendency and Variability** *Weisberg*
84. **ANOVA: Repeated Measures** *Girden*
85. **Processing Data** *Bourque/Clark*
86. **Logit Modeling** *DeMaris*
87. **Analytic Mapping and Geographic Databases** *Garson/Biggs*
88. **Working With Archival Data** *Elder/Pavalko/Clipp*
89. **Multiple Comparison Procedures** *Toothaker*
90. **Nonparametric Statistics** *Gibbons*
91. **Nonparametric Measures of Association** *Gibbons*
92. **Understanding Regression Assumptions** *Berry*
93. **Regression With Dummy Variables** *Hardy*
94. **Loglinear Models With Latent Variables** *Hagenaars*
95. **Bootstrapping** *Mooney/Duval*
96. **Maximum Likelihood Estimation** *Eliason*
97. **Ordinal Log-Linear Models** *Ishii-Kuntz*
98. **Random Factors in ANOVA** *Jackson/Brashers*
99. **Univariate Tests for Time Series Models** *Cromwell/Labys/Terraza*

Quantitative Applications in the Social Sciences

A SAGE PUBLICATIONS SERIES

100. **Multivariate Tests for Time Series Models** *Cromwell/Hannan/Labys/Terraza*
101. **Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models** *Liao*
102. **Typologies and Taxonomies** *Bailey*
103. **Data Analysis: An Introduction** *Lewis-Beck*
104. **Multiple Attribute Decision Making** *Yoon/Hwang*
105. **Causal Analysis With Panel Data** *Finkel*
106. **Applied Logistic Regression Analysis, 2nd Edition** *Menard*
107. **Chaos and Catastrophe Theories** *Brown*
108. **Basic Math for Social Scientists: Concepts** *Hagle*
109. **Basic Math for Social Scientists: Problems and Solutions** *Hagle*
110. **Calculus** *Iversen*
111. **Regression Models: Censored, Sample Selected, or Truncated Data** *Breen*
112. **Tree Models of Similarity and Association** *Corter*
113. **Computational Modeling** *Taber/Timpone*
114. **LISREL Approaches to Interaction Effects in Multiple Regression** *Jaccard/Wan*
115. **Analyzing Repeated Surveys** *Frebaugh*
116. **Monte Carlo Simulation** *Mooney*
117. **Statistical Graphics for Univariate and Bivariate Data** *Jacoby*
118. **Interaction Effects in Factorial Analysis of Variance** *Jaccard*
119. **Odds Ratios in the Analysis of Contingency Tables** *Rudas*
120. **Statistical Graphics for Visualizing Multivariate Data** *Jacoby*
121. **Applied Correspondence Analysis** *Clausen*
122. **Game Theory Topics** *fink/Gates/Humes*
123. **Social Choice: Theory and Research** *Johnson*
124. **Neural Networks** *Abdi/Valentin/Edelman*
125. **Relating Statistics and Experimental Design: An Introduction** *Levin*
126. **Latent Class Scaling Analysis** *Dayton*
127. **Sorting Data: Collection and Analysis** *Coxon*
128. **Analyzing Documentary Accounts** *Hodson*
129. **Effect Size for ANOVA Designs** *Cortina/Nouri*
130. **Nonparametric Simple Regression: Smoothing Scatterplots** *Fox*
131. **Multiple and Generalized Nonparametric Regression** *Fox*
132. **Logistic Regression: A Primer** *Pampel*
133. **Translating Questionnaires and Other Research Instruments: Problems and Solutions** *Behling/Law*
134. **Generalized Linear Models: A Unified Approach, 2nd Edition** *Gill/Torres*
135. **Interaction Effects in Logistic Regression** *Jaccard*
136. **Missing Data** *Allison*
137. **Spline Regression Models** *Marsh/Cormier*
138. **Logit and Probit: Ordered and Multinomial Models** *Borooah*
139. **Correlation: Parametric and Nonparametric Measures** *Chen/Popovich*
140. **Confidence Intervals** *Smithson*
141. **Internet Data Collection** *Best/Krueger*
142. **Probability Theory** *Rudas*
143. **Multilevel Modeling, 2nd Edition** *Luke*
144. **Polytomous Item Response Theory Models** *Ostini/Nering*
145. **An Introduction to Generalized Linear Models** *Dunteman/Ho*
146. **Logistic Regression Models for Ordinal Response Variables** *O'Connell*
147. **Fuzzy Set Theory: Applications in the Social Sciences** *Smithson/Verkuilen*
148. **Multiple Time Series Models** *Brandt/Williams*
149. **Quantile Regression** *Hao/Naiman*
150. **Differential Equations: A Modeling Approach** *Brown*
151. **Graph Algebra: Mathematical Modeling With a Systems Approach** *Brown*
152. **Modern Methods for Robust Regression** *Andersen*
153. **Agent-Based Models, 2nd Edition** *Gilbert*
154. **Social Network Analysis, 3rd Edition** *Knoke/Yang*
155. **Spatial Regression Models, 2nd Edition** *Ward/Gleditsch*
156. **Mediation Analysis** *Iacobucci*
157. **Latent Growth Curve Modeling** *Preacher/Wichman/MacCallum/Briggs*
158. **Introduction to the Comparative Method With Boolean Algebra** *Caramani*
159. **A Mathematical Primer for Social Statistics** *Fox*
160. **Fixed Effects Regression Models** *Allison*
161. **Differential Item Functioning, 2nd Edition** *Osterlind/Everson*
162. **Quantitative Narrative Analysis** *Franzosi*
163. **Multiple Correspondence Analysis** *LeRoux/Rouanet*
164. **Association Models** *Wong*
165. **Fractal Analysis** *Brown/Liebovitch*
166. **Assessing Inequality** *Hao/Naiman*
167. **Graphical Models and the Multigraph Representation for Categorical Data** *Khamis*
168. **Nonrecursive Models** *Paxton/Hipp/Marquart-Pyatt*
169. **Ordinal Item Response Theory** *Van Schuur*
170. **Multivariate General Linear Models** *Haase*
171. **Methods of Randomization in Experimental Design** *Alferes*
172. **Heteroskedasticity in Regression** *Kaufman*
173. **An Introduction to Exponential Random Graph Modeling** *Harris*
174. **Introduction to Time Series Analysis** *Pickup*
175. **Factorial Survey Experiments** *Auspurg/Hinz*
176. **Introduction to Power Analysis: Two-Group Studies** *Hedberg*
177. **Linear Regression: A Mathematical Introduction** *Gujarati*
178. **Propensity Score Methods and Applications** *Bai/Clark*
179. **Multilevel Structural Equation Modeling** *Silva/Bosancianu/Littvay*
180. **Gathering Social Network Data** *adams*
181. **Generalized Linear Models for Bounded and Limited Quantitative Variables** *Smithson/Shou*
182. **Exploratory Factor Analysis** *Finch*
183. **Multidimensional Item Response Theory** *Bonifay*

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

SOCIAL NETWORK ANALYSIS

Third Edition

David Knoke
University of Minnesota

Song Yang
University of Arkansas

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London, EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
18 Cross Street #10-10/11/12
China Square Central
Singapore 048423

Acquisitions Editor: Helen Salmon
Editorial Assistant: Megan O'Heffernan
Production Editor: Tracy Buyan
Copy Editor: Celia McCoy
Typesetter: Hurix Digital
Proofreader: Jennifer Grubba
Indexer: Jeanne Busemeyer
Cover Designer: Candice Harman
Marketing Manager: Shari Countryman

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: Knoke, David, author. | Yang, Song, author.

Title: Social network analysis / David Knoke, University of Minnesota, Song Yang, University of Arkansas.

Description: Third edition. | Thousand Oaks, California: SAGE, [2020] |

Series: Quantitative applications in the social sciences; 154 | Includes bibliographical references and index.

Identifiers: LCCN 2019031097 | ISBN 9781506389318 (paperback) | ISBN 9781506389295 (epub) | ISBN 9781506389301 (epub) | ISBN 9781506389325 (web pdf)

Subjects: LCSH: Social networks.

Classification: LCC HM741 .K66 2020 | DDC 302.3—dc23

LC record available at https://lccn.loc.gov/2019031097

This book is printed on acid-free paper.

19 20 21 22 23 10 9 8 7 6 5 4 3 2 1

CONTENTS

SERIES EDITOR'S INTRODUCTION

Much has happened since the publication of *Social Network Analysis*, 2nd Edition. Perhaps most importantly from the standpoint of its content, "social network" has entered the modern lexicon. Facebook and YouTube started in the mid-2000s, quickly followed by Twitter, Snapchat, WhatsApp, and others. Social media applications have exploded. The percentage of U.S. adults using at least one social media site increased from 5% in 2005 to 25% in 2008, to 50% in 2011, and is now nearly 75% according to Pew Research Center estimates. Of course, social networks are not new. They have formed the organizational backbone of social life for many millennia and have been a focus of social science research for almost a century. What is new is broad public interest in social networks, including how they can be manipulated, for good or ill. Also new is the creation and accumulation of massive online datasets reflecting and recording participation in social media. These trends have inspired David Knoke and Song Yang to issue a new edition of their classic text.

As with the earlier editions, *Social Network Analysis*, 3rd Edition, provides a concise introduction to the concepts and tools of social network analysis. The authors are highly regarded technical experts, and the field itself can be quite complicated, but, as was the case with the earlier editions, this "little green cover" is readily accessible. Professors Knoke and Yang convey key material while at the same time minimize technical complexities. The examples are simple—sets of five or six entities such as individuals, positions in a hierarchy, political offices, and nation-states. The set or sets of relations between them include friendship, communication, supervision, donations, and trade.

As with earlier editions, *Social Network Analysis*, 3rd Edition, would serve well as a course supplement at the undergraduate or graduate level. The authors have gone to great lengths to keep the math simple in all but the final chapter of the monograph. The volume is organized in a clear and straightforward manner. After a brief introduction in the first chapter, which situates the study of social networks in a broader context, the second chapter takes up "network fundamentals," defines central concepts, and demonstrates multiple perspectives on how networks can be viewed and studied. Chapter 3 addresses social network data collection, specifically, how the choices made at the design phase such as how to define membership and where to set the boundary, how to sample network entities, and which relations to measure affect subsequent analysis and inference. This chapter also discusses missing data and data quality more generally. In Chapter 4, Professors Knoke and Yang introduce basic methods for analyzing networks, presenting measures of nodes (e.g., degree centrality), dyads (e.g., reachability), subgroups (e.g., cliques), and whole networks (e.g., centralization). They describe and explain strict and more relaxed forms of structural equivalence at the end of the chapter. Level of difficulty increases in Chapter 5. Matrix algebra is needed for parts of this chapter, whereas basic algebra is all that is needed for Chapters 1 through 4. Chapter 5 introduces readers to advanced analytic methods such as clustering, multidimensional scaling, blockmodeling, community detection, and exponential random graph models (ERGMs), preparing them to read the technical literature on these topics.

In comparison with earlier editions, *Social Network Analysis,* 3rd Edition, reflects developments and changes in practice over the past decade. To begin with, Professors Knoke and Yang update the specific language used by network researchers (e.g., whole networks rather than complete networks). In addition, they expand coverage of some topics. For example, whereas the earlier edition presented affiliation models in terms of bipartite models alone, the third edition provides a more general discussion, covering tripartite as well as bipartite models. The authors also describe important recent developments in network analysis, especially in the fifth chapter. ERGMs are a prime example. Analysts interested in statistically modeling network ties as an outcome need to account for clustering and endogeneity. When the second edition was published, P* models were the recommended approach for this, but they have been replaced by ERGMs since then. Finally, throughout the volume, Professors Knoke and Yang comment on the challenges and opportunities offered by Internet and social media data.

Social Network Analysis is one of the most popular "little green books" in the *Quantitative Applications in the Social Sciences* series. It draws on the authors' years of experience to provide an initial entrée into a highly complex area of study, laying a firm foundation on which readers at all levels can continue to build. With the publication of the third edition, if anything, its popularity will increase.

—Barbara Entwisle

Series Editor

ABOUT THE AUTHORS

David Knoke (Ph.D., University of Michigan, 1972) is a professor of sociology at the University of Minnesota, where he teaches and does research on diverse social networks, including political, economic, healthcare, intra- and interorganizational, and terrorist and counterterror networks. In addition to many articles and chapters, he has written seven books about networks: *Network Analysis* (1982, with James Kuklinski), *The Organizational State* (1985, with Edward Laumann), *Political Networks* (1990), *Comparing Policy Networks* (1996, with Franz Pappi, Jeffrey Broadbent, and Yutaka Tsujinaka), *Changing Organizations* (2001), *Social Network Analysis* (2008, with Song Yang), and *Economic Networks* (2012).

Song Yang (Ph.D., University of Minnesota, 2002) is a professor of sociology and criminology at the University of Arkansas. His teaching and research areas are social network analysis, including business, economic, and organizational networks; work and organization studies; and social statistics. He has published many articles and chapters, with the most recent ones appearing in *Journal of Business Research* and *Nonprofit and Voluntary Sector Quarterly*. He has written several books, including *Social Network Analysis* (2008, with David Knoke), *The Invisible Hands of Political Parties in Presidential Elections: Party Activists and Political Aggregation From 2004 to 2012* (2013, with Andrew Dowdle, Scott Limbocker, Patrick Stewart, and Karen Sebold), and *Social Network Analysis: Methods and Examples* (2016, with Franziska Keller and Lu Zheng).

ACKNOWLEDGMENTS

SAGE gratefully acknowledges feedback from the following reviewers for this edition:

Alexandra Marin, University of Toronto
Allan Watson, Staffordshire University
Ge Jiang, University of Illinois at Urbana-Champaign
Howard Lune, Hunter College, CUNY
Michael Levin, Otterbein University
Brian G. Southwell, RTI International and Duke University

Chapter 1

INTRODUCTION TO SOCIAL NETWORK ANALYSIS

Social networks are as old as the human species. As small bands of hunter-gatherers spread around the globe, their survival depended on cooperative strategies for pursuing game and finding good foraging grounds. Ties of family and extended kin were crucial to raising the next generations. With increased size and density of agrarian settlements, succeeded by expanding urban civilizations, networks grew increasingly complex and indispensable for merchants involved in long-distance commerce and armies engaged in conquest. Palace and court intrigues ran on gossip, rumor, and favor-trading among political factions. Scientific and technological advances necessitated information flows through invisible colleges of experts. Social networks have a truly ancient lineage yet are seldom noted nor well understood by their participants.

People today commonly envision social networking as clusters of coworkers going for lunch or coffee, teams of dormmates playing basketball or softball, and bunches of friends chewing the fat. Yes, those small groups are all social networks. To give a formal definition, a *social network* is a set of actors, or other entities, and a set or sets of relations defined on them. In the three preceding examples, the first actors are coworkers and the relations are lunchmate and coffeemate; the second actors are residents of the same dorm and playing sports is the relation; the third network is friends gossiping leisurely. Applying the definition to diverse social settings, we can easily uncover numerous social networks, some more formal than the three previously described. For example, a college academic unit has a social network composed of faculty members, staff, students, and administrators. Multiple sets of relations suffuse such networks: collegial relations among faculty members, faculty advising graduate students, faculty instructing undergraduates, and administrators supervising faculty and staff. A police department is also structured as a formal social network, in which officers at the same rank are colleagues, whereas a quasimilitary chain of command establishes hierarchical authority relations. Typical order from top down would consist of chief of police, deputy chief, captain, lieutenant, sergeant, corporal, patrol officer.

Although people typically conceive the actors in social networks as human beings, they can just as well be collective entities or aggregated units, such as teams, groups, organizations, neighborhoods, political parties, and even nation-states. For example, corporations can engage in cooperative and competitive relations to pursue many outcomes, such as jointly developing new technologies and products or acquiring greater market shares (Knoke, 2001). Interorganizational relations take many governance forms, from contractual agreements to equity stakes (Child, 2005; Yang, Franziska, & Lu, 2016). Inside organizations, work groups and teams often engage in knowledge transfers or information sharing to facilitate innovation and improve task performance (Tsai, 2001). International relational networks also emerge and evolve, including military alliances and conflicts, trade partnerships and disputes, human migrations, intelligence exchanges, and technology sharing and embargoes (Yang et al., 2016, Chapter 8).

Nonsocial networks are prevalent in many domains: technology networks, computer networks and the Internet, telephone networks and electrical power grids, transportation and logistics networks, food delivery, and patent-citation networks. They share some similarities with social networks, except that instead of actors their units are physical entities, such as computers and transformers, and their relations are transmission and delivery lines such as Ethernet cables, wireless connections, airline routes, and interstate highways. We mention nonsocial networks primarily to note that networks are the subjects of studies by many disciplines besides the social sciences. Those investigations illuminate and inspire one another, engendering strong momentum to improve network knowledge, including social network analysis (Knoke & Yang, 2008). For example, after mathematicians developed graph theory, computer scientists applied it to construct optimal computer networks. Social network scholars can borrow algorithms from computer and mathematical sciences to decipher communication networks among friends, coworkers, and organizations.

Sociology built a long tradition of examining the social contexts of social networks. Founding fathers such as Georg Simmel, Émile Durkheim, and Max Weber promoted a structural perspective in the study of human behaviors. Social psychologist Jacob Moreno (1934) was directly responsible for laying the foundation of modern social network analysis. With Helen Jennings, Moreno invented *sociometry* to draw maps visualizing individuals and their interpersonal relations, revealing complex structural relations with simple diagrams. Moreover, Moreno and other pioneering social network scholars endeavored to explain how network structures affect human behaviors and psychological states (Freeman, 2004). On the one hand, we can better understand people's actions and decisions by examining their social networks because networks provide participants with both opportunities and constraints. On the other hand, the formation and change of social networks themselves have been the object of many research projects. An important sociological principle is social *homophily*, which asserts that people tend to form positive relations with others similar to themselves.

Actors could be attracted to others based on similarity of attributes—such as gender, age, race, ethnicity, or socioeconomic status—or similarity of behaviors—such as life experiences, political preferences, religious beliefs, or hobby interests. In this perspective, social relations are outcomes, or dependent variables, occurring because actors share some of the independent variables listed previously.

Social network analysis was vitally important to the inception of economic sociology, a major specialty in sociology. In his classical article applying sociology to economic actions, Mark Granovetter (1985) criticized the undersocialized view of economists in which human decision making is driven solely by subjective expected utility maximization. Surprisingly, Granovetter likewise disapproved of the oversocialized view of sociologists in which human actions are determined solely by norms and social roles. So how does one avoid both under- and oversocialized explanations of human behaviors? The answer, quite obviously, is by using social network analysis: by looking at actors' social networks, we can better understand their decisions and actions. Social networks generate localized norms, rules, and expectations among their members, which reinforce mutual trust and sanction malfeasance. Thus, by examining how social networks actually operate as both causes and consequences of human perceptions and actions, theorists and researchers avoid accepting either oversocialized or undersocialized perspectives. More importantly, although Granovetter (1985) emphasized economic behaviors, his arguments are very relevant to many social pursuits, such as making friends, casting votes, looking for a job, seeking promotion, finding a therapist, searching for emotional support, and locating instrumental help.

Early sociological and anthropological research on social networks inspired other disciplines to investigate the mechanisms instigating network formation in those fields. Over the past half century, mass communication, strategic management, marketing, logistics, public administration, political science, international relations, psychology, public health, criminology, and even economics begin introducing ideas and methods of social network analysis into those disciplines. For example, Zeev Maoz (2012) analyzed international trade and military alliances as network processes. He found that international trade follows a preferential attachment or bandwagon process: all nations want a quick and short connection to a few key nations in the global trade network, resulting in a highly condensed, single-core structure. In contrast, for military alliances, nations tend to partner with countries sharing similar political ideologies and regime structures. This homophily preference produces a network configuration consisting of multiple small military alliance clusters that are only sparsely interconnected (see also Yang et al., 2016, p. 198).

We would be remiss not to mention social media as an explosively growing component of social networks. Facebook, Twitter, LinkedIn, WeChat, and other apps facilitate a massive amount of daily information exchange among billions of users. Much social networking nowadays occurs in virtual spaces as users contact one another via computers, laptops, iPad tablets, and smartphones linked together by Ethernet cables or wireless. Computer communication networks and human social networks converge, engendering innumerable research opportunities and challenges for social and computer scientists. How does one best search, capture, aggregate, store, share, process, reduce, and visualize vast volumes of complex data generated by online social networkers (Press, 2013; Lohr, 2013)? John Mashey, chief scientist at Silicon Graphics, is often credited with coining the term Big Data, which he described in a slide presentation as "storage growing bigger faster" (1998, p. 2). Exponentially bourgeoning quantities of structured and unstructured information have revolutionized businesses, nonprofits, and governments. For social network researchers, Big Data is a trove of rich relational databases and a smörgåsbord of computer tools for data mining, information fusion, computational intelligence, machine learning, and other applications (de Nooy, Wouter, Mrvar, & Batagelj, 2018). Although Big Data enhances organizational operations and outcomes, it also raises numerous ethical and privacy challenges, such as the rise of surveillance state capacities to predict and control populations (Brayne, 2017; Madden, Gilman, Levy, & Marwick, 2017). Russian manipulation of the 2016 U.S. presidential election was only the most notorious of innumerable criminal abuses of Big Data on social media platforms. Calls for governmental regulation of social media companies encounter conundrums of how to protect platforms and safeguard free speech while prohibiting dangerous content (Berman, 2019). The fate of our democracy hangs in the balance.

In sum, social network analysis is a vibrant multidisciplinary field. Peter Carrington and John Scott called it "a 'paradigm', rather than a theory or a method: that is, a way of conceptualizing and analyzing social life" (2011, p. 5). We believe the network paradigm has roots in and thrives on the integration of three elements: theories, methodologies, and applications. For theories, network analysis demands serious commitment that prioritizes actor interdependence and connectivity, emphasizing structured relations among social entities. For methodologies, network analysis borrows eclectically from diverse disciplines, collaborating across the aisles to create innovative procedures. For applications, people increasingly use their networking skills to navigate along complex interorganizational pathways to acquire desired goods and services, such as better healthcare, shopping bargains, and recreational experiences.

This volume updates the second edition of *Social Network Analysis* by Knoke and Yang (2008). In addition to providing a general overview of fundamental methodological topics, we cover new developments of the past decade. Our approach is didactic, aimed primarily at graduate students and professionals in many social science disciplines, including sociology, political science, business management, anthropology, economics, psychology, public administration, public health, and human resources. College faculty could assign it as a text in graduate-level courses, use it for workshops at professional association meetings or summer instructional institutes, or study it to learn more about networks on their own. Graduate and advanced undergraduate students interested in social network analyses can read it to get a jump-start on their social network skills and intellectual aspirations. Professionals face many challenges in developing social network research, such as how to design a social network project, details and problems that may arise during network data collection, and alternative techniques for analyzing their social network data. Social network scholars may find this volume a useful brief refresher or reference book. For more advanced texts, we suggest Easley and Kleinberg (2010); Dorogovtsev and Mendes (2014); Lazega and Snijders (2015); de Nooy, Mrvar, and Batagelj (2018); and Newman (2010).

We frequently illustrate concepts and methods by referring to substantive social network research problems, citing examples from children's playgroups to organizations, communities, and international systems. We tried to write with a precision and freshness of presentation using concise language that minimizes technical complexities. The book consists of five substantive chapters. Chapter 2 introduces fundamental network assumptions and concepts, as applied to a variety of units of observation, levels of analysis, and types of measures. It contrasts relational contents and forms of relations and distinguishes between egocentric and whole networks. The structural approach emphasizes the value of network analysis for uncovering deeper patterns beneath the surface of empirical interactions. Chapter 3 concerns issues in collecting network data: boundary specification, data collection procedures, cognitive social structures, missing data, measurement error, and collecting online social media and Big Data. In Chapter 4, we discuss basic methods of network analysis, including graphs and matrices; centrality, prestige, and power; social distance, paths, walks, and reachability; transitivity and cliques; and size, centralization, density, and different measure of equivalence for pairs of actors or entities. Chapter 5 gives an overview of more-advanced methods of network analysis, including ego-nets; clustering, multidimensional analysis, and blockmodels; 2-mode and 3-mode networks; community detection; and exponential random graph models. The final section concludes with some speculations about future directions in social network analysis.

After years of painstaking efforts, network analysts developed several computer packages to facilitate social network data collection and analyses. Softwares vary on many dimensions, such as operating systems, affordability, learning curves, and strengths and weaknesses. We attached an Appendix that summarizes some useful packages and contrasts them on those dimensions. We remain most impressed, however, with the breadth and user-friendly qualities of UCINET (Borgatti, Everett, & Freeman, 2002) as both a teaching and a research tool for smaller-scale social network analyses. Consequently, we used it to make this edition whenever we demonstrated social network analysis methods.

Chapter 2

NETWORK FUNDAMENTALS

In this chapter, we discuss fundamental concepts for understanding social network analysis methods. We use terms and definitions most widespread and accepted by academic researchers but in instances of disagreement defer to sociological perspectives. We cite many examples from diverse disciplines that illustrate these basic concepts. Interested readers should read numerous publications to deepen their understanding of how network analysis methods can be applied to investigate substantive problems in their fields.

To clarify the distinctive social network perspective on social action, a contrast to individualistic, variable-based approaches may be insightful. Many social science theories, possibly a large majority, assume that actors make decisions and act without regard to the behavior of other actors. Whether analyzed as utility-maximizing rational calculations or as drive-reduction motivation based on causal antecedents, such explanations primarily consider only the characteristics of persons while ignoring the broader interaction contexts within which social actors are embedded. In contrast, network analysis explicitly assumes that actors participate in social systems connecting them to other actors and that their relations comprise important influences on one another's behaviors. Central to the theoretical and methodological agenda of network analysis is identifying, measuring, and testing hypotheses about the structural forms and substantive contents of relations among actors. This distinctive structural-relational emphasis sets social network analysis apart from the individualistic, variable-centric traditions still prevalent in much social science theory and research. We see encouraging signs that many social science disciplines are increasingly embracing structural-relational explanations of social action.

2.1 Underlying Assumptions

The network perspective emphasizes *structural relations* as its key orienting principle. Siegfried Nadel, the great British anthropologist, proposed a relational definition of social structure: "We arrive at the structure of a society through abstracting from the concrete population and its behaviour the pattern or network (or 'system') of relationships obtaining 'between actors in their capacity of playing roles relative to one another'" (Nadel, 1957, p. 12). By network, he meant "the interlocking of relationships whereby the interactions implicit in one determine those occurring in others" (p. 16). By separating structural forms from their empirical contents, structural analysts can uncover the underlying systems of roles that arise from interdependent activities of the persons performing those roles. Nadel further contributed to nascent network science by suggesting that matrix methods could graphically depict network relations. Nadel's conceptualization of networks as relational social structures was widely adopted by social network theorists and researchers over the ensuing decades of development. For example, Harrison White and his colleagues defined social structure as "regularities in the patterns of relations among concrete entities; it is not a harmony among abstract norms and values or a classification of concrete entities by their attributes" (White, Boorman, & Breiger, 1976, pp. 733–734). More recently, the core mechanisms in Crossley and Krinsky's (2016) relational approach to sociology are interactions, relations, and networks. In network analyses, the entities may be individual natural persons, small groups, organizations, or even nation-states. Some types of network entities lack agency, such as documents posted on websites and participatory events such as sports matches and social movement protests. The patterns of relations connecting members of one or more sets of entities comprise the macrosocial contexts, or overall relational structures, that influence actor perceptions, attitudes, beliefs, decisions, and actions. The primary objectives of network analysis are to measure and represent these structural relations accurately and to explain both why they occur and what their consequences are.

Social network analysis rests on three underlying assumptions about structural relations and their consequences. First, structural relations are often more important for understanding observed behaviors than are such characteristics as race, gender, age, socioeconomic status, and political ideology. For example, research on voting behavior and social movement participation found that egocentric network structures more strongly influence people's choices than respondent attributes (Diani, 2004; Huckfeldt & Sprague, 1987; Knoke, 1990). Many actor attributes remain unaltered across the numerous social settings in which they participate (a woman's age, race, and education remain unchanged whether at home, at work, and at church). In contrast, many structural relations occur only at specific time-and-place locales and either vanish or are suspended when participants are elsewhere (e.g., student-teacher and doctor-patient relations do not exist outside school and clinic settings, respectively). A man holding a menial factory job requiring little initiative may be the dynamic leader of his church and an enthusiastic softball team player. Such behavioral differences are difficult to reconcile with unaltering gender, age, and status attributes but comprehensible on recognizing that people's structural relations can vary markedly across social contexts within which they are embedded. The structural-relational explanations favored by network analysts depart markedly from substantialist approaches premised on static "thing-concepts" as their primary units of analysis: essences, self-action, norm-based conformity, rational choice, and variable-centric and social identity approaches (Emirbayer, 1997). In assuming that patterned relations influence social entities apart from their attributes, network analysis offers distinctive theoretical and empirical explanations of the origins of social action.

Second, social networks affect actor perceptions, beliefs, and actions through diverse structural mechanisms that are socially constructed by relations among entities. Direct contacts and more-intensive interactions dispose people and organizations to be better informed, more aware, and more susceptible to influencing or being influenced by others. Indirect relations through intermediaries (in popular imagery, agents who broker connections for their clients) also bring exposure to new ideas and potential access to useful resources that may be obtained through exchanges with others. For example, in a classic network study by Mark Granovetter (1973), job seekers typically obtained less useful information from their intimate circles, whose members already shared and circulated the same intelligence, than from their weaker and more distant social contacts. Relational structures provide complex pathways for assisting or hindering flows of knowledge, gossip, and rumor through a population (Fang, McAllister, & Duffy, 2017). A variety of structural-relational factors explains racial differences in the spread of HIV/AIDS infections among young men who have sex with men (Mustanski, Birkett, Kuhns, Latkin, & Muth, 2015) and the propagation of financial distress through the international banking network during the global financial crisis of the aughts (Kojaku, Cimini, Caldarelli, & Masuda, 2018). Physical illness, mental health, and recovery from substance abuse are strongly affected by people's social support networks (Cullen, Mojtabai, Bordbar, Everett, Nugent, & Eaton, 2017; Stevens, Jason, Ram, & Light, 2015), with social media exerting some unusual impacts (Lu & Hampton, 2017; Pallotti, Tubaro, Casilli, & Valente, 2018). Structural relations are vital to building cohesion and solidarity within a group but may also reinforce prejudices and intensify conflict with outgroups (Bliuc, Faulkner, Jakubowicz, & McGarty, 2018; Roversi, 2017). Competitive and cooperative relations enable innovation in corporate supply chains (Delgado-Márquez, Hurtado-Torres, Pedauga, & Cordón-Pozo, 2018), mobilization for collective action by social movements (Diani, 2016), and the operation of "dark networks" for drug trafficking, immigrant smuggling, and terrorist campaigns (Wu & Knoke, 2017). By channeling information, money, and other types of resources to particular structural locations, networks help to create interests and shared identities and to promote shared norms and values. Network analysts seek to uncover the mechanisms through which social relations affect social entities and to identify the contingent conditions under which particular mechanisms operate in specific empirical contexts.

The third underlying assumption of network analysis is that structural relations should be viewed as dynamic processes. This principle recognizes that networks are not static structures but are continually changing through interactions among people, groups, or organizations. In applying their knowledge about networks to leverage advantages, network entities also transform those structural relations, both intentionally and unintentionally. For instance, in an intervention experiment to reduce conflict and bullying among students in 56 schools, experimenters comprehensively measured every school's networks, then randomly selected "seed groups" of 20 to 32 students to be encouraged to take public stands against conflict (Paluck, Shepherd, & Aronow, 2016). Disciplinary reports of conflict fell by 30% in the treatment schools compared to control-group schools, but the effect was stronger for seed groups containing more students who attracted greater student attention. Apparently, those popular students changed their network peers' beliefs and behaviors by publicly stigmatizing conflict and bullying as less socially normative. Such dynamics exemplify the more general "micro-to-macro problem" in the theory of social action (Coleman, 1986). The core issue is how large-scale systemic transformations emerge out of the combined preferences and purposive actions of individuals. Because network analysis simultaneously encompasses both structures and entities, it provides conceptual and methodological tools for linking changes in actors' microlevel choices to macrolevel structural alterations. The increased availability of longitudinal datasets, especially large online networks, coupled with methodological developments for analyzing multilevel relations, are accelerating research on cross-level dynamic processes (Lazega & Snijders, 2015; Snijders, Steglich, & Schweinberger, 2017). Likewise, developments in temporal exponential random graph models (TERGMs) and stochastic actor-oriented models (SAOMs), such as SIENA, hold great promise to advance our understanding of network dynamics (Leifeld & Cranmer, 2019; Leifeld, Cranmer, & Desmarais, 2018).

2.2 Entities and Relations

The two indispensable elements of any social network are entities and relations. Their combination jointly constitutes a social network, as described in the next subsection. *Entities* may be individual natural persons or collective actors such as informal groups and formal organizations. Common examples of individual actors include children on a playground, high school students attending a prom, employees in a corporate work team, staff and residents of a nursing home, and terrorists operating in a covert cell. Collective actors might be firms competing in an industry, voluntary associations raising funds for charities, political parties holding seats in a parliament, and nations signing a military alliance. Other types of entities lack human agency, such as bills debated in a legislature, dances attended by students, and books read by library patrons. Sometimes networks are comprised of diverse types of entities, such as a healthcare system consisting of doctors and nurses, patients, clinics, hospitals, laboratories, insurance companies, and governmental regulations.

A *relation* is generally defined as a specific kind of contact, connection, or tie between a pair of entities, or *dyad.* Relations may be either *directed,* where one actor initiates and the second actor receives (e.g., advising, selling), or *undirected,* where mutuality occurs (e.g., conversing, collaborating). A relation is not an attribute of one entity but is a joint dyadic property that exists only so long as both participants maintain their association. An enormous variety of relations among individual and collective entities may be relevant to representing network structures and explaining their effects. At the interpersonal level, children befriend, play with, fight with, and confide in one another. Employees work together, discuss, advise, trust, undermine, and betray. Among collectivities, corporations exchange goods and services, communicate, compete, sue, lobby, and collaborate. In healthcare systems, physicians refer patients to specialty clinics, pharmacies, laboratories, hospitals, imaging centers, nursing homes, and hospices. Which specific type of relation a network researcher should measure depends on the particular objectives of the research project. For example, an investigation of community networks will likely examine various neighboring activities, whereas a study of banking networks would investigate financial transactions. Of course, some analyses scrutinize multiple types of relations, such as the political, social, and economic ties among corporate boards of directors. We present a general classification of relational contents in the next subsection.

Social science researchers rely heavily on measuring and analyzing the attributes of individual or collective units of analysis, whether through survey, archival, or experimental data collection. Although attributes and relations are conceptually distinct approaches to investigating social behavior, they should not be viewed as mutually exclusive options. Instead, many entity attributes can be reconceptualized as relations connecting dyads. For example, a nation's annual volumes of exports and imports are characteristics of its economy. But, the amount of goods and services exported and imported between all pairs of nations represents the structure of trading networks in the global economy. Patents awarded to scientists employed at high-tech firms indicate companies' research innovations, but patent-citation networks reveal how knowledge flows through industries (Zhang, Kong, Zheng, Wan, Wang, Hu, & Shao, 2016). The number of friends indicates a child's popularity, but only network analyses of all dyadic friendship choices can uncover important cliques and clusters. Relations reflect emergent dimensions of complex social systems that cannot be captured by simply displaying a variable's distribution or averaging its members' attributes. Structural relations potentially influence both individual behaviors and systemic outcomes in ways not reducible to entity characteristics. For example, efforts to control sexually transmitted infections among injection drug users and sex workers require knowledge of both social and geographic distances among street people. Researchers identified 101 "hot-spots" of high-risk activities in Winnipeg, Canada, where "the combination of spatial and social entities in network analysis defines the overlap of vulnerable populations in risk space, over and above the person to person links" (Logan, Jolly, & Blanford, 2016). An experiment in a large environmental nongovernmental organization found that "boundary spanners"—individuals who cross internal boundaries, such as departmental or geographic location, via their informal social networks—were more likely to diffuse innovations, although positions in a formal organizational hierarchy mediated this activity (Masuda, Liu, Reddy, Frank, Buford, Fisher, & Montambault, 2018). The strong inference is that exclusively focusing on actor attributes loses many important explanatory insights provided by network perspectives on social behavior.

2.3 Networks

A *social network* is a structure composed of a set of entities, some of whose members are connected by a set of one or more relations. These two fundamental components are common to most network definitions; for example: "a network contains a set of objects (in mathematical terms, *nodes*) and a mapping or description of relations between the objects or nodes" (Kadushin, 2012, p. 14). Different types of relations identify different networks, even where observations are restricted to the same set of entities. Thus, the friendship network among a set of office employees very likely differs from their advice-seeking network. Stating that connections exist among members of a network does not require that all members have direct relations with all others; indeed, sometimes very few dyads have direct links. Rather, network analysis considers both present and absent ties and possibly also variation in the intensities or strengths of the relations. A configuration of empirical relations among entities identifies a specific *network structure,* the pattern or form of that network. Structures can vary dramatically in form, ranging from isolated structures where no actors are connected to saturated structures in which everyone is directly connected. More typically, real networks exhibit intermediate structures in which some entities have more numerous connections than others. A core problem in network analysis is to explain the occurrence of different structures and, at the entity or nodal level, to account for variation in linkages among entities. The parallel empirical task in network research is to detect and represent structures accurately using relational data.

The first researcher credited with using the term *social network* was John A. Barnes (1954), an anthropologist who studied the connections among people living in a Norwegian island parish. Barnes viewed social interactions as a "set of points some of which are joined by lines" to form a "total network" of relations (Barnes, 1954, p. 43). The informal set of interpersonal relations composed a "partial network" within this totality. Barnes drew on the work of Jacob Moreno (1934), whose hand-drawn *sociograms* of lines and labeled points displayed children's likes and dislikes of their classmates. We discuss methods for representing networks visually as graphs and mathematically as matrices in Chapter 4. From anthropology and sociology, network ideas and methods diffused over the past half century to many disciplines, which adapted them to prevailing theories and problems. For historical overviews of the origins and diffusion of network principles, see Freeman (2004, 2011); Knox, Savage, and Harvey (2006); Kadushin (2012); and Scott (2017).

If network analysis were merely a conceptual framework for describing how a set of actors is linked together, it would not have excited so much interest and effort among social researchers. But, as an integrated set of theoretical concepts and analytic methods, social network analysis offers more than accurate representations. It proposes that, because network structures affect actions at both the individual and systemic levels of analysis, network analysis can explain variation in structural relations and their consequences. J. Clyde Mitchell's (1969, p. 2) definition of social networks emphasized their impacts on outcomes: "a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of the persons involved." The first edition of this book underscored this perspective: "The structure of relations among actors and the location of individual actors in the network have important behavioral, perceptual, and attitudinal consequences for the individual units and for the system as a whole" (Knoke & Kuklinksi, 1982, p. 13). Similarly, Barry Wellman

(1999, p. 94) wrote, “Social network analysts work at describing underlying patterns of social structure, explaining the impact of such patterns on behavior and attitudes.”

2.4 Research Design Elements

Three elements of network research design shape the measurement and analysis strategies available to researchers: social settings, relational form and content, and level of data analysis. Every network data collection project must involve making explicit choices about these elements before beginning fieldwork. Varying combinations of them generate the wide range of social network investigations published in the research literatures of numerous disciplines.

Social Settings. The first steps in designing a network study are to choose the most relevant social setting and to decide which entities in that setting comprise the network entities. Ordered on a roughly increasing scale of size and complexity, a half-dozen basic units from which samples may be drawn include individual persons, groups (both formal and informal), complex formal organizations, classes and strata, communities, and nation-states. Some two-stage research designs involve a higher-level system within which lower-level entities comprise the actors. Common examples are hierarchical social settings such as corporations with employees, schools with pupils, hospitals with physicians, municipal agencies with civil servants, and universities with colleges with departments with professors.

The earliest and still most common network projects select small-scale social settings—classrooms, offices, factories, gangs, social clubs, schools, villages, artificially created laboratory groups—and treat their individual members as the actors whose relations comprise the networks for investigations. Recent examples include bullying and homophobic teasing among middle school students (Merrin, De La Haye, Espelage, Ewing, Tucker, Hoover, & Green, 2018), helping and gossip networks among employees of a Turkish retail clothing company (Erdogan, Bauer, & Walter, 2015), and the effects of ethnic diversity on the spread of word-of-mouth information in two matched rural Ugandan villages (Larson & Lewis, 2017). Small settings have considerable advantages in sharply delineated membership boundaries, completely identified populations, and usually researcher access by permission from a top authority. However, network analysis concepts and methods are readily applied to larger-scale formations, many of which have porous and fuzzy boundaries, including clandestine networks. Examples include peer network origins of adolescent dating behavior (Kreager, Molloy, Moody, & Feinberg, 2016), criminal organizations in communities of Calabria, Italy (Calderoni, Brunetto, & Piccardi, 2017), and strategic alliances among multinational corporations in the Global Information Sector (Knoke, 2009).

Relational Form and Content. Network researchers must decide on which particular relations to collect data. Relations among pairs of social actors have both form and content, a dichotomy that Georg Simmel (1908) proposed in his classic analyses of association. The two elements are empirically inseparable and only analytically distinguishable. *Contents* are the interests, purposes, drives, or motives of individuals in an interaction, whereas *forms* are modes of interaction through which specific contents attain social reality. Simmel argued that the task of sociology is to identify a limited number of forms—sociability, superiority, subordination, competition, conflict, cooperation, solidarity—that occur across a wide range of concrete settings, social institutions, and historical contexts. A particular form can vary greatly in content. For example, the basic forms of superordination and subordination are ever present in government, military, business, religious, athletic, and cultural institutions. Conversely, diverse contents like economic interests and drives for power are manifested through forms of competition and cooperation.

The form-content dichotomy also applies to social network analysis. *Relational form* is a property of relations that exists independently of any specific contents. Two fundamental relational forms are (a) the intensity, frequency, or strength of interaction between pairs of entities and (b) the direction of relations between both dyad members—null, asymmetric, or mutual choices. *Relational content* refers to its ‘‘substance as reason for occurring’’ (Burt, 1983, p. 36). Substantive content is an analytic construct designed by a researcher to capture the meanings of a relation from the informants’ subjective viewpoints. When people are asked, ‘‘please identify your close friends, friends, and acquaintances” in some social setting, the intended relational content is “friendship.” The results of this query depend on how each actor first conceptualizes the meanings of the three proffered response categories and then classifies the other actors according to recollections of diverse interpersonal interactions. Obviously, people may vary markedly in their interpretations of both the friendship labels and those activities that they consider to be indicators of greater or lesser intimacy. Friendship dyads are never precisely reciprocated and the level of intimacy may be very unequal; for example, one dyad member considers the second person a “best friend,” but the second member views the first person as a “friend.” The National Study of Adolescent Health (Add Health) found that girls and Asian Americans were most likely to have reciprocated friendships, whereas interracial friendships were much less common than friendships between students of the same race (Vaquera & Kao, 2008).

The choice of relational content, also called *type of tie,* is largely determined by a project's theoretical concerns and research objectives. A study of healthcare networks could inquire into people's interpersonal sources of trusted information and advice about health-related matters, whereas a project on political networks might ask them to identify others with whom they discussed or participated in political affairs. Some substantive problems imply that more than one analytically distinct relational content should be investigated, in which case measuring and simultaneously analyzing two or more types of ties (i.e., *multiplex networks*) is an appropriate strategy. For example, psychologists asked 132 undergraduates at Midwestern University to list their Facebook friends who fulfilled each of five social functions (i.e., types of ties): sharing social activities, discussing personal matters, providing instrumental support, providing emotional support, and sharing success and happy events (Gillath, Karantzas, & Selcuk, 2017). Students with higher attachment avoidance were likely to ascribe fewer multiplex social roles to their networks' members, implying a lower degree of social trust.

Inexplicably, network analysts have conducted little research on the connections among diverse domains of relational contents. Ronald Burt (1983) examined survey respondents' perceptions of relational contents and uncovered substantial confusion, redundancy, and substitutability among the 33 questions posed to a sample of Northern Californians. He concluded that just five key questions would suffice to recover the principal structure of relational contents in the friendship, acquaintance, work, kinship, and intimacy domains. However, we still need much more research on the similarities and differences of meanings that people attach to commonly used relational terms and labels in a wide variety of network settings. A cognitive map of the structural connections among relational content domains would enable researchers efficiently and accurately to select specific contents most relevant to their theoretical and substantive concerns.

Until that desideratum arrives, in the spirit of Simmel we propose a small typology of generic contents:

- Transaction relations: Entities exchange control over physical or symbolic media, for example, in gift giving or economic sales and purchases.
- Communication relations: Linkages between entities are channels through which messages may be transmitted.
- Boundary penetration relations: Ties consist of membership in two or more social formations, for example, voluntary associations or social movement organizations.

- Instrumental relations: Actors contact one another in efforts to obtain valued goods, services, or information, such as a job, an abortion, political favors, or religious salvation.
- Sentiment relations: Perhaps the most frequently investigated networks involve actors expressing their feelings of affection, admiration, deference, loathing, or hostility toward one another.
- Authority/power relations: These types of ties, usually occurring in formal hierarchical organizations, indicate the rights and obligations of position holders to issue and obey commands.
- Kinship and descent relations: These bonds of blood and marriage reflect relations among family roles.

Levels of Analysis. After deciding the social setting and the relational forms and contents, researchers have several alternative levels at which to analyze the structures in data that they collect for social network projects. Details of appropriate measures and methods appear in Chapters 3 through 5, but here we summarize four conceptually distinct levels of analysis that analysts could investigate.

The simplest level is the *egocentric* network, consisting of one actor (*ego*) and all other actors (*alters*) with which ego has direct relations as well as the direct relations among those alters. This set is also called ego's "first zone," in contrast to second and higher zones consisting of all the alters of ego's alters, and so on. If a network's size is N actors, an egocentric analysis would have N units of analysis. Each ego actor can, in turn, be described by the number, intensity, and other characteristics of its linkages with its set of alters, for example, the proportion of reciprocated relations or the density of ties among its alters. An egocentric analysis of incarcerated California youths indicated that respondents reporting no close friendships within the facility had lower postinterview misconduct than those who nominated peers, suggesting an influence or amplifying effect of friends on misbehavior (Reid, 2017). In some respects, egocentric analysis resembles typical attribute-based survey research, with a respondent's individual characteristics such as gender, age, and education supplemented by measures derived from that person's direct network relations. Egocentric network research designs are well suited to surveys of respondents who are unlikely to have any contact with one another. The 1985 General Social Survey of the adult U.S. population (Marsden, 1987) pioneered procedures for identifying and eliciting information about a respondent's alters, which we describe in some detail in Chapter 3.

A second level of analysis is the *dyadic network,* consisting of pairs of actors. If the order of a pair is irrelevant—as in marital status where persons are either unmarried, cohabiting, married, separated, or divorced—a sample of N actors has $(N^2 - N)/2$ dyadic units of analysis. But, if the direction of a relation matters, as in giving orders and taking advice, then the sample contains $(N^2 - N)$ ordered dyads. The most basic questions about a dyad are whether a specific type of tie exists between two actors, and, if so, what is the intensity, duration, or strength of that relation? A closely related issue is whether a dyad without a direct tie is nevertheless indirectly connected via ties to intermediaries (e.g., brokers, go-betweens). Typical analyses seek to explain variation in dyadic relations as a function of pair characteristics, for example, the homophily hypothesis that "birds of a feather flock together" or the complementarity hypothesis that "opposites attract." Dyadic empathy—"a combination of perspective taking and empathic concerns for one's romantic partner"—is associated with higher sexual satisfaction, relationship adjustment, and sexual desire of first-time parents (Rosen, Mooney, & Muise, 2017, p. 543).

A third level of network analysis is, unsurprisingly, *triadic relations*. A set of N actors has $\left(\frac{N}{3}\right)$ triples, the number of ways to take N actors, three at a time. All possible combinations of present and absent directed binary relations among the actors in a triple generates a set of 16 distinct triad types. A basic descriptive question for empirical network analysis regards the distribution of observed triads among the 16 types, a summary tabulation called the *triad census.* Substantive research on triadic structures concentrated on sentiment ties (liking, friendship, antagonism), with particular interest in balanced and transitive triadic relations (e.g., if A chooses B and B chooses C, does A tend to choose C?). Because we lack space to review triad analysis methods, interested readers should consult the research program of James Davis, Paul Holland, and Samuel Leinhardt (Davis, 1979) and a comprehensive treatment by Wasserman and Faust (1994, pp. 556–602) for details.

Beyond the three microlevels, the *whole network* (also called *complete network*) is the most important macrolevel of analysis. Researchers use the information about every relation among all N actors to represent and explain an entire network's structural relations. Typical concerns are the presence of distinct positions or social roles within the system that are jointly occupied by the network actors and the pattern of ties within and among those positions. Although a whole network has N actors and $(N^2 - N)$ dyads (assuming directed relations and self-relations are generally ignored), these elements add up only to a single system. Examining the causes or consequences of structural variation at the whole network level of analysis typically involves measures of the global structural properties.

An example is a Dutch online social network of more than 10 million users living in 438 municipalities (Norbutas & Corten, 2018). Communities with higher network diversity were more economically prosperous than less-diverse communities, whereas greater network density at the community level was negatively associated with prosperity.

The four levels of network analysis imply that emergent phenomena at one level cannot be simply deduced from knowledge of the relations at other levels. For example, transitivity of choice relations is a substantively important variable for theories of friendship formation ("a friend of my friend is my friend"), which can be observed at the triadic level but not at the egocentric or dyadic level. For another illustration, Mark Newman (2001) found that coauthorship networks in biomedical research, physics, and computer science were each structured as "small worlds," where only five or six steps were necessary to connect random pairs of scientists. However, biomedical research was dominated by many people with few coauthors, in contrast to other disciplines characterized by a few people with many collaborators (see also, e.g., Ebadi & Schiffauerova, 2016; Maggioni, Breschi, & Panzarasa, 2013). The adaptability of network principles and procedures to investigate structural relations across multiple levels of analysis underlies its bourgeoning popularity for theorizing about social action and guiding empirical research.

Chapter 3

DATA COLLECTION

Before collecting data about a network, researchers must first answer three important questions: where to set the boundary, how to sample network entities, and which relations to measure? Because deciding the set of nodes to include is an obvious starting point for any network project, we begin with boundary specification issues.

3.1. Boundary Specification

The boundary specification question asks: Where does an investigator set limits when collecting data on social relations that, in reality, may have no clearly demarcated limits (Barnes, 1979, p. 414)? In the seminal article on network bounding, Laumann, Marsden, and Prensky (1983, p. 19) framed their answer around contrasting realist and nominalist strategies for specifying "the inclusion rules in defining the membership of actors in particular networks and in identifying the types of social relationships to be analyzed." They subsequently expanded the dichotomy to three generic approaches to identifying network boundaries: positional, relational, and event-based (Laumann, Marsden, & Prensky, 1989). A recent assessment concluded that each strategy has strengths and limitations about which network researchers should be better informed when deciding where to draw the boundary (Nowell, Velez, Hano, Sudweeks, Albrecht, & Steelman, 2018). This section highlights various aspects of alternative boundary specification procedures.

Realist and Nominalist Strategies. In the *realist strategy* for boundary specification, a researcher attempts to capture the subjective perceptions of network actors, defining boundaries as the limits that are consciously experienced by all or most participants (e.g., members of a family, congregation, or social movement). Actors and their relations are included or excluded to the extent that the other actors judge them to be relevant. For example, to identify the core organizations of the U.S. energy and health national policy domains, Knoke and Laumann (1982, p. 256) took the realist approach: "a policy domain is a subsystem identified by specifying a substantively defined criterion of mutual relevance or common orientation among a set of consequential actors concerned with formulating, advocating, and selecting courses of action (i.e., policy options) that are intended to resolve the delimited substantive problems in questions." They excluded any organizations that domain informants did not perceive as influential in national energy or health policy making.

To use the *nominalist strategy,* an investigator imposes an a priori conceptual framework serving an analytic or theoretical purpose for a particular project. Legal or other formal membership requirements typically draw clear distinctions between persons inside and outside an organization. Often a membership roster or list of employees identifies current organizational participants. For example, research on the diffusion of physical activity guides for patients with spinal cord injuries investigated information-sharing relations among 78 staff within a Canadian community-based organization (Gainforth, Latimer-Cheung, Athanasopoulos, Moore, & Ginis, 2014). Another example of a nominalist approach bounded a 2012 Czech criminal network as the 11 persons "being charged" in the Rath Affair of bribery, kickbacks, abuse of European Union subsidies, and public contract manipulation (Diviák, Dijkstra, & Snijders, 2018). The degree to which both the subjective-perception and analytic-imposition nominalist strategies produce equivalent network boundaries is, of course, always an empirical question.

Positional Strategies. This strategy uses actor attributes, membership in formal organizations, or occupancy of a well-defined position as a criterion for inclusion in a network. Heemskerk, Fennema, and Carroll's (2016) research on the changing global corporate elite exemplified the positional approach. Defining the elite as persons who sat on the boards of the 176 largest global corporations in 1976, 1996, 2006, and 2013, they identified 10,164 board positions. Over time, the number of interlocks (directors with multiple board seats) showed "a steady trend of decline" (p. 74), and the network became less hierarchical and more transnational. Feldman-Savelsberg, Ndonko, and Yang (2005) investigated collective memory choices among the members belonging to six Cameroon women's hometown associations in the city of Bamileke, that is, women originating from the same village or chiefdom. For a study of environmental sustainability and corporate social responsibility in 49 large healthcare organizations, Senay and Landrigan (2018) required inclusion on one of five lists, including Fortune 500, Standard & Poor (S&P) 500, and Becker's Hospital Review. Analysts who use a positional strategy may discover that an organization's membership list or roster is outdated, incomplete, or otherwise inaccurate. They may need to conduct their own census to assemble a complete list of network participants.

Positional strategies commonly generate a list of actors occupying similar positions in a formal social structure, even when those actors lack direct ties to one another. Researchers should be acutely sensitive to how representative are network structures uncovered using positional criteria. Thus, the connections among business elites differ substantially from ties among lower-echelon employees. Strategic alliances formed between multinational corporations may be quite dissimilar from partnerships of small firms. Another issue emerging from networks bounded using the positional strategy is that the actors are often disconnected, comprising many small, densely connected positions lacking ties to one another. For example, the Cameroon women belonging to the same hometown association were often complete strangers (Feldman-Salvesberg, Ndonko, & Yang, 2005).

Network analysts using a positional approach should provide explicit justification for including or excluding particular positions (Laumann et al., 1983). Researchers may apply nominalist criteria, setting an arbitrary threshold for their inclusion rules even where positions vary continuously. For example, to study very large business firms, researchers might restrict the population to 50, 100, or 500 firms from the Fortune 1000 companies list. Where to draw a boundary may depend more on time and budget constraints than on some "natural" division between the included and excluded actors. Galaskiewicz's (1979) selection of organizations in the small city of Towertown vividly illustrated this process. First, he applied a territorial criterion that restricted the population to a geographic area. Then, an industry criterion excluded commercial establishments, transportation facilities, public utilities, real estate, block clubs, community organizations, and elementary schools because of time and budgetary constraints.

Relational Strategies. This approach relies on knowledgeable informants or the network actors themselves to nominate additional actors for inclusion. Relational approaches embrace several procedures, including the reputational method, snowball sampling, fixed-list selection, expanding selection, and k-core methods. This subsection briefly describes the requirements and limitations of these relational strategies.

In the *reputational method,* researchers ask the most knowledgeable informants or experts to nominate a set of actors for their study. For example, Michael Heaney (2014) investigated interest group influence in the U.S. health policy domain by compiling a preliminary list of organizations that had lobbied the federal government, testified at Congressional hearings, or appeared in a previous network project. The list was then "circulated to a panel of experts from academia and the policy world to solicit additional recommendations. Any interest group recommended by at least two experts was included in the study." These procedures identified 171 "most active" organizations, which Heaney contacted for interviews about their network ties and political influence reputations.

Reputational methods rely heavily on key informants to provide accurate and complete information, which raises concerns about the researcher's ability to locate informants capable of enumerating all important network players (Scott, 1991, p. 59). Morrissey, Tausig, and Lindsey (1985, p. 35) cautioned, “there are no standards by which the accuracy of this boundary-drawing [reputational method] criterion can be assessed. It is conceivable that different criteria could result in different actors being included and the subsequent analyses affected.” Assessing the reliability and completeness of network enumeration using reputational methods is a formidable task. Often an assessment is possible only after data collection is completed. Therefore, network researchers should always justify their choice of key informants with strong theoretical and empirical reasons that are independent of the particular social relations under investigation (Scott, 1991, p. 59). But sometimes the problem of *sampling bias*—defined as a sample that is not representative of its target population due to not-at-random missing cases (Allison, 2001, pp. 78–81)—goes beyond the issue of identifying suitable informants. Key informants, no matter how knowledgeable, always produce data systematically different from those collected through snowball sampling.

Snowball sampling begins with an initial convenience sample of network actors (“seeds”) who are asked to nominate others with whom they have a specified relation. In turn, those actors recruit another wave of network participants, so the sample expands wave by wave like a snowball growing as it rolls downhill. The process continues until few or no additional names surface (Frank, 2005; Wasserman & Faust, 1994, p. 34). Recent projects used the method to identify the number of geriatric emergency departments in the United States (Hogan, Olade, & Carpenter, 2014), to locate cannabis users in two Spanish provinces (Brañas, Barrigón, Garrido-Torres, Perona-Garcelán, Rodriguez-Testal, Lahera, & Ruiz-Veguilla, 2016), and to study the risk of fatal and nonfatal gunshot injuries in Boston’s Cape Verdean community (Papachristos, Braga, & Hureau, 2012). In an earlier version, each wave of a snowball sample rigidly generated the same number of actors (e.g., “name your three best friends”), and every actor at every stage was asked the identical questionnaire item (Goodman, 1961). More recently, these two conditions typically are relaxed to elicit differing numbers of nominated actors using different questionnaire items (Wasserman & Faust, 1994, p. 34). Because snowball sampling uses network actors’ social relations to construct the sample, each round of nominations typically uncovers new participants who have relations with the extant actors. Thus, snowball sampling usually generates strongly connected social networks (Laumann et al., 1983) and is also called the *chain-referral method* (Heckathorn, 1997). A complex extension is *respondent-driven sampling,* which combines snowball methods with a mathematical model that weights the sample to compensate for its nonrandom collection (Heckathorn & Cameron, 2017).

Snowball sampling is particularly powerful for finding members of a hard-to-reach population for social network analysis, such as drug dealers and users, illegal immigrants, HIV-positive sex workers, and violent white supremacists. Because sampling frames for such hidden populations don't exist, standard survey sampling methods often yield insufficient numbers of respondents. In contrast, snowball sampling starts by interviewing a handful of informants, for example, illegal-drug users in a needle-exchange program or sellers in an open-air drug market. These seeds are asked to give locational information about other actors with whom they have the specified relationship, for example, the sellers' regular customers. In turn, these nominees are contacted and asked to name additional participants. Because network data collection usually requires knowing the identities of egos' alters, obtaining informed consent, protecting anonymity, assuring confidentiality, and securing stored data raise serious ethical concerns (Curtis 2014). Human subjects committees may disapprove snowball sampling designs in which researchers recruit alters directly. As an alternative, respondent-driven sampling designs typically ask the seed egos to give coupons to their alters describing a small reward, such as a $20 gift card, that they would receive after voluntarily participating in the project (Scheim & Bauer, 2015).

A procedural issue in implementing network sampling—whether to provide informants with a list of names or to permit respondents to generate their own nominees—distinguishes fixed-list selection and expanding selection (Doreian & Woodard, 1992, 1994). In *fixed-list selection,* respondents can only report their ties with a set of alters identified a priori by the researchers and their informants. In *expanding selection*, respondents identify as many actors as they wish without referring to a list of names. The implementation of expanding selection closely resembles snowball sampling procedures. In Doreian and Woodard's (1992) project to identify child and adolescent service system program agencies, any organization that was added to the network had to receive at least three nominations from the directors or five nominations from the staff. Doreian and Woodard reported that fixed-list selection and expanding selection yielded radically different networks on several dimensions, including numbers of actors, numbers of ties, density, and quality of ties. Fixed-list selection generated only 50% of the organizations, and 40% of the dyadic ties, uncovered by expanding selection procedures. The fixed-list approach is more prone to nonrandom sampling bias; that is, it always produces a core set of actors and systematically excludes peripheral actors. This method yields an inferior result, a network without structural context. Unless conditions guarantee that both methods produce equivalent networks, fixed-list selection, despite its low administrative cost, should not be used as a surrogate for expanding selection.

The *k-core method* finds subsets of actors within a large network that typically contains many subgroups weakly connected to one another but densely connected within the subsets (Seidman, 1983; Yang & Hexmoor, 2004). A subset is a k-core if every actor has ties with at least *k* other actors in the subgroup. By changing the value of *k,* a researcher can set more or less restrictive criteria for bounding a network. Doreian and Woodard (1994) applied expanding selection to demonstrate how the k-core concept could be used to define and locate network boundaries. By changing the *k threshold*, researchers can redraw the boundaries of a very large and sparsely connected network to make it either more restrictive (high k) or less restrictive (low k). For practical purposes, Doreian and Woodward (1994) recommended using a low value of k to establish the overall network boundary. A more inclusive network is less susceptible to selection bias, whereas analysts can always subsequently apply a higher k to create a more restrictive network. However, a more inclusive network imposes a greater data collection burden: Although a low k threshold produces a more inclusive network, researchers must interview large numbers of respondents at each nomination round, a costly and error-prone task. Lowering the k threshold also generates an exponential, rather than linear, increase in total number of network nodes. Recent research used k-core methods to identify influential spreaders in online social networks (Al-garadi, Varfathan, & Ravana, 2017), to locate core areas in many science fields (Liu, Tang, Zhou, & Do 2015), and to find power elites (Larsen & Ellersgaard, 2017).

Event-Based Strategies. This method draws a network's boundary around actors who participate in specified types of activity occurring at particular times in real or virtual locations. The classic Southern Women network, 18 women who attended 14 informal gatherings and civic events in the 1930s, continues to be a much-reanalyzed benchmark (e.g., Alzahrani & Horadam, 2016). A daunting task for event-based researchers is to provide a sufficient rationale for identifying and selecting important events capable of answering a specific research question. Researchers might include events that are either noteworthy to a neutral observer or identified as important by knowledgeable network participants. The event-based boundary method is particularly vulnerable to incomplete or missing data from failure to include some crucial activities and the actors who attend them. This problem is particularly acute for researchers who rely on a single event to locate network boundaries because many important players may fail to attend that event. Hence, observing multiple events would normally produce a more comprehensive network. Because each event involves a potentially unique subset of network actors, a multievent approach produces malleable boundaries: Every event yields a distinct network whose participants only partially overlap with those attending other events. Aggregating participants across all events should yield a more inclusive network that is better able to answer the research questions. For example, to study student interactions in a college dormitory, Freeman and Webster (1994) observed participants at events occurring in two visible settings, a cafeteria and dorm social meetings. A complete network of children's play activities requires observational or self-reported data collected at school playgrounds, homes, and out-of-school clubs (King & Howard, 2014).

Smartphone-driven event-based social networks (EBSNs) and location-based social networks (LBSNs)—such as Foursquare, Google Local, and Meetup—attracted the attention of computer scientists who designed algorithms that try to predict who will participate in which events (Du, Yu, Mei, Wang, Wang, & Guo, 2014; Frith, 2014; Li, Westerholt, Fan, & Zipf, 2018; Zhang, Zhao, & Cao, 2015). The sheer volume of alternative EBSNs "often undermines users' ability to choose the events that best fit their interests" (Macedo, Marinho, & Santos, 2015). To cope with overload, recommender software tools create lists of books, merchandise, restaurants, hotels, movies, concerts, and other activities personally tailored to users' interests and prior choices (Ricci, Rokach, & Shapira, 2015). Could increasing integration of personalized EBSN, LBSN, and recommender systems be social networker paradise (Purushotham & Kuo, 2016), and the death knell of privacy?

3.2. Data Collection Procedures

We discuss a variety of data collection methods for producing network data, including single- and multiple-name generator procedures, measures of total personal networks, position and resource generators, and archival documents.

Single- and Multiple-Name Generators. Name generators are prevalent in research on egocentric networks, typically using survey questionnaires to collect information from an ego respondent about relations among a set of alters with whom ego has direct contact (Marsden, 1987). Egocentric network research requires two survey instruments: a *name generator* to identify all alters and a *name interpreter* to obtain information from ego about each alter and network ties among them and with ego (Marsden, 2005). First, ego is asked to name persons with whom she or he has a specified type of relation, such as friendship or political discussion. Next, ego is asked whether each pair of alters also has that relation and how strong or frequent is that tie. Finally, ego also provides information about every alter's attributes, such as age, sex, race, and education. The reliabilities of ego's reports are unknown because the alters are not interviewed to obtain their self-reports.

Egocentric network studies may use single- or multiple-name generators, depending on a project's objectives. A single-name generator relies on one questionnaire item to elicit the alters' names. The pioneering 1985 General Social Survey (GSS) module on core discussion groups of Americans exemplified the single-name generator design (Burt, 1985; Marsden, 1987). An interviewer first asked a respondent:

> From time to time, most people discuss important matters with other people. Looking back over the last 6 months—who are the people with whom you discussed matters important to you? Just tell me their first names or initials.

The interviewer recorded as many as six names, then asked, "Do you feel equally close to all these people? (IF NO): Which of these people do you feel especially close to? (PROBE: Anyone else?)" They next asked about ties among all pairs of alters:

> Please think about the relations between the people you just mentioned. Some of them may be total strangers in the sense that they wouldn't recognize each other if they bumped into each other on the street. Others may be especially close, as close or closer to each other as they are to you. First, think about NAME1 and NAME2. Are they total strangers? Are they especially close? (PROBE: As close or closer to each other as they are to you?)

Next, the interviewer asked how long the respondent had known each alter, how often they talked to one another on average, and various types of roles they played in relation to the respondent (e.g., spouse, parent, child, neighbor, coworker, friend, advisor). Finally, the respondent also reported each alter's gender, race, education, age, religion, and political party identification.

On average, the 1985 GSS respondents had 2.94 alters with whom they discussed important matters, and 55% of them were kin (Marsden, 1987). The mean egocentric density was 0.61, indicating that a majority of ego's alters knew one another. (Density of an egocentric network measured ego's perceived strength of relation for each pair of alters as 0 = *total strangers,* 1 = *very close,* and 0.5 = *other*.) American core discussion groups were more homogenous in age and education than the general population. Mean gender diversity was 0.68, suggesting a heterogeneous mix of men and women in most core discussion networks. In contrast, the low racial/ethnic heterogeneity of 0.05 indicated that most alters were the same race, which Marsden (1987) attributed to the high volume of kin nominations.

The 2004 GSS replicated the “discuss important matters” name generator and found a precipitous drop in the mean network size, from 2.94 in 1985 to 2.08 in 2004 (McPherson, Smith-Lovin, & Brashears, 2006). In 1985, the modal respondent had three confidants, but the modal response 2 decades later was zero. The percentage of respondents who listed no names increased from 10 in 1985 to almost 25 in 2004. A substantial reduction in nonkin ties resulted in networks centered more on spouses and parents, with fewer connections to alters in voluntary associations and neighborhoods. Homogeneity among egos and alters remained very high, with educational heterogeneity decreasing and racial heterogeneity increasing. The analysts speculated about how changing U.S. demographics might explain Americans’ increasing social isolation over time. (See the dispute between Fischer [2009] and McPherson, Smith-Lovin, and Brashears [2009] about questionnaire “anomalies” allegedly producing an artifactual decrease in ego network size.)

Some network researchers have explored name generators in other societies. A study of core discussion groups in The Netherlands asked a national sample with whom they “discussed important personal matters” in the past 6 months (van Tubergen, 2014). Dutch majority respondents had the highest mean number of discussion partners (2.81), followed by second-generation Turks (2.27) and Moroccans (2.11) and trailed by first-generation Turks (1.91) and Moroccans (1.74). A four-nation analysis of five network datasets from Canada, Switzerland, Chile, and The Netherlands found substantial variation in mean ego network size, ranging from 11.9 in Zurich to 23.8 in Toronto (Kowald et al., 2013). Ruan, Freeman, Dai, Pan, and Zhang (1997) replicated the GSS single-name generator in 1986 and 1993 surveys of Tianjin, China. The respondents named more alters in their core discussion groups (4.58 and 3.30 persons in the two surveys, respectively) than the 1985 U.S. mean of 2.94 alters, and they were less likely to nominate kin. Compared to 1986, the 1993 Tianjin respondents named fewer coworkers, many fewer relatives, but more friends in their core discussion groups. These changes reflected macrostructural transformations since 1978, in which China increasingly replaced lifelong employment at one workplace with more flexible market-based employment. As the transformation progressed, people came to know more contacts outside their work spheres, which increased the chances of including such alters in core discussion networks (Ruan et al., 1997).

The 1985 GSS questionnaire did not give respondents any cues about the content of the “discuss important matters” name generator but left them the burden of interpreting that phrase. This ambiguity triggered some concerns among network analysts that the structure or composition of egocentric networks varies according to respondent interpretations of the key phrase (Bailey & Marsden, 1999; Brashears, 2014). For example, asking respondents with whom they discuss important matters generates a set of alters that only partly overlaps with the list of whom they discuss health issues (Perry & Pescosolido, 2010). To scrutinize respondents' cognitive processes in interpreting "discuss important matters," Bailey and Marsden (1999) used concurrent "think-aloud" probes with 50 persons, who were asked the GSS name generator questions, followed immediately by some probes into their thought processes. Most respondents interpreted it as asking about personal matters, such as familial or interpersonal problems. Although preceding survey items apparently induced respondents to interpret "important matters" in alternative ways, these varying interpretations did not produce substantially different network compositions. Bailey and Marsden (1999) proposed four alternative strategies for future implementation of a single-name generator. The first strategy separates the definition of the content from the elicitation of alters: a respondent is first asked to define important matters in his or her own terms, then to name alters according to that definition. The second strategy involves exemplifying important matters, in which the researcher provides some examples that facilitate a respondent's definition of important matters. In the third strategy, the researcher explicitly specifies the meaning of important matters for every respondent. The fourth strategy involves rearranging the questionnaire sequence to attenuate any contextual impacts on respondent interpretations of the phrase.

The single-name generator method elicits only a fraction of the number of alters produced by using multiple-name generators. Researchers are often interested in a wider range of routine activities beyond the core relation captured using a single-name generator. For example, Fischer (1982) reported that people use the term "friend" to describe quite diverse relations. Instead of relying on a single-name generator, Fischer's multiple-name generators consisted of nine items such as house care, asking for a sizable loan, socializing, and discussing jobs, hobbies, and other personal matters. His survey of Northern California respondents elicited widely different numbers of alter names, ranging from 2 to 65, with a mean of 18.5 alters named by ego. Moreover, respondents considered 11 of these contacts to be friends, evidently interpreting "friend" indiscriminately to encompass a broad spectrum of interpersonal relations. Two survey replications in Israel generated means of 14 and 11 names (Fischer & Shavit, 1995). Both surveys found that Israelis had higher-density egocentric networks than the Northern California survey.

To compare single- to multiple-name generators, Ruan (1998) administered 11 name generators to Tianjin respondents. In addition to the GSS important matters item, she included instrumental ties such as house care and money borrowing and expressive relations such as socializing, confiding, and advice seeking. The GSS generator yielded a mean of 3.30 Tianjin alters, but the other 10 generators together produced 8.17 alters. Moreover, the Chinese respondents interpreted "discuss important matters" as social expressive issues, with most respondents identifying the same set of persons with whom they socialized (going out to dinner, shopping, or visiting) as members of their discussion network. In contrast, persons nominated by instrumental name generators were least likely to be included in core discussion networks. A study of microfinance diffusion in 75 villages of rural Karnataka, India, used 12 name generators to elicit 80,838 directed social ties from 16,403 women and their spouses (Shakya, Christakis, & Fowler, 2017). The mean total alters named across all 12 questions was 9.55. Visiting alters in their homes elicited highest number of alters, whereas "go to temple with" generated the fewest names. Domestic interaction name generators ("visit their home" and "invite home") resulted in highly clustered and centralized networks, but asking whom respondents "talk to" uncovered more egalitarian relations. Because some questions appeared to identify networks specific to cultural context, the authors suggested that researchers "should balance local relevance with generalizability when choosing network generators" (p. 157).

The largest egocentric networks can be produced by the "knowing" name generator, which asks respondents to report all persons known to them. Killworth, Johnsen, Bernard, Shelley, and McCarty (1990) showed that the knowing name generator could produce as many as 1500 acquaintances in the United States and Mexico. However, to elicit a manageable list of persons with whom a respondent has significant contact, stringent limits are often imposed. Campbell and Lee (1991) summarized four types of constraints typically built into name generators: (1) *role/content constraint* restricts respondents to focus on only one or a few types of relations in nominating their contacts, (2) *geographical constraint* asks respondents to name only those persons residing within a specified area, (3) *temporal constraint* requires respondents to name their contacts within a certain retrospective period, and (4) *numerical constraint* limits respondents to naming only *N* persons who fit the name generator criteria (e.g., "your three best friends"). Many projects use some combination of the four name generators' constraints. For example, Campbell and Lee (1991) presented 690 Nashville respondents with maps of their neighborhoods and asked them to list all neighbors in the nearest nine or 10 houses whom they know by name. They next asked them to identify everyone with whom they had either chatted for at least 10 minutes or whose homes they had visited in the previous 6 months. Compared to several other name generators, the 1985 GSS "discuss important matters" name generator had the most restrictive content, numerical, and temporal limits and elicited the smallest egocentric network size (mean of 3.01 alters) with the highest network density. Fischer's Northern California study, with its multiple-name generators, produced the largest networks (mean of 18.5 alters) but with lower density. The duration of the relations was the longest for the Nashville study, which did not impose spatial constraints, and contact frequency was the highest for the GSS. Campbell and Lee's (1991) study extended knowledge of how the restrictions imposed by name generators shape the configurations of resulting networks.

Positional Generators and Resource Generators. Nan Lin's social resources theory posited that social structures—defined by wealth, power, and status—are pyramidal and hierarchical formations in which social resources and access to these resources are embedded (Lin, 1982). To solve the empirical problem of measuring actors' social resources, *positional generators* ask respondents to report whether they have contacts with particular social positions or roles. To the extent that positions in a hierarchical occupational structure are reasonable indicators of social resources, investigating personal contacts with those positions discloses not only the types of social resources to which people may have access but also how they gain access to their alters' resources.

The selection of social positions affects how well positional generators capture ego's access to a broad range of social resources. Lin and Dumin (1986) selected a list of 20 occupations with the greatest frequency in the 1970 U.S. Census Classified Index of Occupations. Those occupations spanned the upper and lower white- and blue-collar occupational divisions. Respondents reported whether any of their personal contacts, such as relatives and friends (strong ties) or acquaintances (weak ties), held each occupational position. Using a 1975 dataset containing positional generators to examine the job-seeking process, Lin and Dumin (1986) reported that the relationship between tie strength and access to prestigious occupations was contingent on a respondent's origin, as indicated by the father's occupation. When origin was high, strong ties and weak ties provided equal access to prestigious occupations. But, when origin was low, weak ties provided better access than strong ties to more prestigious occupations. Subsequently, Lin, Fu, and Hsung (2001) applied positional generators to study job prestige and income in Taiwan. They reported that Taiwanese society exhibited gender-based inequality in access to social capital. Social capital was more useful to men in obtaining prestigious occupations and higher incomes, whereas women relied more on human capital (education) to gain better jobs and pay. Several projects deployed positional generators to examine social capital formation in occupational milieus (e.g., Hällsten, Edling, & Rydgren, 2015; Kmetty, Tardos, Albert, & Dávid, 2018; Roth, 2018). Reliability tests of 13 position generator measures found that only the volume measure (total number of accessed occupations) had good reliability, whereas measures based on occupational prestige, status, and social class fared poorly (Verhaeghe, Van de Putte, & Roose, 2013).

The choice of which occupations to include on a positional generator list obviously depends on the research question. To examine how cultural differences between social classes affect networks, Erickson (1996) selected occupations varying on three major class dimensions: control of property, organization, and skill. Because the respondents worked in the security industry, the 19 occupations based on control of skill reflected their security relevance (eight professional workers, four blue-collar workers, four policemen, two business managers, and one business owner). Interviewers instructed respondents, "Now I am going to ask you whether you know anyone in a certain line of work at all in the Toronto area, for example, whether you know any lawyers. Please count anyone you know well enough to talk to, even if you are not close to them." If a respondent knew someone with an occupation, interviewers asked about the closeness of the relations, ranging from "just knowing" to "knowing as a close friend" to "knowing as a relative." In a study of gendered social capital, Erickson (2004) purposely selected 15 male-dominated or female-dominated occupations, based on a Canadian census. Respondents were asked to identify whether they knew any men or women in each occupation. Erickson (2004) reported that men were more likely to know people in female-dominated occupations than women were to know people in male-dominated occupations. Moreover, because men were more strategically located in many social spheres, men's advantages in social networks were difficult to change, whereas both genders had more diverse ties to men than to women.

Positional generators produce egocentric networks measuring personal connections to several occupations in hierarchical ladders. To the extent that social resources are distributed within a pyramidal and hierarchical structure of occupations with differing status levels, positional generators accomplish the goal of capturing people's access to different occupations, hence to varying social resources. However, people often receive instrumental and expressive help from alters beyond those enumerated by positional generators that restrict selections to a handful of occupations. Researchers often broadly define individual social capital to encompass all forms of assistance that people may receive from their contacts. In particular, a resource generator typically captures individual social capital expressed as $SC = \Sigma_j S_j$ whereby j refers to resource items and S_j measures the availability of this type of resource (Van der Gaag & Snijders, 2004). Unlike positional generators that ask whether respondents have contacts with selected occupations, a resource generator asks whether people know anybody useful for specific resources. Van der Gaag and Snijders (2004) asked their Dutch respondents, “Do you know anybody who can. . .” help with a list of 35 items, ranging from “repair a bicycle” to “visit socially.” They reported that 17 resource generator items formed four subscales that were internally homogeneous and weakly correlated. One subscale that measured access to incumbents of prestigious occupations correlated strongly with Lin and Dumin’s (1986) positional generator. However, two subscales—access to information and access to instrumental support (e.g., help with house moving)—correlated only weakly with the positional generator, thus comprising distinct dimensions of social capital.

Measuring Total Personal Network. Various name generations are used to calculate the *total personal network,* defined as all alters known to ego. Depending on the procedure, total personal network size ranges from 250 to 5000 alters (Dunbar, 2016; Freeman & Thompson, 1989; Killworth et al., 1984). We briefly examine the checklist, reverse small-world, and network scale-up methods for measuring total personal network size.

The *checklist method* first randomly generates several names (either first or last names), then interviewers read those names to randomly sampled respondents, asking, “Do you know anybody with that name?” If a respondent recognizes a name, the interviewer asks name interpreter questions to elicit information about that alter. McCarty, Bernard, Killworth, Shelley, and Johnsen (1997) implemented this design by interviewing 793 Floridians about a list of 50 first names. Interviewers asked respondents to report any contacts having the same first name as those on the list. The respondents were told that they should know a contact by sight or name and have interacted within the last 2 years. Recording a maximum of 14 alters with this method, McCarty et al. (1997) found that the elicited alter sample underrepresented blacks, Hispanics, and Asians. In a follow-up, the researchers asked a national sample of respondents to estimate the number of people they know in subpopulations (diabetics, Native Americans) and people in relational categories (kin, coworkers). Both methods generated mean total network sizes of 291 persons (McCarty, Killworth, Bernard, Johnsen, & Shelley, 2001). Both methods “yield valid and reliable proxies for actual network size, but questions about accuracy remain.” A reanalysis of the data using a latent nonrandom mixing model resolved some problems with the earlier approach and showed that, “if the first names asked about are properly chosen, the estimates from the simple scale-up model enjoy the same bias-reduction as the estimates from our more complex latent nonrandom mixing model” (McCormick, Salganik, & Zheng, 2010).

The *reverse small-world* (RSW) *method* starts with a fictitious target, a person with an invented name who is randomly assigned geographic locations, ages, sexes, hobbies, organizational memberships, occupations, and other socioeconomic characteristics (Killworth & Bernard, 1978). Researchers ask informants to identify all alters whom they believe could either directly deliver a message or be a link in a chain to the target. After eliciting the informant alters' names, researchers ask name interpreter questions about those alters' demographic characteristics, their relations with the informant, and the extent to which the alters know one another. Killworth et al. (1984) implemented an RSW method by interviewing 15 Jacksonville, Florida, informants about 100 American and 400 international targets. The informants named an average of 134 alters to reach these 500 fabricated targets. Most of the alters (86%) were friends of the informants, and more than half were male. Freeman and Thompson (1989) concluded that RSW approaches capture only a portion of an individual's total network because informants cannot name more alters than the number of targets. To address this concern, Killworth et al. (1990) combined the RSW with the checklist method and the GSS name generator. They found that the mean size of total personal networks was 1700 for Floridians and 600 for Mexico City residents. Drawing from the same dataset, Bernard, Johnsen, Killworth, McCarty, Shelley, and Robinson (1990) investigated the overlap among these different methods. They reported that the GSS name generator and the multiple social support name generator together accounted for only 18% of the total personal network generated by the RSW method. The checklist method (based on last-name matching) produced the largest personal network.

The *network scale-up method* enables researchers to estimate the sizes of hidden-populations—for example, people at-risk of HIV infections, such as drug injectors, female sex workers, and men who have sex with men—from sampled network data (e.g., Maltiel, Raftery, McCormick, & Baraff, 2015). Using a collection of Florida subpopulations of known size, Killworth, Johnsen, McCarty, Shelly, and Bernard (1998) also asked about a subpopulation of unknown size (persons who were HIV seropositive). Their best estimates were "108 members of the network defined by 'having been in contact with during the previous two years,' and (approximately unbiased) 1.6 million for the seropositive subpopulation." Feehan and Salganik (2016) proposed a *generalized scale-up estimator,* which requires two samples, one from the frame population and another from the hidden population. They demonstrated the advantages of their approach over basic scale-up methods with a series of simulation studies.

Archival Documents. Documents, tape recordings, videos, music, maps, and other multimodal historical records provide raw information for social network analysis, after recoding into suitable formats. Compared with surveys, archival data are relatively inexpensive, pose no burden on informant time and efforts, and may contain high-quality longitudinal information when maintained over time. Archival data come in various forms, including personal letters and diaries, webpages, patent citations, book and article references, and computer network communications. Several online data aggregators and providers—such as Acxiom, Edgar-Online, LexisNexis, SDC Platinum, the Bureau of Labor Statistics, and the U.S. Patent and Trademark Office (USTPO)—store vast amounts of information about corporate boards of directors, mergers and acquisitions, strategic alliances, ownership of subsidiaries, and patent citations, which greatly facilitates research on business organization networks. Archival data are especially valuable but underutilized in micro-organizational research, that is, analyses of individual-level activities occurring within organizations (Barnes, Dang, Leavitt, Guarana, & Uhlmann, 2018).

When the contents of letters, diaries, and emails carry information about interpersonal ties, scouring those messages provides fruitful data about personal networks as well as larger social structures. This approach is particularly beneficial for historians whose subjects from bygone eras render survey methods impossible. For example, Edwards and Crossley (2009) analyzed 26 letters and eight speeches, written between 1909 and 1914 by an English suffragette, Helen Watts, to reconstruct her egocentric network ties to other militants in the Votes for Women social movement. Combining content analysis with the network data extracted from the documents enabled the researchers "to sharpen and make more precise a vague qualitative impression regarding centrality of Helen's parents within 'her' ego-net" and then to explore questions raised by each method (p. 58). (See also Crossley, Edwards, Harries, & Stevenson, 2012.) In another example, Alexander and Danowski (1990) investigated ancient Rome's social structure through the letters of Cicero, the renowned orator and influential politician who straddled two elite social classes: the "knight class," which was without office-holding, and the "senatorial class" of officeholders. They reviewed 280 letters between Cicero and his acquaintances, friends, and relatives spanning 18 years. Their data management program recorded 1914 relations among 524 individuals, including the name and rank of both persons, and such relational contents as giving, ordering, serving as intermediary, helping, describing in a negative way, and visiting. A major finding was that, although senators and knights often opposed one another on particular issues, they appeared to make up a single, well-integrated, and interlocked social class.

Journal publications and patents, which both report extensive citation lists, provide another good data source for network analysis. In contrast to traditional content analyses, citation network analysts cluster articles based on who cites whom, thus revealing how scholars construct their fields and how new research areas emerge and evolve. Recent citation network analyses of research domains include physical activity and health (Verala, Pratt, Harris, Lecy, Salvo, Brownson, & Hallal, 2018), educational administration (Wang & Bowers, 2016), fuel cell technologies (Ho, Saw, Lu, & Liu, 2014), and digital humanities (Gao, Nyhan, Duke-Williams, & Mahony, 2018). Persons and organizations filing for patents generate masses of data that chart interorganizational knowledge flows. A focal innovation contains a patent citation list that provides information about its connections to prior foundational innovations, while subsequent innovations building on that focal innovation must cite it. Technological domains are comprised of interlocking egocentric networks among sets of focal innovations and their ties to other patented innovations. Using the USPTO's database, Aharonson and Schilling (2016) developed a system for mapping technological space that enabled them to develop and apply novel measures of technological capabilities. At the firm level, "researchers can use the measures to assess the diversity of a firm's technological footprint, and how that footprint changes over time. Temporal changes in a firm's trajectory, in turn, can be useful for studies on topics such as organizational learning, responses to environmental shocks, mimetic isomorphism, and more" (p. 93). Other researchers have analyzed patent citation networks to investigate knowledge flows in such fields as organic solar cells (Choe, Lee, Kim, & Seo, 2016), semiconductors (O'Reagan & Fleming, 2018), genetically modified organisms (Ho & Cheo, 2014), and unmanned aerial vehicles a.k.a. drones (Kim, Lee, & Sohn, 2016).

The citation lists in academic publications and patents differ in two important ways. First, journal articles tend to cite a broad range of books and articles, whereas patent citations have a much narrower focus on only the prior inventions that contribute significantly to an innovation (Meyer, 2000). Second, journal citations serve a broad array of purposes, such as giving credit to related work, correcting one's past work, and disputing previous claims. In contrast, the more restrictive purpose of patent citation is to acknowledge previous works serving as the building blocks for the current invention. Thus, the networks constructed using patent and journal citations may differ fundamentally in their contents, necessitating careful distinctions between them.

Governments and business organizations routinely collect massive amounts of information about their relations with citizens, employees, and other organizations. Two examples where network analysis can be conducted on archived data that were originally gathered for other purposes are healthcare (Chambers, Wilson, Thompson, & Harden, 2012) and law enforcement (Hollywood, Vermeer, Woods, Goodison, & Jackson, 2018; Liu, Patacchini, Zenou, & Lee, 2012). Electronic medical records (EMRs), federally mandated for all U.S. practitioners, are intended to improve physician efficiency in documenting patient diagnosis and treatment. The immense store of information locked away in EMRs is a challenge for network analysts to design new tools for its extraction and analysis (Aickin, 2011). Tracking patients' trajectories and outcomes as they are referred among healthcare organizations could reveal bottlenecks and gaps in the complex U.S. healthcare "system" where reform and restructuring are urgently needed. Procedures to protect patient privacy by anonymizing EMRs must be created, for example, using blockchain technology (Ekblaw, Azaria, Halamka, & Lippman, 2016). Law enforcement organizations apply network analysis to solve crimes by identifying and collecting information from known associates of suspects. A study of 83 corporate frauds involving 436 defendants extracted data from indictments and secondary sources on corporate conspiracy networks (Steffensmeier, Schwartz, & Roche, 2013). Women were rarely involved in conspiracy groups, but, when they were, women had more minor roles and less profit than their male coconspirators. Research on terrorist organizations relies on public news sources to reconstruct clandestine networks (Basu, 2014; Ouellet, Buchard, & Hart, 2017). With social life increasingly online, cyberspace hosts a variety of crimes and creative efforts to detect and prevent them (Al-garadi, Varfathan, & Ravana, 2016; Kamat & Gautam, 2018).

Data mining methods and tools have proliferated with the exponential expansion of Big Data and rapidly falling costs of data storage (Aggarwal, Kapoor, & Srivastava, 2017; Leech, Collins, & Onwuegbuzie, 2017; Rafiei, Agichtein, Baeza-Yates, Kleinberg, & Leskovec, 2018). In addition to social media on the World Wide Web, major sources of Big Data are economic transactional records—from stock prices, to bank transfers, to business sales and purchase data—and computer-connected electronic and digital sensors, a.k.a. the Internet of Things—from security monitors, to industrial production systems, to wearable medical devices, and mobile metadata. Big Data come in diverse formats: spreadsheets, texts, pdfs, photographs, audio recordings, video clips, social media posts. Data miners sift through enormous raw datasets to detect patterns and apply machine-learning and statistical methods to predict trends and outcomes. Businesses and government agencies can analyze customer- and citizen-generated information to identify and respond to problems, such as housing bubbles and natural disasters. For example, Akay, Dragomir, and Erlandsson (2015) used network methods to analyze opinions about a lung cancer drug posted on a social media site, to identify influential users, and to "provide rapid, up-to-date information for the pharmaceutical industry, hospitals, and medical staff, on the effectiveness (or ineffectiveness) of future treatments" (p. 210). Other applications of data mining include predicting political orientation and homophily from Twitter messages (Colleoni, Rozza, & Arvidsson, 2014), optimizing renewable energy network design (Cai, 2016), and predicting customer churn from electronic banking services (Keramati, Ghaneei, & Mirmohammadi, 2016). With the growing complexity of relations in Big Data deluges, cloud computing, and the Internet of Things, a challenge for social network analysts is understanding how they can contribute value to a discipline dominated by algorithms written by computer scientists and engineers.

3.3. Cognitive Social Structure

The key question in mapping cognitive social structure (CSS) is "Who knows who knows whom?" CSS analysts investigate variations in network participants' perceptions of the relations among other network members (Brands, 2013). To create a CSS dataset, researchers ask each person to report her or his subjective perceptions of every dyadic relation, for example, "who is friends with whom?" In a complete network of N actors, CSS data collection of one type of relation results in *N cognitive maps,* each map showing an ego's beliefs about the presence or absence of ties among the N network members. Given the potentially huge time and recall burdens (an informant is asked to make $N^2 - N$ judgments about directed dyadic connections), CSS data collection is often attempted only for very small networks or for a few salient linkages (e.g., "Who are X's three closest friends?" or "Which five persons does Y trust most?"). For example, Krackhardt (1987) collected CSS data on two types of relations—advice seeking and friendship—from 21 managerial employees, resulting in two sets of 21 cognitive maps each displaying $(21)^2 - 21 = 420$ directed ties. An egocentric network is a subset of CSS, in which ego describes only the perceived relations among her or his alters but not the ties among other network participants.

Two frequent uses of CSS are (1) to measure consensus on a network's structures as perceived by its participants and (2) to assess biases in perceived network structures in comparison to alternative criteria. *Consensus* is the extent of agreement between two or more informants' judgments or assessments of social relations. Important theoretical propositions are that consensus is greater among actors with similar attributes (such as race, gender, or age), similar relational ties, and similar network locations. *Bias* refers to the accuracy of an informant's perceptions of network relations when compared to a specified criterion. One criterion is a comparison to the aggregate of all the other participants' perceived networks. Another criterion is relational data obtained from direct observations of participant interactions. Network analysts have long observed major discrepancies between respondent self-reports and behavioral measures of relation and have proposed methods to deal with those inconsistencies. Although the large majority of network data is collected using face-to-face or telephone interviews, few researchers investigate the CSSs used by participants in forming their reports. Investigators typically treat all responses as unequivocally "objective" social facts. However, a few network methodologists tried to theorize and measure the sources of divergence between information based on respondent perceptions and behavioral observations (e.g., Batchelder, 2002; Carley & Krackhardt, 1996; Casciaro, 1998; Daniel, Silva, Santos, Cardoso, Freitas, & Ribeiro, 2017; Johnson & Orbach, 2002; Krackhardt, 1987; Neal, Cappella, Wagner, & Atkins, 2011; Yenigün, Ertan, & Siciliano, 2017).

CSS also helps researchers identify systematic perceptual errors in respondent self-reports. Kumbasar, Romney, and Batchelder (1994) constructed a CSS by interviewing 25 computer hardware engineers. Each informant was shown a randomly selected engineer's name and asked to mark, on a list of all 25 engineers, which persons they perceived as a friend of the selected engineer. This process was repeated for all names, including the informant's own name. The resulting 25 cognitive maps could then be aggregated across all the informants into a global network, using matrix algebra methods (see Chapter 4). Comparing the self-perceived friendship networks with the global network, Kumbasar et al. (1994) concluded that individuals tended to see themselves as closer to the center of their own network representations than to the global network's center. Furthermore, individuals tend to construct a cognitively consistent network; that is, they viewed themselves as surrounded by friends, who also had numerous friendships among themselves. An analysis of perceived network accuracy in a technical call center found that employees with more formal and informal power were more accurate about negative relations, more accurate about their own incoming ties, and that network accuracy was related to employee outcomes, including subsequent transfers, promotions, and exits from the organization (Marineau, Labianca, Brass, Borgatti, & Vecchi, 2018).

The notion that individuals try to construct consistent mental images of their personal social worlds is a longstanding principle in cognitive psychology. People experience psychological or emotional distress when their CSSs have unbalanced relations (Heider, 1958; Newcomb, 1961).

For example, if you perceive that two close friends dislike one another, your cognitive triad is unbalanced and very likely stressful. You might balance the triad by denying any antagonism between your friends. Unbalanced relations tend to trigger applications of a “balance schema,” wherein people change their perceptions to achieve a rebalanced image of network relations (e.g., Betancourt, Kovács, & Otner, 2018; Sentis & Burnstein, 1979). However, because cognitive correction occurs primarily among friends, especially among close friends, people seem relatively unstressed by perceived imbalances among others who are not their friends. In trying to achieve and maintain psychological balance, informants are more prone to reporting distorted and biased information about their friends' relations than about their more distant contacts. In contrast to this cognitive balance explanation, Krackhardt and Kilduff (1999), drawing from Taylor and Fiske's (1978) cognitive miser model, argued that network participants typically possess little knowledge about socially distant persons, that is, to whom they are connected only through several intermediaries. By activating a balance schema to fill in knowledge gaps, respondents give biased reports about both close and distant relations. But their motives differ: For close relations, the balanced schema is activated to achieve psychological and emotional comfort, but, for distant relations, the schema compensates for the lack of knowledge about the relations. Krackhardt and Kilduff (1999) found a curvilinear relationship between network structure and perceived balance: employees perceived both their immediate friendship circle and their most peripheral contacts as more balanced than were those persons at intermediate distances. A major implication is that both close and distant ties can be subject to high informant bias.

Selective exposure to sources of information and advice is one mechanism to construct balanced networks (Metzger, Hartsell, & Falangin, 2015; Stroud, 2017). The increasingly polarized political atmosphere in the United States may arise from voters deliberately constructing networks to avoid exposing themselves to people, mass media, and social media conveying messages incompatible with their CSSs (Medders & Metzger, 2018). Partisans of left- and right-wing ideologies cocoon themselves inside echo chambers where they hear only what they want to hear and disregard the rest (Guo, Rohde, & Wu, 2018; Tsang & Larson, 2016). Donald Trump's successful campaign for the presidency was enhanced by his constantly tweeting bombastic tropes—“rapist and criminal” Mexican immigrants, “fake news” media, and “Crooked Hillary”—that arrived unfiltered to reinforce his core constituents' dispositions to outraged grievances (Haynes & Sattler, 2017). Whether a voter believes that global climate change is caused by human activity, or is a “Chinese hoax” designed to hurt American manufacturing, depends substantially on whether those narratives fit, or fail to fit, with the views held by others in their CSSs (O'Gorman, 2018; Riley, Wang, Wang, & Feng, 2016). In any event, the Commander in Tweet's memes will continually bombard the nation for the remainder of his one or two terms.

3.4. Missing Data

Social network studies are especially sensitive to missing data. In egocentric data collection, an ego with *N* alters is asked to report on C_N^2 undirected dyadic relations. Specifically $C_N^2 = \frac{N!}{2!(N-2)!}$ where *N*! (pronounced "N factorial") is the product of all the positive integers from 1 to *N*. For example, if ego has five alters, $C_N^2 = \frac{5!}{2!\times(5-2)!} = \frac{(1\times2\times3\times4\times5)}{(1\times2)\times(1\times2\times3)} = \frac{120}{12} = 10;$ therefore, ego must give information on 10 undirected ties. For directed relations, the number of ties among ego's alters is twice as large: 2 C_N^2 in the example; ego would have to report on 20 directed ties. The relational response rate (*R*) for egocentric networks is calculated by dividing the number of reported ties by the total number of possible dyadic relations among the alters. For example, if ego reported about eight of the 10 undirected relations, then *R* = 0.80, or 80%; if ego failed to report on six of the 20 directed relations, then *R* = 0.70, or 70%.

Calculating the response rate for a complete social network is more complicated. A complete network consists of the dyadic relations among all pairs of the *N* actors in the network. For a undirected network, *R* is less attenuated because a report by one member of a dyad suffices when the measure is reliable. For example, to measure friendship between actors A and B, information provided by either informant could be used to determine whether that relation is present or absent. That is, unless both A's and B's reports about one another are missing, we measure their friendship with a single report. In general, for a complete undirected network of *N* actors with no alter reports from *M* actors, the response rate for a particular relation is calculated as

$$R\begin{cases} = 100 \text{ percent when } M = 0 \text{ or } M = 1 \\ = \left(1 - \dfrac{C_M^2}{C_N^2}\right) \times 100 \text{ percent when } 1 < M < N \\ = 0 \text{ percent when } M = N \end{cases}$$

Table 3.1 Missing Response in a Network With Five Actors

Number of Missing Nodes	*Nodal Response Rate (%)*	*Relational Response Rate (%)*
0	100	100
1	80	100
2	60	90
3	40	70
4	20	40
5	0	0

For example, in a network of five actors, the nodal response rate and the relational response rate for varied numbers of missing nodes are the following (see Table 3.1):

To illustrate, assume that the network's five actors are labeled A, B, C, D, and E. The 10 undirected dyadic relations among these five actors are AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE. If actor A fails to report its relations, those dyadic ties can be obtained from the other four actors' reports about A. Thus, the relational response rate is 100% despite missing reports from one node. When the missing nodes range between 2 and 4 ($1 < M < N$), the relational response rate is $\left(1 - \frac{C_M^2}{C_N^2}\right) \times 100$ percent. For example, if three nodes (A, B, and C) do not report their relations with anyone, the response rate is 70% $\left[\left(1 - \frac{C_3^2}{C_5^2}\right) \times 100 = 70\%\right]$. In this case, three undirected relations are missing (AB, AC, and BC), but the other seven dyads are reported by at least one member. If no actors provide information, both nodal and relational response rates fall to 0%. Because the nodal response rate is computed as $\left(1 - \frac{M}{N}\right) \times 100$ and $\frac{M}{N}$ is always greater than $\frac{C_M^2}{C_N^2}$, the relational response rates for undirected networks are always higher than the nodal response rates at every level of missing nodal reports.

In contrast, missing nodal information has a substantial impact on the relational response rates of directed networks. Asymmetries occur in many types of ties, such as giving advice, trusting, and liking. Actors A and B have two directed relations: AB denotes A's report of its relation to B and BA represents B's reports of its relation to A. Therefore, each missing node results in missing relational information about that node's ties directed toward all other actors. Assuming a network of N actors with M missing nodes, the relational response rate for a directed network is the following:

$$R\begin{cases} = 100 \text{ percent when } M = 0 \\ = \left(1 - \dfrac{M \times (N-1)}{2 \times C_N^2}\right) \times 100 \text{ percent when } 0 < M < N \\ = 0 \text{ percent when } M = N \end{cases}$$

To demonstrate, the five-actor network has $(5^2 - 5) = 20$ directed dyadic relations: AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE, and ED. If every actor reports its relations with all others, we obtain a 100% relational response rate. If one node (e.g., A) is missing, all of A's reports are also missing, resulting in four unreported relations (AB, AC, AD, and AE) among the 20 dyadic relations, a response rate of 80%. Because each node in a directed network must assess its relations with the other nodes, M missing nodes ($0 < M < N$) generate $M \times (N - 1)$ missing relations. Of course, when no nodes report any relations ($M = N$), both nodal and relational response rates equal 0%.

With a bit of arithmetic deduction, we now prove that relational response rate always equals nodal response rate in a directed network. Suppose we have a directed network of N nodes and M missing nodes; the nodal response rate is $1 - \dfrac{M}{N}$, whereas the relational response rate is $\left(1 - \dfrac{M \times (N-1)}{2 \times C_N^2}\right)$. In particular,

$\dfrac{M \times (N-1)}{2 \times C_N^2} = M \times (N-1) \div \left(2 \times \dfrac{N!}{2! \times (N-2)!}\right) = \dfrac{M \times (N-1)}{N \times (N-1)} = \dfrac{M}{N}$. Thus, in a directed network with five nodes, the response rate for both nodes and relations are 100% with no missing node, 80% with one missing node, 60% with two missing nodes, 40% with three missing nodes, 20% with four missing nodes, and 0% with all five missing nodes.

In the preceding discussion, we argued that a report by one member of a dyad in a undirected network could be treated as a reliable measure of the dyadic relation. Scholars typically assert that this practice must be exercised with great caution because nonreciprocal relations often occur (Stork &

Richards, 1992). For example, in a communication network, person A reports talking with person B, but B claims never to have talked with A. This contradiction calls into question the practice of reconstructing the relation between a responding actor and a missing actor from information provided only by a nonmissing node. Stork and Richards (1992) suggested that similarity in a pair of actors' characteristics, such as their age, sex, and education, may be a good indicator of their dyadic tie confirmation. In addition, network reliability, measured as the proportion of all dyadic relations described identically by both members, indicates how well a single report characterizes the link between respondents and nonrespondents. When network reliability is high and both members of a dyad have identical or very similar characteristics, one member's report should serve as a reliable proxy for their relation. Kossinets (2003) endorsed Stork and Richards's method of reconstructing a dyadic tie from the respondent actor's report, provided that the overall number of nonrespondents is low.

Egocentric network studies are vulnerable to missing data because people often fail to describe ties among their alters. Researching the 1985 GSS discussion networks, Burt (1987) reported that the missing data for the GSS egocentric studies are less severe: among the total 1534 respondents who enumerate 4483 discussion partners, only 66 respondents provided incomplete network data involving 195 discussion partners. Burt compared two kinds of discussion partners—those with complete network data and those with only partial network data. He found that 35% of the former group's alters were not close to any other alters, as reported by the respondent, but that figure increased to 59% for the latter group. The major implication is that the missing relations among alters in an egocentric network tend to be weak ties. Egos seem more likely to report relations for alters perceived as well connected to other alters than they are to report relations for alters who are seen as isolated. This finding also explains why, contrary to other methodological research, the college-educated GSS respondents produced more incomplete egocentric network data than the less-educated respondents. College graduates tended to enumerate egocentric networks consisting of many weakly tied alters, which in turn engendered more missing reports about these alters' relations.

Missing nodes and/or relations may be particularly problematic in whole network research. If a key entity having important relations with others is omitted, the resulting network structure may be severely distorted. For example, the political network of movers-and-shakers in a city would look quite different with or without the mayor (whereas the dogcatcher's absence would likely have little impact). Or, the investigation of a criminal network may be unable to detect crucial covert communications due to the criminals' evasive actions, painting an inaccurate picture of the enterprise (e.g., Berlusconi, Calderoni, Parolini, Verani, & Piccardi, 2016). Unlike survey samples, where any randomly missing case has little effect on estimating a population mean, whole network analysis may produce misleading structures due to a single missing case or a handful of absent links. Analysts must take exceptional care to identify all important whole network members and their connections.

As no foolproof *post facto* remedy to the missing data problem exists, the solution to this problem lies in convincing more respondents of the importance of participating in the research (Knoke & Kuklinski, 1982, p. 35). To elicit higher participation, such as the 90% response rate achieved by some network studies, extraordinary research efforts often require a combination of different persuasion techniques such as personal letters, phone contacts, and monetary inducements. The missing data problem is not a unique plague on social network studies using survey methods. Archival studies, for example, are also susceptible to the curse of missing data. Granted, the completeness of archival data is at the mercy of the data repository's maintenance practices. However, given a certain level of data availability, the competency of the coders and the efficacy of the software in data mining can also make a substantial difference in the amount and quality of information extracted. Poorly trained coders and ill-designed software definitely contribute to sizable amounts of missing data.

3.5. Measurement Error

Theories and methods of network analysis implicitly assume that all nodes and relations are measured without error in collecting and coding data. But, as in all scientific fields, that assumption is simply untenable. "Almost all studies of real-world network structures contain experimental error of some kind, and frequently of many kinds simultaneously" (Newman, 2017, p. 1; see also Newman, 2018a). Using data with substantial errors to estimate network parameters, such as degrees and centralities, may lead to incorrect inferences. Fortunately, network methodologists have made considerable progress toward understanding and detecting network errors (Borgatti, Carley, & Krackhardt, 2006; Huisman & Krause, 2014; Kossinets, 2006; Zipkin, Schoenberg, Coronges, & Bertozzi, 2016). Typical procedures start with an artificial or real network, randomly add or delete varying numbers of nodes or relations, then observe changes in network parameters estimated with the modified datasets. We lack the space to review in detail the large and growing research literature on network measurement errors. Interested readers may wish to begin with Smith and Moody (2013) and Smith, Moody, and Morgan (2017) on node-level missingness; Žnidaršič, Ferligoj, and Doreian (2012, 2017) on the stability of blockmodels; and Huisman and Steglich (2008) and Krause, Huisman, and Snijders (2018) on longitudinal network data. This section extends the preceding section on missing data by discussing measurement error scenarios in the typology presented by Wang, Shi, McFarland, and Leskovec (2012).

Missing or Spurious Nodes. Entities that should be included in a network may be absent. Primary sources of such *false-negative nodes* are nonresponses, such as respondent refusals, and, in organizational research, not-up-to-date membership rosters. In bibliometric research, large portions of a citation network may be missing because of coding errors for digital libraries due to author-name ambiguities, that is, many authors having identical names (Hussain & Asghar, 2017; Levin & Heuser, 2010; Smallheiser & Torvik, 2009). One example of name disambiguation in graphs applied cluster analysis to documents within a low-dimension network space (Zhang & Hasan, 2017). The researchers concluded that their algorithm "clusters the documents belonging to a single person better than other existing network embedding methods" and is particularly useful "in anonymized network where node attributes are not available due to the privacy concern" (p. 192).

False-positive nodes are entities that are incorrectly present in a network. Some respondents may be prone to exaggerate their number alters, for example, reporting others as friends rather than as acquaintances. Russia's malicious interference in the 2016 U.S. presidential election revealed the pervasiveness of fake Twitter, Facebook, and other social media platform accounts, which distort the size and structure of online political communities (Badaway, Ferrara, & Lerman, 2018; Lazer et al., 2018). Separating true-positive from false-positive nodes is crucial to improving network data quality. Mark Newman proposed an expectation-maximization (EM) method for making optimal estimates "in the presence of both richly textured data and significant measurement uncertainty" (Newman, 2018b, p. 1). Applied to two U.S. student networks, the EM algorithms found very few false positives but a much larger proportion of false negatives (about 20% for physical proximity between pairs of people, as measured by mobile phone software, and 33% for mutually reported friendships). EM posterior probability distributions can be used to estimate network parameters of substantive interest to researchers, along with their standard errors.

Missing or Spurious Relations. *False-negative edges* occur when a respondent omits relations that should have been reported. Omissions are more likely to occur when name generator instructions rely on recalling alters from memory (Bernard, Killworth, Sailer, & Kronenfeld, 1984) than where recall is aided by giving respondents a name list. Questionnaires that restrict the number of choices ("Name your three best friends") unnecessarily place an artificial constraint on network size, density, and connectivity. *False-positive edges* result when relations are incorrectly reported that actually don't exist. Most of Facebook's 2.2 billion users know that a substantial percentage of their so-called "friends" never make any subsequent contact after an initial friend-request. The mean and median numbers are 338 and 200, respectively, indicating a highly skewed distribution (Mazie, 2018). Compounding the problem is the prevalence of fake accounts, estimated by Facebook to comprise more than a quarter of the total (Ng, 2018). Obviously, researchers should try to purge phony friends before analyzing the properties of online ego-network data.

Tie decay (or its converse, *tie persistence*) is important for determining whether a prior relation no longer exists. If a social setting is disrupted—by graduating from school, moving to a new community, receiving a job promotion—people faced opportunities for making new connections but also choices about whether to retain or drop their prior contacts (Kleinbaum, 2017; Shipilov, Rowley, & Aharonson, 2006). Among academics, three indicators of tie strength that slow the rate of decay are frequency, closeness, and scientific contributions (Mohdeb, Boubetra, & Charikhi, 2016). Bridges, which connect people not otherwise connected, tend to "decay at an alarming rate," threatening the stability of social capital (Burt, 2002). Researchers studying the duration of interorganizational collaborations, such as strategic alliances and joint ventures, have difficulty determining whether collaborations have been terminated, due to substantial amounts of missing data in the widely used Securities Data Corporation Platinum database (Lee, 2017). One expensive option is to contact firms directly by email or phone, but high nonresponse rates may result in a biased subsample (Cui, Calantone, & Griffith, 2011). More commonly, investigators estimate contemporaneous alliance portfolios by assuming that every alliance ceases after some specific period, typically using moving widows of 3 to 5 years' duration (Andrevski, Brass, & Ferrier, 2016; König, Liu, & Zenou, 2014; Lee, 2017).

Falsely Aggregated or Disaggregated Nodes. These forms of network measurement error either mistakenly combine two or more nodes as a single entity or incorrectly treat a single node as separate entities, respectively. An example from co-citation network research is a database in which an author with multiple name spellings is treated as different nodes, whereas different authors with identical name spellings are treated as a single node. Great care must be taken to disambiguate identities, but sometimes the task is impossible (Wang, Shi, McFarland, & Leskovec, 2012, p. 398). In particular, huge online databases pose formidable obstacles to cleaning dirty data manually, and automated methods using artificial intelligence have their own challenges (Chu, Ilyas, Krishnan, & Wang, 2016).

Correcting Errors. Based on simulations of errors on four node-level network properties using a pair of very large networks, Wang, Shi, McFarland, and Leskovec (2012, p. 408) recommended that concentrating on cleaning high-degree nodes would improve measurement reliability more than labor- and time-intensive efforts to collect and clean an entire large dataset. “Also, when setting a threshold for tie strength, using a lower threshold, which yields more false positive edges, results in more reliable measures than higher thresholds, which yield more false negative edges.” Faced with missing network data due to respondent nonresponse, analysts may impute and replace values from available information (Folch-Fortuny, Villaverde, Ferrer, & Banga, 2015; Huisman, 2014; Stork & Richards, 1992). For example, if a person did not respond to an item about discussing work-related matters, just fill in those missing row values with the corresponding column values. The assumption is that, if one member of a dyad reports a discussion partner, the other person took part in the conversations. However, Wang, Shi, McFarland, and Leskovec (2012) cautioned that missing-value imputations of nodes or relations are a viable error correction strategy only in “scenarios where false negative nodes are more detrimental than false positives” (p. 408). In the contrary situation “imputation could introduce even greater measurement error with the presence of spurious nodes.” The bottom line, as always, is to proceed with caution and run analyses using alternative ways of handling missing values (including simply removing all nodes with missing data) to see whether outcomes are robust. If not, the results may reflect the method of handling measurement errors more than actual network structures.

3.6. Collecting Network Data

Recent years have witnessed tremendous growth in computer-supported or online social networks. Commercial social media sites, such as Facebook, Twitter, Instagram, and LinkedIn, draw billions of users, including persons, organizations, social events, and other entities. Other commercial products, such as Facetime and WeChat, use smartphones and other mobile devices to deploy as phone apps. They achieved phenomenal success in attracting users and revenue. For example, WeChat boasted more than 1 billion users by 2018 and raked in $11 billion annually from advertising. A couple of reasons lie behind such tremendous growth of online social networking services. First, computers and smartphones are connected almost 24/7 via Ethernet, WiFi, and cellular signals. Second, social networking websites and social networking apps operating on computers or smartphones enable multiplex connections among more than 2.5 billion people worldwide (Statista, 2018), creating a vast network among this huge population that sustains billions of messages and images exchanged, posted, and shared every day.

Massive online social networking activities provide a goldmine to scholars interested in online networking behaviors. Focusing on the network nodes, scholars can extract valuable information on nodal attributes (individual persons or collective entities), nodal centralities, and network density and centralization. Studying the contents of network ties, scholars have glimpsed at what draws people and organizations together and what sets them apart. Investigating the whole network, researchers understand the process of network formation and dissolution. However, before investigators can examine those issues in much depth, the unique features of those data—their sheer massive volume and problematic storage requirements—present challenges to scholars trying to obtain access. Subsections that follow discuss those challenges and progress in tackling those issues.

Big Data Social Media Networks. Big Data became a core topic for network researchers since the turn of the millennium. In 2001, Doug Laney discussed the Big Data opportunity with regard to data volume, data velocity, and data variety. Laney's (2001) talk inspired many to examine the Big Data revolving around the 3-Vs (Volume, Velocity, and Variety). In particular to social media networks, Big Data is omnipresent. All major commercial social network sites (Facebook, LinkedIn, Twitter) and smartphone apps (WeChat, Facetime) far exceed 100 million users, and messages exchanged among the users in billions. The rate of information flowing through those networking sites and apps is 1 or more petabytes every hour (Lee & Sohn, 2015, Chapter 1). One petabyte is 1024 terabytes, which in turn is 1024 gigabytes (a gigabyte is 1 billion bytes), and one terabyte equals the data volume of the entire Library of Congress collections, consisting of 167 million items. In other words, the volume of messages exchanged or shared among users of all social networking sites and apps is at least 1024 Libraries of Congress every hour.

The second dimension of Big Data is velocity, or the rate of data accumulation. The key phrase in the preceding paragraph is "every hour," which means that, on a typical day, the data accumulated exceeds 24,576 Libraries of Congress. How does one keep up with such gushers of data pouring in hourly and daily? Scholars wishing to stay on top of issues concurrent with ongoing networking activities face daunting challenges. The third Big Data dimension is variety. In traditional data format, data are stored digitally or, on some occasions, in syntax forms. However, online social networking data can be video, audio, locational information, music clips, images, pictures, text, screen shots, and so on. They come in different shapes, forms, and structures. Existing analytical tools do not provide sufficient storage to inventory such large quantities of unstructured data, let alone to analyze them. Putting the 3-Vs together, we can truly see that social media networking data are Big Data insofar as they are large, rapidly accumulating, and highly diverse, all of which make data handling (storage, processing, analytics, transfers) very difficult and costly.

Social Science Studies of Social Media Networks. Facing such challenges, social scientists fall back on their original strength: using questionnaire items in surveys to capture respondents' online activities. Many researchers focus on social networking usage and its consequences for an array of outcome measures such as mental health, physical health, life satisfaction, civic engagement, political activity, co-curricular participation (learning experiences complementing school), and general social trust (Bouchillon, 2018; Junco, 2012; Shakya & Christakis, 2017; Valenzuela, Park, & Kee, 2009). One commonality of these studies is that the data came from surveys tapping into respondents' self-reported online networking activities (Facebook usage) and related variables. For example, Valenzuela, Park, and Kee (2009) sent emails to 40,360 students in two large public universities in Texas and collected information with Survey Monkey software from 2603 respondents between 19 and 28 years of age. The researchers asked the students questions about their life satisfaction, social trust, civic and political participation, and intensity of Facebook use. Also using Survey Monkey, Junco (2012) sent questionnaires to 5414 students in a northeastern public university, achieving a final sample of 2368 respondents. Bouchillon (2018) used Survey Sampling International (SSI) to identify respondents. SSI first asked respondents to complete a short set of demographic questions, with which SSI drew the sample of respondents whose demographic features (age, sex, race/ethnicity, and region) approximated the U.S. population. Bouchillon collected information from 1005 respondents regarding their civic engagement, general trust, and Facebook usage.

Shakya and Christakis's (2017) study differed slightly from the research designs described previously. They started with the 2013, 2014, and 2015 waves of the Gallup Panel Social Network Study. They conducted a survey at each wave, a total of 10,680 respondents. They asked for permission to access their Facebook accounts, but only 3195 respondents agreed. They measured such variables as mental health, physical health, and life satisfaction with respondent self-reports. However, the respondents' Facebook networking activities—their number of Facebook friends, number of likes clicked, number of links clicked the past 30 days, and frequency of updating their status in the previous month—were collected objectively by examining the Facebook accounts.

Interestingly, these studies found very different effects of Facebook usage and outcome variables. The most optimistic was Bouchillon's (2018) finding that Facebook users had great levels of civic participation and generalized trust. Junco (2012) discovered that time spent on Facebook increased co-curricular activities, but playing games on Facebook was detrimental to co-curricular activity. Valenzuela, Park, and Kee (2009) found that intensity of Facebook use and life satisfaction, social trust, and civic/political participation were positively correlated, although the effects are small. In light of their results, the authors called for a relaxing of concerns about the negative impact of Facebook exposures on the one hand but also for a quenching of enthusiasm about Facebook as a spur to social trust, civic engagement, and political involvement on the other. In contrast to those studies, Shakya and Christakis's (2017) research was more pessimistic about Facebook. They reported that using Facebook is associated with reduced mental and physical health. In particular, higher rates of "likes clicked," "links clicked," and "status updates" were associated with a 5% to 8% of standard deviation decrease in self-reported mental health. They further contended that the negative effects of Facebook usage may outweigh the positive impact of offline activities on those outcomes.

Although studies discussed previously all investigated Facebook activities, several studies used survey methods to examine other social networking platforms such as Twitter, LinkedIn (Boyd & Ellison, 2008), and WeChat (Lien, Cao, & Zhou, 2017). WeChat is unique in starting as a popular social networking app in smartphones (it subsequently added a website counterpart), and it operates primarily inside China. Using conventional survey questionnaires, Lien, Cao, and Zhou (2017) collected responses from hundreds of informants at four major Chinese cities (Beijing, Shanghai, Guangzhou, and Shenzheng). Their research provided information about customer satisfaction and continuous usage intent by WeChat members.

Online Data Mining of Social Media Networks. Traditional data collection with survey questionnaires to elicit respondents' self-reports helps social scientists to gain a glimpse of the massive online social media network. At the same time, computer scientists and information engineers work in parallel to develop various data mining techniques for collecting data directly from online social media networks. They commonly use online crawlers or scrapers to extract information straight from webpages (Russell, 2011). These approaches present two clear advantages over survey methods: (1) the information is more objective than self-reports because they are scraped directly from respondents' online archival records, and (2) online data mining is less costly and more feasible for collecting information on millions of records within short time spans.

Computer scientists use Python programming language, various crawlers, and webpage scrapers, in connection with an application programming interface (API), to parse immense amounts of data in Twitter, Facebook, and LinkedIn. For example, users can mine data in Twitter with trending topics, hashtags, tweets, and retweets. In Facebook, analysts can delve into likings and friendings to detect the popularity of Facebook members. Foster, Ghani, Jarmin, Kreuter, and Lane (2017, Chapter 2) illustrated how to use API to scrape information from the web. They particularly pointed out that the researchers' network can be revealed by mapping their "hidden" networks. Details about Python programming, API coding, webpage scraping, and crawling are beyond the scope of this introductory book. We advise readers to consult special volumes (e.g., Foster et al., 2017; Ignatow & Mihalcea, 2018) to acquire in-depth knowledge of those topics.

In addition to the data being massively huge, Big Data also challenges years of social science studies and experiences on data cleaning, which transforms messy, noisy, and unstructured data into a well-defined, clearly structured and quality tested dataset. Big Data can be very messy as computer scientists and information engineers commonly mesh different datasets from diverse sources together to get a complete picture of the activities under investigation. Big Data can also be very unstructured, as video and audio clips, symbols, screen shots, and traditional numerical data can all be poured in to overwhelm data storage and processing capabilities. Are social scientists so outpaced by the development of Big Data and analytics that they have to unlearn much of what they have learned over years? Of course not. Also, social scientists are rightly concerned about Big Data's inaccuracies (erroneous, duplicated, or missing links), incompleteness, and inconsistencies. In addition, the traditional inferential statistics based on probability sampling don't apply to Big Data—literally every nonzero correlation and regression slope will reject a null hypothesis, simply due to the massive data size, irrespective of whether they are truly substantively unimportant. So, what are the good barometers to measure data representativeness for Big Data?

All those great challenges present great research opportunities for data scientists to work across the aisles (e.g., social scientists working with computer scientists and information engineers) to capitalize on the goldmine of Big Data residing on social media networking sites and apps. On the one hand, computer scientists and information engineers are overly focused on data collection processes (e.g., designing smart crawlers with machine learning) and pay less attention to the postcollection data analytics of substantive social issues. Social scientists, on the other hand, lack the techniques for mining, storing, and analyzing Big Data, resorting to conventional survey methods to glimpse at online social networking activity. Only by working together can social scientists obtain the benefits of Big Data access and analytics, while computer scientists acquire better understanding of social issues implicated in their machine designs. Through cross-disciplinary teams, computer and information scientists can help social scientists with Big Data collection and data parsing, and social scientists can provide guidelines to computer scientists to improve their designs of data mining techniques.

We conclude with a brief look at ethical concerns that network researchers routinely confront in designing, collecting, and analyzing data and reporting the results (for a more detailed examination, see Chapter 4 in adams, 2020). As in all human subjects research, safeguarding at-risk participants takes precedence over other considerations. Protocols for securing informed consent and protecting privacy and anonymity, especially for underage or vulnerable participants, have been developed and enforced in all disciplines, from medicine to labor relations to education. Social network research has a distinctive problem, in that, rather than aggregating and summarizing variables for anonymous participants, the heart of the enterprise is displaying graphs revealing detailed connections among participants. Even if the points are labeled with pseudonyms, participants in small social systems may be able to deduce some nodal identities. Social network projects conducted inside organizations, such as corporations and government agencies, are fraught with ethical dilemmas. Management controls researchers' access to their employees and, as a condition for granting access, may require the right to see and use the data when making personnel decisions. For example, supervisors could insist on seeing fully labelled graphs showing workers' friendship, trust, social support, and antagonism networks. Consequently, researchers cannot promise anonymity and confidentiality to employees. In turn, workers may be reluctant to answer truthfully if they believe that network data could be used by management to discipline or fire "bad apples." Borgatti and Molina (2003) proposed ethical principles for organizational research when one purpose of network analysis is to make decisions affecting employees. Survey participation must be truly voluntary, not coerced by managers. Informed consent requires being "extremely explicit" about possible adverse consequences of answering survey questions, for example, showing workers samples of network diagrams and explaining possible interpretations managers could draw from them. Researchers might rely more on nonsurvey data collection, such as email logs or project collaborations "to avoid asking employees to incriminate themselves." As with all types of social research, network studies should provide feedback directly to the respondents "as payment in kind for their participation" (p. 348). The overarching ethical principles that network researchers must strive to uphold are to minimize potential harms to respondents and to safeguard continued researcher access to organizations for future network studies.

Chapter 4

BASIC METHODS FOR ANALYZING NETWORKS

This chapter presents some basic methods for analyzing social networks. Specifically, we discuss representing networks as graphs and matrices and measuring nodes, dyads, subgroups, and whole networks.

4.1. Network Representation: Graphs and Matrices

Social networks can be represented with two common forms: graphs and matrices. Graphs are effective communicative tools for visually informing audiences about the social network structures, but they do not permit computer processing and arithmetic calculation. Conversely, matrices facilitate calculation and computer processing but are not as intuitive as graphs at revealing social network structural features. In this section, we illustrate various social networks with both graphs and matrices, covering four fundamental types of social networks: binary undirected, binary directed, valued undirected, and valued directed graphs of networks.

A fourfold typology of networks results from cross-tabulating of two basic dimensions of social networks: whether network ties are binary or valued and whether ties are directed or undirected. Table 4.1 shows this typology, but two points are worth noting. First, this typology, which stresses the positive effects of social networks, does not consider the sign of social relations, which may be either positive (e.g., like, collaborate, assist) or negative (e.g., dislike, oppose, undermine). Although the overwhelming majority of research on social networks emphasizes positive relations, a few network researchers have investigated negative social connections, such as antagonism, gossip, and social undermining (Fang, Duffy, & McAllister, 2015; Greetham, Hurling, Osborne, & Linley, 2011; Oberst, Wegmann, Stidt, Brand, & Chamarro, 2017). Second, social network analysis software programs, such as UCINET, typically contain a wide range of algorithms that place various prerequisites on appropriate data structures. Analysts should check their software's user manual to understand those prerequisites and make accurate interpretations of the program outputs.

Social network Type I in Table 4.1 denotes a binary undirected network whose relations indicate only the presence or absence of connections between pairs of nodes. Binary ties have only two values: 1 for a relation, 0 for no relation. The intermarriage network of 15th-century Florentine ruling families represents such a network (Padgett & Ansell, 1993). Type II is

Table 4.1 Typology of Social Networks

Binary or Values of Social Networks	*Direction of Social Network*	
	Undirected	*Directed*
Binary	Type I Binary undirected social networks	Type II Binary directed social networks
Valued	Type III Valued undirected social networks	Type IV Valued directed social networks

a binary directed graph, also called a *digraph*, that indicates the presence or absence of connections between pairs of nodes and also the direction of those ties. Morimoto and Yang's (2013) study of friendship networks (who nominates whom as friend) among several cohorts of graduate students exemplifies in this type. Type III networks are valued undirected graphs, where the numerical value of a tie shows the intensity, strength, frequency, or volume of connections between pairs of nodes. Knoke's (2001) analysis of the number of concurrent strategic alliances between pairs of high-tech firms indicates that the ties are valued and undirected (see also Yang & Hexmoor, 2004). Finally, Type IV indicates valued directed networks, where ties reveal the intensity or volume of the relations/communications between pairs of nodes as well as the direction of the relationship between pairs of nodes. A prominent example is an international trade network where countries are the nodes and the ties measure the total dollar values of goods and services exchanged (Garlaschelli & Loffredo, 2005; Sopranzetti, 2018). The ties are directed because pairs of nations import and export differing monetary magnitudes to one another.

Figures 4.1 to 4.4 illustrate the four types of networks, using the example of an artificial communication network among six graduate students. In the following notation, g is the number of nodes in a network; hence, $g = 6$ in these four graphs. Figure 4.1 is the simplest Type I network, with lines showing the presence/absence of undirected binary communication ties among pairs of students. Table 4.2 is a square matrix corresponding to Figure 4.1. A matrix is often referred to as a *sociomatrix* in the social sciences and an *adjacency matrix* in computer science and engineering. The six students are listed in alphabetical order in both the rows and columns, with a cell entry showing whether a pair of students communicated. The *order* of a matrix is the number of rows by number of columns, expressed as order g-by-g (also denoted gxg). The example matrices have order 6-by-6 or 6x6.

Figure 4.1 An Undirected Binary Graph of an Artificial Communication Network Among Six Graduate Students

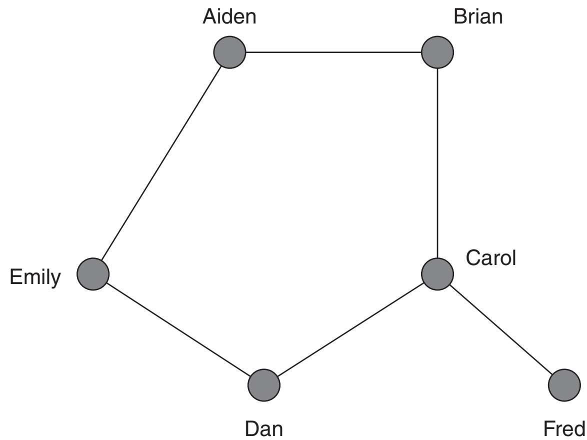


Figure 4.2 A Directed Binary Graph of an Artificial Communication Network Among Six Graduate Students

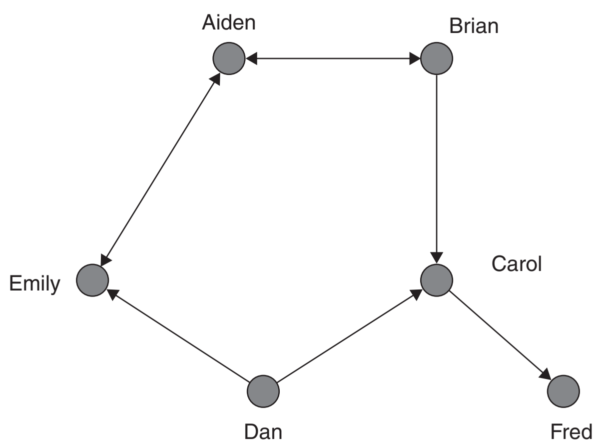


Table 4.2 is a *binary matrix* because the value 1 or 0 in a cell (denoted as x_{ij}) indicates the presence or absence of communication between the pair in row *i* and column *j*. For example, $x_{\text{Aiden, Emily}} = 1$, indicating a tie between Aiden and Emily, whereas $x_{\text{Brian, Dan}} = 0$, indicating the lack of connection between Brian and Dan. An undirected matrix exhibits *symmetry* because

Figure 4.3 An Undirected Valued Graph of an Artificial Communication Network Among Six Graduate Students

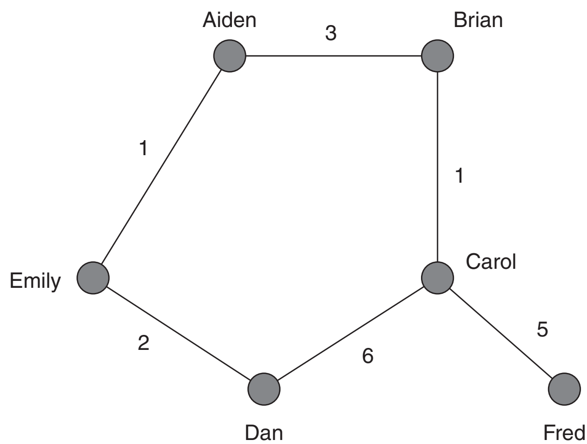
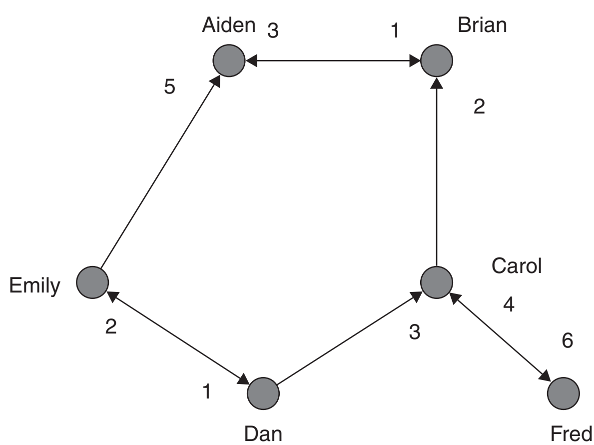


Figure 4.4 A Directed Valued Graph of an Artificial Communication Network Among Six Graduate Students



the cell entries for every pair of nodes i and j have identical values (i.e., $x_{ij} = x_{ji}$). All diagonal values are 0 ($x_{ij} = 0$, if $i = j$) because students do not communicate with themselves.

Row margins, which are not a part of a matrix, are the sums of cell values across all columns within row i ($x_i+ = \sum_{j=1}^{g} x_{ij}$). For example, the row

Table 4.2 Matrix Representation of Figure 4.1

	Aiden	*Brian*	*Carol*	*Dan*	*Emily*	*Fred*
Aiden	0	1	0	0	1	0
Brian	1	0	1	0	0	0
Carol	0	1	0	1	0	1
Dan	0	0	1	0	1	0
Emily	1	0	0	1	0	0
Fred	0	0	1	0	0	0

Table 4.3 Matrix Representation of Figure 4.2

	Aiden	*Brian*	*Carol*	*Dan*	*Emily*	*Fred*
Aiden	0	1	0	0	1	0
Brian	1	0	1	0	0	0
Carol	0	0	0	0	0	1
Dan	0	0	1	0	1	0
Emily	1	0	0	0	0	0
Fred	0	0	0	0	0	0

margin for Carol is 3, showing that Carol communicates with three people in the network (i.e., Brian, Dan, Fred). Column margins are the sums of cell values across all rows within column j ($x+j = \sum_{i=1}^{g} x_{ij}$). The column margin for Fred is 1, revealing that Fred has a tie to just one network member (Carol). When a network is undirected, the corresponding row margin and column margin of a node are always equal.

The Type II network graph in Figure 4.2 adds direction to the relations, specifying who is sending messages to whom. Two reciprocal ties occur, between Aiden and Brian and between Emily and Aiden. Other pairs are unidirectional: Brian sends a message to Carol, but Carol does not send a message to Brian. Dan also sends a message to Carol without receiving any answer. Likewise Dan's message to Emily is unreciprocated. Carol sends a communication to Fred, but he does not reply. The matrix in Table 4.3 exhibits all the relations shown in the graph, with some notable differences from undirected binary graphs. First, senders and receivers of communications must be distinguished and, by convention, the row actors (i) are senders and the column actors (j) are receivers. Second, considering that many social relations are not often reciprocated, reciprocity may be the exception rather than the rule. In network notation, $x_{ij} \neq x_{ji}$. For example, Brian sends a message to Carol, but Carol sends no message to Brian ($x_{\text{Brian, Carol}} = 1$ and $x_{\text{Carol, Brian}} = 0$). Third, the row margin of a node may not equal its column margin ($x_{i+} \neq x_{+j}$ when $i = j$). The row margins reveal the extent to which actors send messages to others in a network, whereas the column margins show the extent to which actors receive messages from others. In the example, Carol sends one message (to Fred), thus her row margin is 1, but she receives two messages (from Brian and Dan, respectively); therefore, her column margin is 2.

The Type III network in Figure 4.3 assigns values to the lines between pairs of nodes, in the example, depicting the number of messages exchanged between the two students. For example, although Dan and Carol exchange six messages, Aiden and Emily only exchange one message. Table 4.4 is the undirected valued matrix corresponding to the graph in Figure 4.3. Table 4.4 bears great similarity to Table 4.2, except that the values in the cells are not restricted to binary codes (0 and 1). Instead, the integer entries show the number of messages exchanged between pairs of nodes. As a result, Table 4.3 exhibits symmetry because the cell entries for each pair of students i and j have identical values.

However, the interpretations of row and column margins are not straightforward because both the number of ties and tie strength are involved in those sums. For example, Dan's row margin (also his column margin) is 8, which is twice as high as the margins of Aiden or Brian (4 each). However, Dan, Aiden, and Brian each have connections to two other students. Dan's higher margins reflect his greater number of messages exchanged with Carol and Emily, in contrast to the fewer messages that Aiden and Brian exchange with their partners. Another example is that Fred's row and column margins = 5, which results from his high exchange volume with Carol. We recommend that social network analysts report the network actors' row margins of both the valued network and its binary version. Then, dividing the valued margins by the binary margins yields the mean communication per node. Thus, Dan's mean number of exchanges per student is (2+6)/2 = 4, Aiden and Brian each have means of (1+3)/2 = 2, and Fred has the highest mean (5)/1 = 5. What are the means of Carol and Emily?

Finally, the Type IV graph in Figure 4.4 is the most complex network of the four types, entailing not only the direction of communications but also

Table 4.4 Matrix Representation of Figure 4.3

	Aiden	*Brian*	*Carol*	*Dan*	*Emily*	*Fred*
Aiden	0	3	0	0	1	0
Brian	3	0	1	0	0	0
Carol	0	1	0	6	0	5
Dan	0	0	6	0	2	0
Emily	1	0	0	2	0	0
Fred	0	0	5	0	0	0

Table 4.5 Matrix Representation of Figure 4.4

	Aiden	*Brian*	*Carol*	*Dan*	*Emily*	*Fred*
Aiden	0	1	0	0	0	0
Brian	3	0	0	0	0	0
Carol	0	2	0	0	0	6
Dan	0	0	3	0	2	0
Emily	5	0	0	1	0	0
Fred	0	0	4	0	0	0

their values as shown by the numbers on the lines. The number near each arrowhead indicates the number of messages sent to the student to whom the arrow points from the student at the arrow tail. For example, the 5 near the arrow pointing to Aiden shows that Emily sends five messages to her (conversely, Aiden receives five messages from Emily). Similarly, Brian sends three messages to Aiden, whereas Aiden only sends one message to Brian. Carol sends two messages to Brian, whereas Brian sends none to Carol. Table 4.5 is the matrix representation for the directed valued network, where the senders are in the rows, the receivers are in the columns, and the cell entries are the number of communications sent from actor *i* to actor *j*. In addition, the row margin is the total number of outbound messages from an actor. For example, Carol's row margin is 8, and a close examination of her outbound ties reveals that it's the sum of two messages to Brian plus six messages to Fred. Similarly, the column margin indicates the total number messages received by a student. For example, Carol's column margin is 7, comprised of four messages from Fred plus three from Dan. We again recommend that network researchers report the network actors' row margins of both the valued network and its binary version. Then, dividing the valued margins by the binary margins yields the mean messages sent and received per node.

4.2. Nodes: Centrality, Power, Prestige

In this section we discuss a basic element in a social network, the node, also called *actor* or *vertex* in different disciplines. One of the most important topics in analyzing nodes is to identify the important and prominent entities in a network. Often *node centrality* reveals actors' power, influence, visibility, or prestige. To quantify node centrality, three fundamental measures are available: *degree centrality, closeness centrality,* and *betweenness centrality*. Each measure is premised on different principles for identifying importance or prominence in a network.

The first and simplest node centrality measure is **degree centrality**, which examines the extent to which a specific node is connected with other nodes in the network. In a binary undirected network, the degree centrality of node *i* is its row margin (or column margin). The formula for calculating degree centrality in a matrix is

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij} (i \neq j) \tag{4.1}$$

where $C_D(N_i)$ denotes degree centrality of node *i* and $\sum_{j=1}^{g} x_{ij}$ counts the number of direct ties that node *i* has to the $g - 1$ other nodes *j*. ($i \neq j$ excludes *i*'s relation to itself; i.e., the main diagonal values of the matrix are ignored.) The computation of $C_D(N_i)$ involves simply adding all the cell entries in either actor *i*'s row or column (because undirected relations have a symmetric data matrix, the corresponding row and column cell entries must be identical).

However, for binary directed graphs and matrices, every node has two distinct degree centralities: one is *in-degree centrality*, which measures the extent to which a node receives relations or nominations from the $g - 1$ other nodes in the network. The second measure is *out-degree centrality*, which indicates the extent to which a node sends out relations or nominations to the $g - 1$ other nodes. Formally, in a matrix of a binary undirected network, node *i*'s out-degree centrality equals its row margin (formula 4.1), whereas its in-degree centrality is its column margin, or

$$IDC\left(N_i\right)=\sum_{j=1}^{g} x_{ji} \quad \left(i \neq j\right) \tag{4.2}$$

where IDC(N_i) stands for the in-degree centrality of node *i*. The formula aggregates all the values for a given node *i* (in the *i*th column) across its different rows. In directed networks, the in-degree centrality (column margins) often does not equal the out-degree centrality (row margins) for network nodes. What do the in-degree and out-degree centralities mean substantively? Knoke and Burt (1983) discussed the directed relations in directed networks. They stated that, in directed networks, the mere participation or involvement in certain relations is less important than distinguishing between senders and receivers of relations. For example, in a reporting network of a workplace, rank and file employees routinely report to their supervisors and managers about their work activities, whereas managers and supervisors send commands that direct the work activities of employees under their direct management.

Following Knoke and Burt's (1983) discussion, we define *prestige* as the extent to which a social actor in a network "receives" and "serves as the object" of relations sent by others in the network. Such reasoning separates sender from receiver, or source from target, emphasizing inequalities in control over resources, authority, and deference. Senders of commands attempt to exert authority, power, or influence over the behaviors of the receivers of the commands. Directed network ties not only reflect the factual inequality in power, authority, and control of resources but also can help to identify new leaders emerging over time. For example, in an advice-seeking network, persons who are asked by others to give advice can exert tremendous influence over the advice-seekers' actions. Therefore, despite the lack of any "official title," actors who are frequently asked to give advice to others are actually very powerful in the sense of charismatic leadership. In addition to the resources, power, and prestige, a node's in-degree centrality can also reflect its popularity. For example, in a classroom environment, students who are frequently chosen as friends are popular and welcomed members of the class, but students who receive few or no friendship nominations are unpopular or even isolates (Morimoto & Yang, 2013).

Node degree centrality has a major flaw when it comes to comparing nodes across different social networks; it's susceptible to network size. For example, node A with degree centrality of 10 in a network of 11 actors is drastically different from node A with the same degree centrality (10) in a network of 1001 actors. The former is a highly central node connected to all other nodes in the small network, whereas the latter has no connections to the vast majority of nodes in the large network. To eliminate the effect of network size on degree centrality of nodes, researchers should use normalized degree centrality,

$$C'_D = \frac{C_D\left(N_i\right)}{g-1} \tag{4.3}$$

Normalized degree centrality simply divides the degree centrality formula (4.1) by the total number of nodes (*g*) in the network minus 1, which is the maximum number of direct connections any node can have in a network of size *g*. For the preceding example, the normalized degree centrality for the node A is 1.00, whereas the measure for node B is .01 (or as percentages, 100.0% and 1.0%, respectively). Normalized degree centrality is particularly useful to compare degree centrality of nodes across different networks with drastically different sizes. Computation of degree centrality of directed networks is the same as that shown in formula 4.3, except for the distinction between normalized in-degree centrality (in-degree centrality divided by $g - 1$) and normalized out-degree centrality (out-degree centrality divided by $g - 1$).

Computations of degree centrality for valued graphs/networks are more complicated than for binary graphs. The chief complication is that high degree centrality for a node in valued graphs results from two competing sources (confounding factors): high connectivity with other nodes and/or high values to the ties with other nodes. For example, a node in a value network with degree centrality of 10 may have a wide span of connections to many other nodes where each connection has a low value (e.g., ties to 10 other nodes with values = 1 apiece), or the node may have few connections to other nodes but those connections have high values (e.g., a single connection with value = 10). Therefore, any measures of degree centrality in valued graphs need to be carefully distinguishing between the two sources and identify to what extent the measure comes from the connections and to what extent it comes from the values of those connections.

A variation on degree centrality is *eigenvector centrality*. It weights each of the nodes with direct ties to an actor by their centralities. Hence, a node's eigenvector centrality is proportional to the sum of centralities of the other actors to whom it is connected. Eigenvector centrality scores can be interpreted as measuring actor influence or popularity: a high score means that an actor is connected to many nodes that also have high scores. Google's PageRank algorithm, used to rank webpages in its search engine results, is a well-known example of eigenvector centrality. Borgatti (2005) and Bonacich (2007) are accessible discussions of eigenvector centrality in comparison to alternative measures.

Closeness centrality measures how fast node *i* can reach all other nodes in a network. Although the emphasis in degree centrality is solely on the number of direct ties, closeness centrality emphasizes speed of connections through both direct and indirect ties to all network nodes. For that reason, closeness centrality may be ideal for measuring communication in social networks, or even computer networks. Closeness centrality of node *i* is computed by taking the inverse of the node's *geodesic distance* to the $g - 1$ other nodes, where geodesic distance measures the length of the shortest path between a pair of nodes (i.e., the smallest number of steps in a path connecting node *i* and node *j*, where a direct tie has a geodesic = 1). The formula for calculating closeness centrality is the inverse of the sum of geodesic distances between node *i* and the $g - 1$ other nodes:

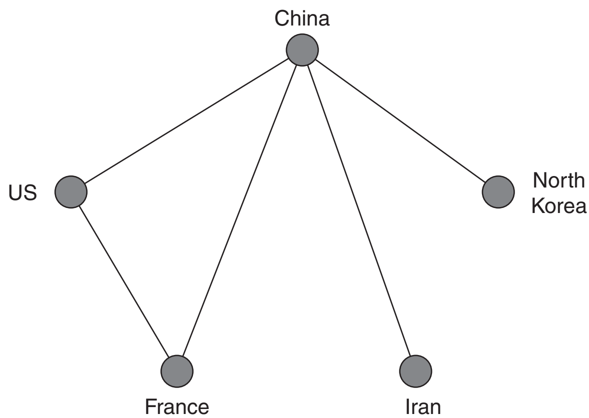
$$C_C\left(N_i\right) = \frac{1}{\left[\sum_{j=1}^{g} d\left(N_i, N_j\right)\right]} \quad \left(i \neq j\right) \tag{4.4}$$

$C_C\left(N_i\right)$ is the closeness centrality for node *i*, and $\left[\sum_{j=1}^{g} d\left(N_i, N_j\right)\right]$ computes the sum of geodesic distances between node *i* and each of the $g - 1$ other nodes. The shorter the distance between node *i* and all others, the larger its value of the closeness centrality. A high value of closeness centrality indicates many short paths and hence speedy communication to others; conversely, low closeness centrality implies long distances between a node and the other nodes and slow transmission of messages. Because the denominator in math functions cannot be zero, a completely isolated node does not have any closeness centrality, i.e., its closeness centrality is undefined. Thus, closeness centrality only applies to the nodes in a connected graph or subgraph.

To illustrate, we use a network of trade relations among five nations from Hafner-Burton, Kahler, and Montgomery (2009, p. 564). Figure 4.5 shows that network (the original graph was valued, but, to simplify, we dichotomized the network to indicate only presence or absence of ties). Using formula 4.4, we calculate the closeness centralities of the United States and China as 1/6 and 1/4, respectively. Thus, China can reach other nodes in the network faster than does the United States. Conversely, the closeness centralities of North Korea and Iran are 1/7, the farthest away from other nations. What is the closeness centrality of France?

Closeness centrality is susceptible to network size. Central nodes in large networks tend to have low closeness centrality scores, in contrast to peripheral nodes in small networks that tend to have high closeness values. To

Figure 4.5 Network of Trade Among Five Nations



illustrate, we conjured up the two artificial networks in Figures 4.6 and 4.7. Using formula 4.4, the closeness centrality for node A in Figure 4.6 is 1/6, and the closeness centrality for node A in Figure 4.7 is also 1/6. However, the A nodes in these two networks occupy very different positions. Node A in Figure 4.6 is the central node with direct connection to all other nodes, whereas node A in Figure 4.7 is a peripheral node with a direct tie with only one other node. So what makes a central node in one network have the same closeness centrality as a peripheral node in another network? The answer is network size: the graph in Figure 4.6 has seven nodes, whereas the graph in Figure 4.7 has only four nodes.

In this situation, we should calculate normalized measures of closeness centrality, which take network size into account in multiplying each node's closeness centrality score by $g - 1$, the total number of nodes in a network minus 1.

$$C'_C\left(N_i\right) = \frac{g-1}{\left[\sum_{j=1}^{g} d\left(N_i, N_j\right)\right]} \quad \left(i \neq j\right) \tag{4.5}$$

Applying formula 4.5 to the graphs in Figures 4.6 and 4.7 yields normalized closeness centrality scores for the two nodes A of 1 and 0.5, respectively. By eliminating the effect of differential network size, we now see that Node A in the Figure 4.6 is indeed much closer to the six other nodes than is Node A in Figure 4.7 to the three other nodes. Much like normalized

Figure 4.6 A Network of Seven Nodes in Star Shape

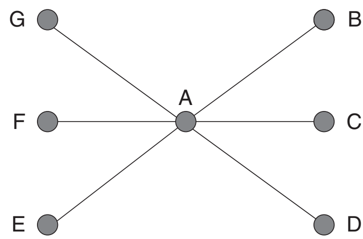


Figure 4.7 A Network With Four Nodes in Chain Shape

degree centrality, normalized closeness centrality is indispensable for comparing nodes across two or more networks of different sizes. To compare nodes within a single network, closeness centrality suffices.

Betweenness centrality measures the extent to which node i is on the geodesic paths of all other pairs of nodes in a network. Let's define g_{jk} as the number of geodesic path(s) between nodes j and k in a network and $g_{jk}(N_i)$ as the number of geodesic path(s) between j and k that includes the node i. (More than one geodesic may exist between two nodes.) Then, dividing $g_{jk}(N_i)$ by g_{jk} measures the proportion of geodesic path(s) connecting j and k in which node i is involved. The sum of the proportions across all pairs measures the extent to which node i is involved in the geodesic distances of all dyads in a network. (In the following formula, g_{jk} indicates geodesics, not to be confused with g, the number of network nodes.)

$$C_B\left(N_i\right) = \Sigma \frac{g_{jk}\left(N_i\right)}{g_{jk}} \quad \left(i \neq j \neq k\right) \tag{4.6}$$

Betweenness centrality is 0 when node i falls on no geodesic path for all pairs of the other $g - 1$ nodes. It reaches a maximum value of 1 when node i falls on every geodesic path for all the pairs in the network. How many geodesic paths exist in a network with g nodes, excluding the node i?

Assuming, for simplification, that every dyad has only one geodesic path, then a network with g nodes would have $\frac{(g-1)(g-2)}{2}$ maximum possible geodesic paths. This maximum exists because $C^2_{g-1} = \frac{(g-1)!}{2\times(g-1-2)!} = \frac{(g-1)(g-2)}{2}$. Much like degree centrality and closeness centrality, betweenness centrality is very sensitive to the network size. For that reason, Wasserman and Faust (1994, p. 190) proposed using normalized betweenness centrality, especially when comparing nodes across networks of different sizes. Normalized betweenness centrality divides a betweenness centrality score (formula 4.6) by its maximum possible value of $\frac{(g-1)(g-2)}{2}$.

$$C'_B(N_i) = \frac{C_B(N_i)\times 2}{(g-1)(g-2)} \tag{4.7}$$

Normalized betweenness centrality ranges between 0 and 1. It is 0 when the original degree centrality is 0, meaning that a given node i is not sitting on the geodesic paths between any pairs of network nodes. Conversely, it is 1 when a given node i is sitting on every geodesic path connecting all dyads. Thus, for node i, the closer the normalized betweenness centrality score is to 1, the more often the node i is part of the network's geodesics, and thus potentially exerting greater influence over network relations.

Applying formula 4.6 to Figure 4.5, we observe that, whereas China has betweenness centrality = 5, all other nations' betweenness centralities = 0. China's high betweenness centrality is due to its presence on five geodesic paths: U.S.-China-North Korea, U.S.-China-Iran, France-China-North Korea, France-China-Iran, and Iran-China-North Korea. If the United States and France didn't have a direct tie, China would be on the geodesic path between the United States and France as well, thus reaching a maximum possible betweenness centrality in a graph of five nodes $\left(\frac{(5-1)(5-2)}{2} = 6\right)$.

Table 4.6 summarizes the three centrality measures for each of the nations in Figure 4.5. China has the highest score in each of these measures, especially on betweenness centrality. Previously, we stated that degree centrality indicates the volume of direct connections, and closeness centrality measures the speed of communication. Here, we assert that betweenness centrality reveals an actor's brokerage position within a network. In his classic essay on structural holes, Ronald Burt (1995) discussed how social

Table 4.6 Centrality Measures for the Five Nations in Figure 4.5

	Centralities		
Nations	*Degree*	*Betweenness*	*Closeness*
United States	2	0	1/6
France	2	0	1/6
China	4	5	1/4
Iran	1	0	1/7
North Korea	1	0	1/7

actors reap the benefits by occupying structural holes also known as brokerage positions (e.g., as an intermediary between a pair of actors who are not directly connected). Burt argued that actors who fill structural holes are thereby able to collect diverse and valuable information from unique sources. Furthermore, those actors are capable of controlling the timing and the content when they pass messages among others who aren't directly connected, thus deriving control benefits. Along this line of reasoning, China gains leverage by receiving diverse and valuable information from the other four nations, reflected in its high betweenness centrality (brokerage positions). Furthermore, China also enjoys control benefits because it can decide whether, when, and to what extent it passes information from one nation to the other. Those decisions are contingent on how they benefit China as the network's information controller.

Other than for degree centrality, we don't discuss applications of betweenness and closeness centrality to binary directed graphs. We also don't present centrality measures for valued graphs. Those topics require elaborate expositions beyond the scope of this introductory volume. Instead, we recommend reading two classic articles examining those topics in depth. White and Borgatti (1994) extended betweenness centrality to directed graphs, and Freeman, Borgatti, and White (1991) described how to apply centrality measures to valued graphs.

4.3. Dyads: Walk, Path, Distance, Reachability

We now consider measures at a different level of network analysis: dyads or pairs. Four concepts are important for measuring dyadic relations: walk, path, distance, and reachability. We examine each concept in some detail, starting with their application to the simplest network form: binary undirected relations.

In a graph, if two nodes are connected by a line representing some type of tie, the nodes are *adjacent*. If node X is adjacent to node Y and Y is adjacent to Z, then the two lines are *incident* on Y. A *walk* is defined as a sequence of nodes and lines, beginning with a node and ending with another node, in which every node is incident with the lines preceding and following it in that sequence. The beginning and ending nodes in a walk may be different or the same, and nodes and lines may appear more than once. A *path* is a walk in which no node and no line appears more than once (Wasserman & Faust, 1994, p. 106). For example, a path in a communication network requires that an actor receives or sends a message just one time. From these definitions, a walk is very loosely defined because any sequences of nodes and lines can comprise a walk. But, a path is a walk with very strong restrictions on its nodes and lines. A *path distance*, or *path length*, is the number of lines in the sequence from the beginning node to the end node. As defined in the preceding section, if more than one path exists between a pair of nodes, the path with the shortest length is called the geodesic. More than one geodesic may exist if two or more paths connecting the same dyad each have the same shortest distances. But, if no path links a dyad, then no geodesic exists, and the distance between the two nodes is considered to be either infinite or undefined.

We illustrate these concepts with the graph in Figure 4.8 representing a small undirected binary communication network. Berto could communicate with Carmen via four walks that are paths because none uses a node or line more than once: Berto-Ana-Carmen, Berto-Diego-Carmen, Berto-Ana-Diego-Carmen, and Berto-Diego-Ana-Carmen. However, the third and fourth paths are longer than the first and second; their distances or lengths are three. Both the first and second paths have distances of two; hence, they're each geodesics of the Berto and Carmen dyad. How many walks and paths does Carmen have connecting her to each of the three others? What are those path distances? And what are those three dyads' geodesics and path lengths?

An important property of a pair of nodes is the dyad's *reachability*: whether two nodes can connect to one another using a walk or a path. Figure 4.9 shows that, in this undirected graph of seven people, all pairs are reachable because every dyad is connected via either direct or indirect paths. Thus, Elaf can reach Badr using the path Elaf-Chana-Ahmed-Badr and Badr can reach Elaf by reversing that sequence. If a dyad is reachable, it has one or more geodesics. What are the two geodesics of Chana and Ghada? In directed graphs, reachability requires that the direction of the arrows be maintained when tracing paths. A directed path must follow a sequence where a node receives an arrowhead from a preceding node, then sends an arrow toward the next node in the path. In other words, all arrowheads in a path sequence must point in the same direction. Although one member of a dyad may reach the second, the reverse direction may or may not exist. For example, Figure 4.10 shows that, although Elaf can reach Badr (via path Elaf→Chana→Ahmed→Badr), Badr cannot reach Elaf by any directed path. Indeed, because she has no outgoing arrow, Badr can't reach anyone! Sometimes two members of a dyad can reach one another by geodesics of differing lengths. What is the shortest path from Chana to Ghada and the shortest path from Ghada to Chana, and what are their lengths?

Figure 4.8 Walks, Paths, and Geodesics

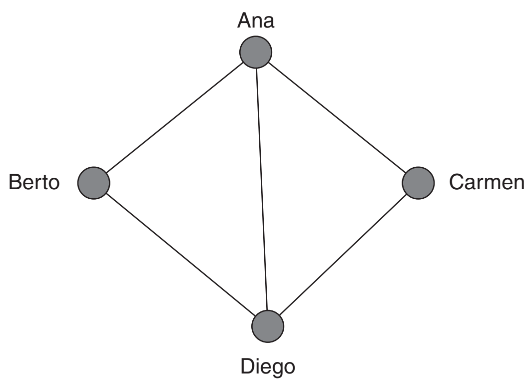


Figure 4.9 Reachability in a Binary Undirected Network

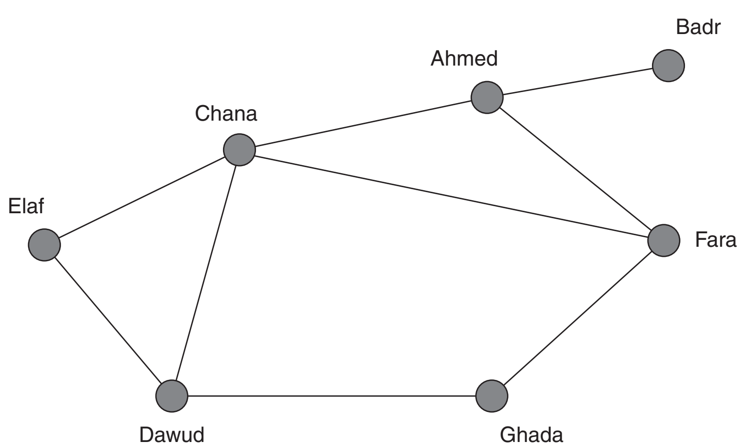
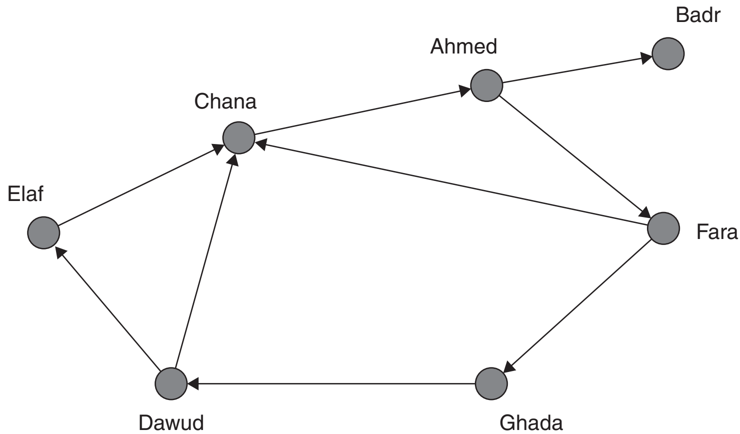


Figure 4.10 Reachability in a Binary Directed Network



4.4. Subgroups: Transitivity and Cliques

Subgroups are an important concept in social network analysis because they indicate the extent to which small, cliquish, and cohesive groups exist within a larger network. In Borgatti, Everett, and Johnson's (2013, p. 156) terms, human social systems are clumpy and compact. To examine empirically just how "clumpy and compact," network analysts developed several measures—among them, transitivity and clique analysis are very important tools.

Transitivity. In social network contexts, transitivity means that, if node A is connected to node B, and node B is connected to node C, then node A and node C are also connected. In an undirected network, the triad is closed. Ancient proverbs attested to transitivity: "the friend of my friend is also my friend" and "the enemy of my enemy is my friend." In real life social networks, if node A is connected with node B, and node B is connected with node C, node A may or may not be connected with node C (that is, transitivity may not occur). Nodes A and C are more likely, but not certain, to connect if each already has a connection with node B. From a longitudinal perspective, if the A-B and B-C ties occur first, then the A-C tie is more likely to follow. For example, a Washington lobbyist is more likely to communicate with another lobbyist if each has a prior relation to a third party, such as a government agency or Congressional staff (Carpenter, Esterling, & Lazer, 2004).

The proportion of *transitive triplets* in a whole network can be measured by dividing the total number of *closed triplets* (three nodes connected by three ties) by the total number of *connected triplets* (the sum of closed triplets plus *open triplets*—triplets connected only by two ties). The ratio, which ranges from 0 when no triplets are closed to 1 when all triplets are closed, indicates the proportion of paths of length two that are closed. In other words, it's the probability that a connected triplet is transitive. Transitive triplets can be analyzed at both the ego level and the whole-network level. To illustrate, Figure 4.11 is an undirected graph of friendship relations among six girls. The network has a total of 15 triplets, identified by taking each person in turn and finding all pairs to which she is connected (e.g., Na is the go-between member of three triplets, Li-Na-Su, Li-Na-Yi, and Su-Na-Yi). But, only one of Na's three triplets is closed (Su-Na-Yi), so the proportion of her triplets that are transitive is 0.33. Table 4.7 lists all 15 triplets, grouped by their egos, classifies them by type (open or closed), and reports the proportion of each girl's triplets that are transitive. For the whole network, the probability that a triplet drawn at random will be closed—that is, that friends of friends are also friends—is .40.

Cliques. Network *cohesion* refers to numerous, intimate relations among members embedded in a social group or tight social circle. A cohesive subgroup consists of actors connected through many direct, reciprocated choice relations that enable them to share information, create solidarity, and act collectively. Many direct contacts among all subgroup members, combined with few or no ties to outsiders, dispose a group toward homogeneity of thought, identity, and behavior. Examples of cohesive groups include religious cults, terrorist cells, criminal gangs, military platoons, sports teams, craft occupations, and work teams. The term clique ("kleek" or

Figure 4.11 A Six-Actor Network

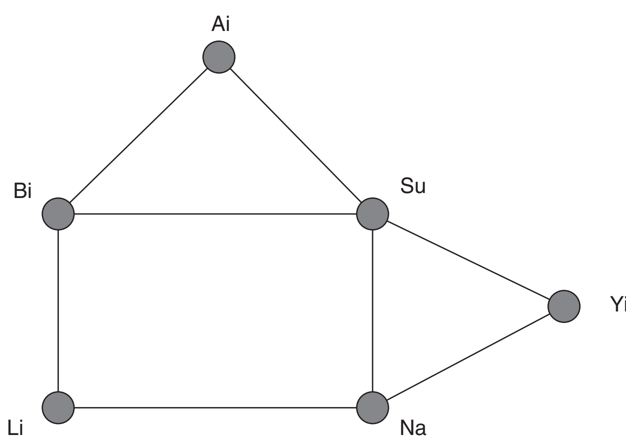


Table 4.7 Proportion of Transitive Triplets in Figure 4.11

Actors	*Triplet*	*Type*	*Proportion Transitive*
Ai	Bi-Ai-Su	Closed	1.00
Bi	Ai-Bi-Su Ai-Bi-Li Li-Bi-Su	Closed Open Open	0.33
Li	Bi-Li-Na	Open	0.00
Na	Li-Na-Yi Li-Na-Su Su-Na-Yi	Open Open Closed	0.33
Su	Ai-Su-Bi Ai-Su-Yi Ai-Su-Na Bi-Su-Yi Bi-Su-Na Na-Su-Yi	Closed Open Open Open Open Closed	0.33
Yi	Su-Yi-Na	Closed	1.00
Total	—		0.40

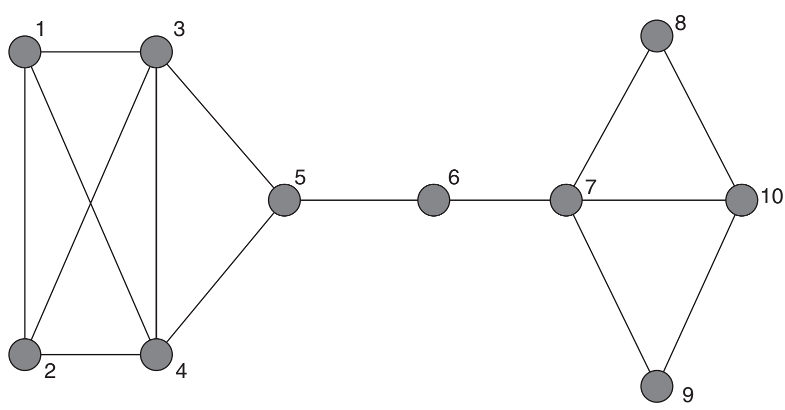
"klick") has passed into everyday language, referring to the high-status in-crowds of schools, churches, and clubs. The concept acquired specificity in network analysis, using measures of density and path length. By investigating subgroup structures, clique analysis enables researchers to understand how cohesion benefits or harms group members, for example, by affecting perceptions of friend support or belonging (Falci & McNeely, 2009; Martí, Bolibar, & Lozares, 2017) and by shielding corporate elites from external shareholder pressures (Benton, 2017). Often sociological concepts such as group, cluster, circle, gang, faction, and clique are used interchangeably without rigorous distinctions (Borgatti, Everett, & Shirey, 1990). Summarizing the vast literature on subgroups in social network studies, Wasserman and Faust (1994, p. 251) extracted four general properties that characterize cohesive subgroups: mutuality of ties, reachability of subgroup members, frequency of ties among members, and relative frequency of ties among subgroup members compared with nonmembers. These characteristics lay the foundation for operationally defining cliques and related measures of network subgroups.

A *clique* is defined in an undirected network as a maximum subset of three or more nodes in which every member of the subset is connected directly to every other member. The word "maximum" means no other node in the whole network can be added to the subset while preserving the property that every node is connected to every other node (Newman, 2010). Cliques may overlap, meaning that two or more cliques may have some nodes in common. Thus, the graph in Figure 4.9 has two cliques: the Chana-Ahmed-Fara clique and the Chana-Elaf-Dawud clique. Both cliques share one member: Chana. In contrast, the absence of direct ties between Chana and Ghada and between Dawud and Fara means that the four-actor subgroup Chana-Fara-Ghada-Dawud is not a clique.

Just as cliques can share common nodes, nodes can co-participate in cliques. Figure 4.12 shows the structure of co-membership among nodes in cliques. Four cliques exist in the figure: Clique 1-2-3-4, Clique 3-4-5, Clique 7-8-10, and Clique 7-9-10. Subgroup 1-2-3 is not a clique because it can add node 4 while still preserving its property of direct connections among all members (in other words, subgroup 1-2-3 is not maximal). Likewise, subgroups 1-3-4, 2-3-4, and 1-2-4 are not cliques because they are not maximal. At the clique level, Cliques 1-2-3-4 and 3-4-5 share two nodes: 3 and 4. Cliques 7-8-10 and 7-9-10 share two nodes: 7 and 10. And at the nodal level, nodes 3, 4, 7, and 10 each co-participate in two cliques.

The extent to which a pair of nodes co-participate in or are co-member of cliques can suggest their proximity. Two nodes co-participating in many cliques are close, as opposed to two nodes co-participating in no cliques. Following this rationale, network analysis programs can generate a clique co-membership matrix of all pairs of nodes in a network. The matrix is basically valued and undirected, with the off-diagonal *i* and *j* cell entries displaying the number of cliques shared by a dyad and diagonal entries showing the total number of cliques with which a node (*i*) is affiliated. The co-membership matrix can then be analyzed by hierarchical clustering and multidimensional

Figure 4.12 Co-Memberships of Cliques



scaling methods to visualize the closeness or the proximity among of all nodes in the network (see Borgatti, Everett, & Johnson, 2013, Chapter 11).

The utility of clique analysis is limited by its highly stringent requirement. Consequently, several alternative measures of subgroup cohesion are available, such as k-plex, k-core, k-clique, and k-components. A k-plex of size n is defined as a maximal subset of n nodes within a network such that each node is directly connected to at least $n - k$ of the others. So, when $k = 1$, the 1-plex is the same as a clique because every node in the subgroup must be directly connected to all the other nodes in the group. When $k = 2$, the 2-plex relaxes the requirement that each node in the subgroup must be directly connected with $n - 2$ nodes. As mentioned previously, the subgroup Chana-Dawud-Ghada-Fara in Figure 4.9 is not a clique because it lacks direct Dawud-Fara and Chana-Ghada connections. However, that subgroup is a 2-plex because every node is directly connected with two other nodes in the group ($n - 2 \rightarrow 4 - 2 = 2$). We refer readers to Newman (2010, Chapter 7) for discussions of k-core, k-clique, and k-component measures.

4.5. Whole Networks: Size, Density, Centralization

At the whole network level of analysis, several emergent properties characterize a network's structure. The most obvious one is *size*, the number of nodes in a network. Many social network analyses examine a handful or tens of nodes. The nodes might be people involved in friendship, collaboration, romantic relations, or advice-seeking. Other nodes may be collective entities such as corporations, work groups, teams, parties, communities, and nation states, engaging in a variety of partnership or competitive relations. Certainly, the number of nodes need not be limited to hundreds or even thousands of entities, as very large networks number in the millions or billions of nodes, for example, hyperlinked pages on the World Wide Web.

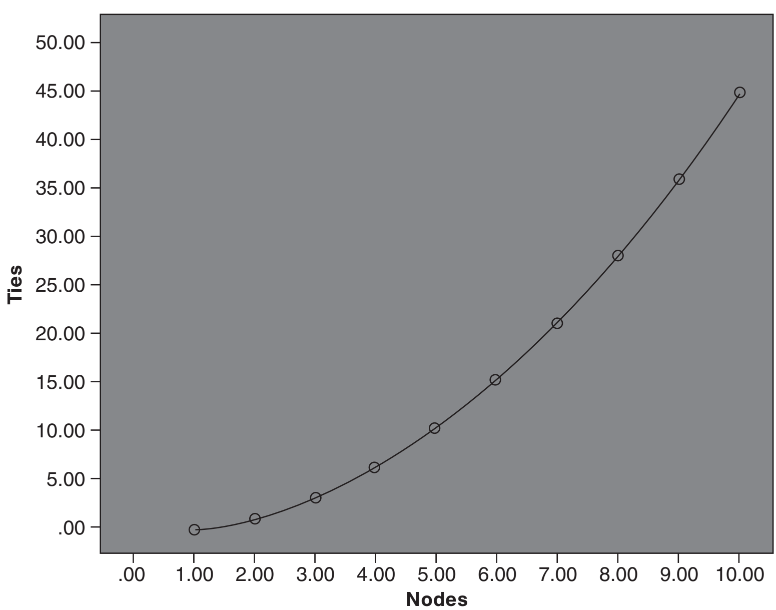
However, analyzing large networks with hundreds or thousands of nodes affords challenges to statistical processing or even computer capacity. Figure 4.13 illustrates that the relation between number of nodes and number of undirected dyads is exponential rather than linear (for directed ties, the relation is more steeply exponential). For a single node, the number of dyads = 0. For two nodes, only one dyad exists; three nodes have 3 dyads; four nodes have 6 dyads; for 10 nodes, 45 dyads must be analyzed; and so on. For networks with hundreds of thousands of nodes, dyadic ties become astronomical, requiring fast algorithms and efficient statistical processing. The good news is that recent developments in data mining and processing in the disciplines of Big Data facilitate social network analysis of huge networks. Social scientists, engineers, and mathematicians are working together to solve analytic issues of large networks with nodes numbering in the millions. Prominent businesses, such as Google, Walmart, and Amazon, as well as the U.S. government, possess techniques able to store, process, and analyze giant datasets, many of which are relational data about the social network activities among tens and hundreds of millions of nodes.

At the whole network level, *density* is an important indicator of the extent to which a network's dyadic ties materialize among the maximum possible number of dyads. The following two formulas show the computation of density for binary graphs, with the first one applying to undirected graphs and the second applying to directed networks.

$$D = \frac{\textit{number of dyadic ties that are present}}{C_N^2} = \frac{2 \times \left(\textit{number of dyadic ties that are present}\right)}{N\left(N-1\right)} \tag{4.8}$$

$$D = \frac{\textit{number of dyadic ties that are present}}{P_N^2} = \frac{\textit{number of dyadic ties that are present}}{N\left(N-1\right)} \tag{4.9}$$

Figure 4.13 Exponential Relation Between Number of Nodes and Number of Dyads



N is the total number of nodes in the network. To illustrate its application, we apply formula 4.8 to Figure 4.5, the network of trade among five nations. Density of this small network is $\left(\frac{2\times 5}{5\times 4}=0.50\right)$, which means that, out of the maximum possible number of dyadic pairs, half occur. Next, we apply formula 4.9 to Figure 4.10, a directed graph of seven nodes. The density is $\left(\frac{9}{7\times 6}=0.214\right)$, indicating that only one fifth of the possible directed ties occur in this directed network.

Density ranges from 0.00 to 1.00, indicating two extreme situations: 0.00 means no one in the network is connected to anyone else; 1.00 means everyone is connected to everyone else. To illustrate those two extreme forms of network, we altered Figure 4.5 to produce two new graphs. Figure 4.14 shows a completely isolated network, in which no nation trades with any other nation, a terrible global network to live in. Figure 4.15 depicts the opposite, in which every nation directly trades with all other nations, an ideal form that's hard to find in real life. In this graph, the dotted lines denote the five connections that, when added to Figure 4.5, convert its 0.50 density to the 1.00 density in Figure 4.15. Most real-world networks are neither totally isolated nor totally saturated but fall somewhere in between. The closer the density is to 0.00, the closer the network resembles Figure 4.14. Conversely, the nearer that density is to 1.00, the more closely the network resembles Figure 4.15.

Figure 4.14 Network of Trade Among Five Nations: Complete Isolation

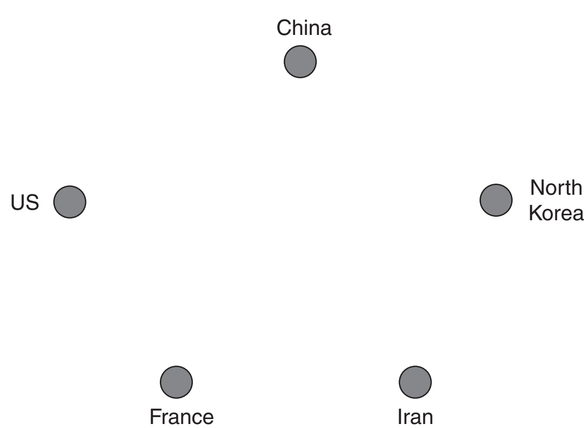
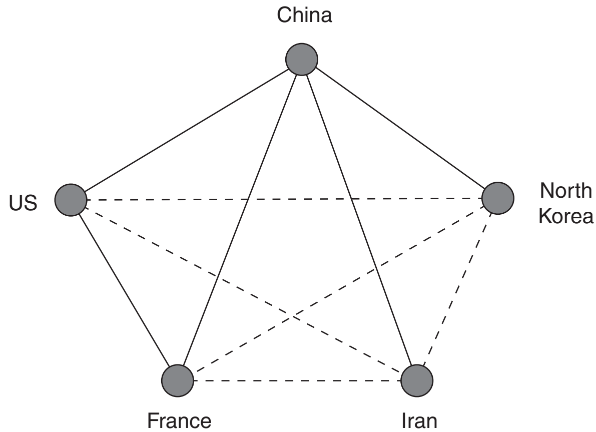


Figure 4.15 Network of Trade Among Five Nations: Total Saturation



Formulas 4.8 and 4.9 only apply to binary networks, not to valued graphs. Valued graphs have scores attached to the dyadic ties, complicating the computation and interpretation of network density. You could still apply the formulas to the binarized versions of a valued network, but doing so would sacrifice rich information conveyed in the values of the dyadic relations. More research is needed to offer satisfactory solutions to computing and interpreting densities in valued graphs.

Another important measure at the whole network level is centralization. *Centralization* differs from centrality, in particular, degree centrality. Degree centrality is a nodal-level measure, the extent a node is connected with other network nodes. Centralization calculates the extent to which the nodal degree centrality differs among all nodes. This formula shows the computation of a whole network's degree centralization:

$$C_D = \frac{\sum_{i=1}^{N}\left[C_D\left(N^*\right) - C_D\left(N_i\right)\right]}{\left(N-1\right)\left(N-2\right)} \tag{4.10}$$

The numerator sums the observed differences in degree centralities between the actor with the largest centrality and the other nodes. $C_D\left(N^*\right)$ denotes the degree centrality for the node with the highest degree. $C_D\left(N_i\right)$ centrality represents degree centrality for the N – 1 other nodes. $\sum_{i=1}^{N}\left[C_D\left(N^*\right) - C_D\left(N_i\right)\right]$ sums up the differences in degree centrality between the node with the highest degree centrality and all other nodes. The denominator measures the maximum possible sum of differences. This value occurs in a *star graph* (Figure 4.16), where one node (A) interacts directly with all the other nodes, but the others have only a tie to node A. Node A has the highest possible degree centrality ($N - 1$), whereas the other nodes each have degree centrality of 1; hence, the difference in centralities between this most central node (A) and any other node is $(N - 1) - 1 = N - 2$. Because this difference occurs ($N - 1$) times in the graph, the value of the denominator is $(N - 1)(N - 2)$.

The index of network degree centralization ranges between 0.00 and 1.00. Degree centrality has maximum centralization when one node has the highest possible centrality ($N - 1$) and all other nodes have degree centrality 1 (such as the star graph in Figure 4.16). The numerator then equals the denominator, and the index of group degree centralization equals 1.00. Conversely, if a network has nodes with the same number of degrees (the wheel graph in Figure 4.17), and hence degree centrality, the numerator will be 0. Hence, the network's degree centralization is 0.00. Therefore, the closer that network degree centralization is to 1.00, the more uneven or hierarchical are the nodes' degree centrality scores.

Size, density, and centralization are important indicators revealing different aspects of social network structures. Size measures the magnitude of a network, which can range from a few graduate students arguing in a classroom to millions of online profiles engaging in political debates. Density shows the level of saturation in a network, ranging from 0.00 (totally isolated network) to 1.00 (everybody is directly connected with everybody else). Centralization divulges the inequality among nodes in their degree centralities. Although 1.00 indicates the greatest hierarchy or inequality in

Figure 4.16 A Network of Six Nodes in Star Shape

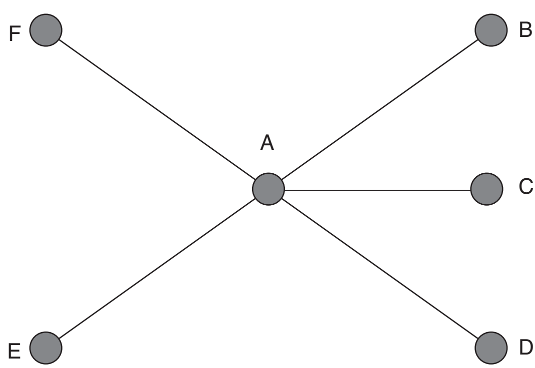
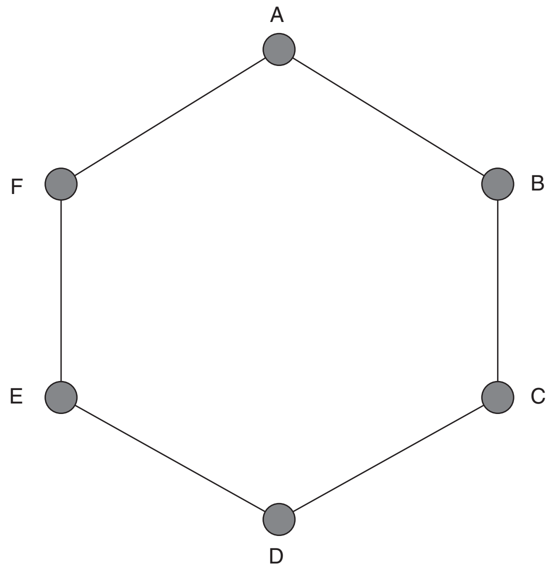


Figure 4.17 A Wheel Graph



degree centralities, 0.00 denotes an egalitarian or democratic structure. Three measures can be applied to uncover structural features of almost all types of social networks: classroom friendships, organizational leadership, or communications among a group of terrorists.

4.6. Structural, Regular, and Automorphic Equivalence

Social scientists are often interested in the *equivalence* of actors, in the sense of two or more actors having identical or very similar relations with others in a network. Structurally equivalent actors typically have a competitive, rather than a cohesive, relation. For example, two cabbage growers who market their produce to the same set of retailers are structurally equivalent and in stiff competition to sell their vegetables. Structurally equivalent actors are completely substitutable for one another. If one farmer were to withdraw from the cabbage network, she could easily be replaced by a structurally equivalent farmer, leaving the original network structure unchanged. Perfect substitutability in a social network often generates fierce competition to obtain favorable responses from other network participants (as is well known to grade-schoolers competing for their teacher's attention). Network scholars who use structural equivalence methods are generally interested in understanding competitive relations rather than group cohesion (Burt, 1992).

Sociological concepts of positions and roles draw from fundamental theories about social classifications that possess distinctive rights and obligations in relation to other categories. Robert Merton's (1957) role theory analyzed the role set, a social status involving an array of interconnected positions within a social system. For example, medical school students play the role of student in relation to their professors, but the med student role set also encompasses relations with patients, physicians, nurses, technicians, and clinic and hospital administrators. This section examines concepts and methods for identifying and measuring different types of equivalence in actors' ties within whole networks. An *equivalence relation* is a partitioning of subsets of nodes into mutually exclusive and exhaustive classes, where the members of an equivalence class (position) are equivalent to one another, whereas members of different equivalence classes are nonequivalent (Wasserman & Faust, 1994, p. 466). Three important equivalence properties, denoted by ≡, are

- Symmetry: $i \equiv j$ if and only if $j \equiv i$
- Reflexivity: $i \equiv i$
- Transitivity: if $i \equiv j$ and $j \equiv k$, then $i \equiv k$

Methods for finding equivalence seek to map the actors from an initial relational set onto a smaller number of equivalence classes. Network researchers developed several approaches to identifying equivalent roles and positions (Borgatti & Everett, 1992; Everett, 1985; Everett, Boyd, & Borgatti, 1990; Faust, 1988; Pattison, 1988). The three types of equivalence examined in this section are, in order of decreasing restrictiveness: structural, automorphic, and regular equivalence.

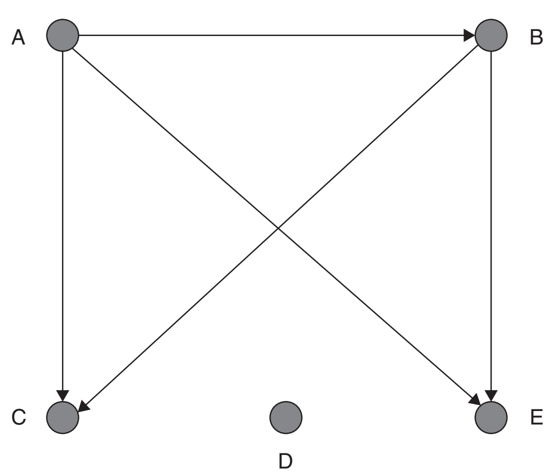
Structural Equivalence. Network analysts are often interested in reducing the complexity of large social systems to simpler structures, whose key features may be more readily grasped. A useful approach is to identify subsets of nodes that are equivalent, then group them together and treat them as jointly occupying a single position in relation to nodes occupying other equivalent positions. In a directed binary graph of a whole network of g actors, two actors are perfectly *structurally equivalent* in a relation if they have exactly identical patterns of ties sent to and received from all the other $g - 2$ actors in the network. More precisely, nodes i and j are structurally equivalent if, for all nodes k in the network (but not including i or j), node i sends a tie to node k, if and only if j also sends a tie to k, and node i receives a tie from k if and only if j also receives a tie from k (Wasserman &

Faust, 1994, p. 356). For multiple networks, this condition must hold precisely in each of the *R* relations for the two nodes to be structurally equivalent. The presence or absence of directed ties between nodes *i* and *j* is irrelevant to determining whether they are structurally equivalent. Rather, their structural equivalence depends only on their patterns of relations with the *g* – 2 other network nodes.

Undirected binary graphs make no distinction between the senders and receivers of relations. Extending the preceding definition of digraph structural equivalence to an undirected graph, actors *i* and *j* are structurally equivalent if, for all other actors *k*, *i* has a tie with *k* if and only if *j* has a tie with *k*. Structural equivalence can also be applied to valued graphs, in which ranking scales rather than binary values measure the ties between nodes. Strictly speaking, two valued-graph nodes are perfectly structurally equivalent only when both have exactly identical values for every tie to all other nodes.

In the digraph in Figure 4.18, the pair of nodes A and B is structurally equivalent because each sends relations to nodes C and E, neither sends a relation to node D, and A and B both do not receive relations from the other nodes. The presence and absence of directed ties between A and B are not included when assessing their structural equivalence. Similarly, nodes C and E are structurally equivalent because both receive relations from A and B and each sends relations to no one. Which nodes are structurally equivalent to D; why?

Figure 4.18 Digraph of a Five-Node Network for Structural Equivalence



The preceding definitions of structural equivalence are usually too rigorous for practical applications in empirical network analyses. Real network data rarely contain any dyads that meet such stringent standards. However, some nodes may be approximately structurally equivalent, in the sense that their patterns of relations with the other nodes are highly similar to one another although not identical. To capture such approximations, researchers use measures of relational similarity rather than applying a rigid, all-or-nothing requirement of perfect structural equivalence. The more similar two nodes are in their respective connections with all the other nodes, the greater is their structural equivalence.

Structural equivalence based on relational similarity for a network dyad only requires that their patterns of present and absent ties to and from the other actors be highly similar. Assuming a binary digraph, two structurally equivalent actors will have entries in the corresponding rows and columns of the sociomatrix that closely resemble one another. Operationalizing this criterion, Burt (1978) proposed Euclidean distance as a measure of the structural equivalence of actors i and j:

$$d_{ij} = \sqrt{\sum_{k=1}^{g-2}[\left(x_{ik} - x_{jk}\right)^2 + \left(x_{ki} - x_{kj}\right)^2]}\left(i \neq j \neq k\right) \tag{4.11}$$

where d_{ij} is the Euclidean distance between actors i and j and the x's are the values (either 1 or 0 for binary relations) in the matrix (the first subscript denotes the row and the second subscript the column of a matrix cell). Because d_{ij} is the positive square root of the sum of two squared difference terms, every $d_{ij} \geq 0$. If actors i and j have exactly identical ties to all others, all the differences $(x_{ik} - x_{jk})$ and $(x_{ki} - x_{kj})$ are zero; therefore, two perfectly structurally equivalent actors have Euclidean distance $d_{ij} = 0$. But, in most empirical cases, observed values of d_{ij} are greater than zero; hence, actors i and j are relationally similar to some varying extent. Euclidean distance is inverse to actor similarity and hence to structural equivalence: The larger the d_{ij}, the less the structural equivalence of actors i and j. In other words, Euclidean distances actually measure the dissimilarities between pairs of actors.

To illustrate how to compute Euclidean distances for a dyad, Figure 4.19 and Table 4.8 depict a five-node network structure in digraph and matrix forms, respectively. Figure 4.19 shows that nodes A and B are perfectly structurally equivalent because they both have direct connections to C and D and no tie to E. In contrast, nodes C and D are not structurally equivalent because, despite each receiving ties from nodes A and B, D also sends a tie to node E but C does not. Using the binary values from Table 4.8, the Euclidean distance between nodes A and B is the following:

$$d_{AB} = \sqrt{\begin{matrix}[(x_{AC}-x_{BC})^2+(x_{CA}-x_{CB})^2+(x_{AD}-x_{BD})^2+ \\ (x_{DA}-x_{DB})^2+(x_{AE}-x_{BE})^2+(x_{EA}-x_{EB})^2]\end{matrix}}$$

$$d_{AB} = \sqrt{\begin{matrix}[(1-1)^2+(0-0)^2+(1-1)^2+ \\ (0-0)^2+(0-0)^2+(0-0)^2]\end{matrix}} = 0 \tag{4.12}$$

$d_{AB} = 0$ indicates that nodes A and B are perfectly structurally equivalent. *Can you show that the Euclidean distance between nodes D and E is* $\sqrt{3}$ *= 1.73*?

Computing Euclidean distances in undirected binary graphs is simpler because the formula makes no distinction between senders and receivers of relations:

$$d_{ij} = \sqrt{\sum_{k=1}^{g-2}(x_{ik}-x_{jk})^2}\ (i \neq j \neq k) \tag{4.13}$$

When multiple relations are present in the network, the Euclidean distance computation involves summing squared differences for a dyad across all R relations:

$$d_{ij} = \sqrt{\sum_{r=1}^{R}\sum_{k=1}^{g-2}\left[(x_{ikr}-x_{jkr})^2+(x_{kir}-x_{kjr})^2\right]}\ (i \neq j \neq k) \tag{4.14}$$

In contrast to Euclidean distance as a measure of dissimilarity, Pearson's correlation coefficient (r_{ij}) directly measures relational similarity, with higher values of dyadic correlation indicating greater structural equivalence. We discuss the use of correlations in blockmodel analysis in Chapter 5.2.

The definitions and formulas of structural equivalence are usually too rigorous for practical applications in empirical social network research. Real networks rarely contain any dyads that meet such stringent standards (i.e., where Euclidean distance = 0). However, some nodes may be approximately structurally equivalent, in the sense that their patterns of relations to the g – 2 other nodes are highly similar but not identical. To capture such approximations, researchers use measures of relational similarity rather than applying a rigid, all-or-nothing requirement of perfect structural equivalence. The more similar two nodes are in their respective connections with all other nodes, the greater is their structural equivalence. Structural equivalence is an important measure for dyads because it reveals positional similarity, hence, the competition between two nodes.

Figure 4.19 Digraph of a Five-Node Network for Structural Equivalence

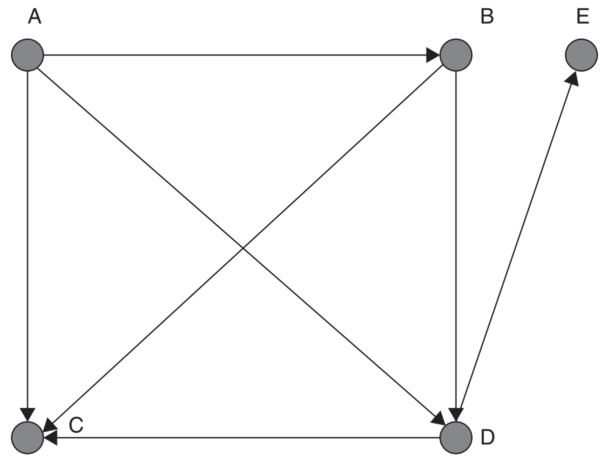


Table 4.8 Matrix of a Five-Node Network for Structural Equivalence

	A	*B*	*C*	*D*	*E*
A	0	1	1	1	0
B	0	0	1	1	0
C	0	0	0	0	0
D	0	0	1	0	1
E	0	0	0	0	0

The most common application of structural equivalence in whole network analysis involves inputting a binary directed or undirected adjacency matrix. The output is a squared matrix whose cell values indicate the structural equivalence of every dyad in the network. Thus, the output matrix is valued and undirected because structural equivalence between each pair is independent of the order of the pair. Chapter 5.2 discusses some applications.

Automorphic and Isomorphic Equivalence. Isomorphic and automorphic equivalence are such closely related concepts that some researchers treat them as interchangeable (Borgatti & Everett, 1992). However, isomorphic equivalence applies to two graphs, whereas automorphic equivalence describes the relational properties of social actors within one graph. Two graphs exhibit structural *isomorphism* if a one-to-one mapping of the nodes from one graph onto the second graph preserves all the nodes' adjacency relations (i.e., the same indegrees and outdegrees). In other words, if two nodes are connected in the first graph, then the corresponding two nodes in the second graph must also be connected in the same way (Borgatti & Everett, 1992, p. 11). Every graph is isomorphic with itself, which is called *automorphism,* a one-to-one mapping of nodes back onto themselves. Two actors are *automorphically equivalent* (jointly occupy the same position) if and only if they are connected to corresponding other positions (but *not* to identical nodes). Automorphic equivalent nodes have identical graph theoretic properties, such as centrality, ego density, and clique size (Borgatti & Everett, 1992).

Automorphic equivalence relaxes the structural equivalence requirement that actors in the same position have identical or very similar ties with the same set of other actors. Instead, automorphic equivalence identifies actors as jointly occupying a position if they have identical ties with different sets of actors that play the same role in relation to that position. To use a familiar example, for two professors to occupy a structurally equivalent position, both must teach the identical set of students, which is virtually impossible. But, to occupy an automorphically equivalent position, the two professors need only teach different sets with the same number of students. The students occupy a second position, defined as persons taught by a professor position. The graphs in Figure 4.20 contrast these two types of equivalence, where directed lines from professors to students represent the ‘‘teach’’ relation. Although both graphs have two positions, automorphic equivalence better captures the idea that social roles involve generalized patterns of relations. To cite another well-known instance, in monogamous marriages, we expect the wife position to be jointly occupied by a set of women who are in nonplural marriages to the same set of men but who are each uniquely paired with a different husband!

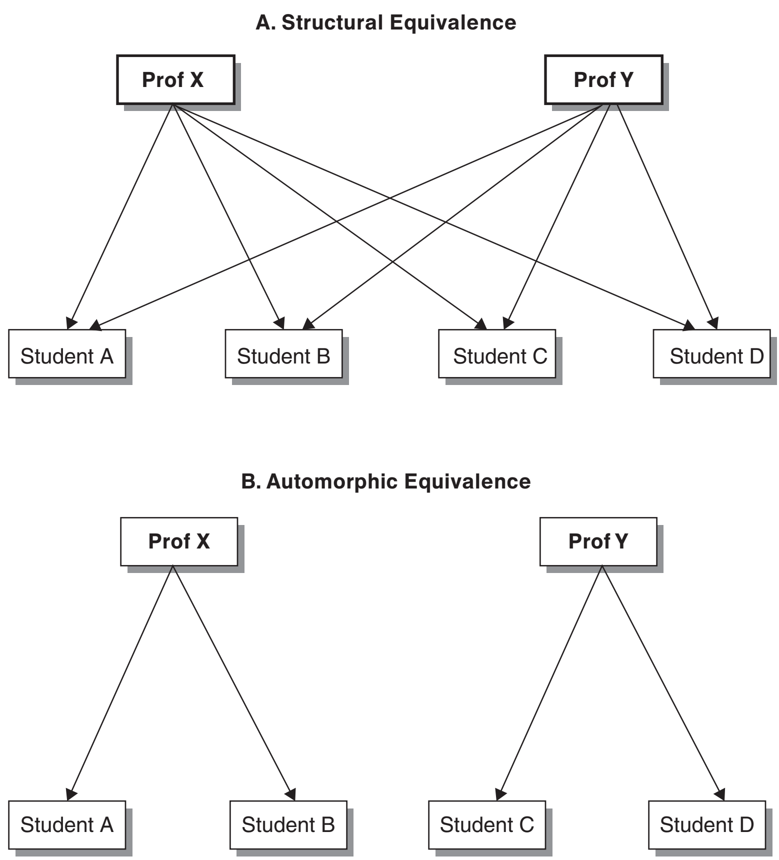
Structurally equivalent actors are also automorphically equivalent but not necessarily vice versa. Automorphically equivalent nodes are indistinguishable if the actor labels are removed from a graph. Thus, if points are substituted for the names in Figure 4.20B, the two subgraphs are indistinguishable. Borgatti and Everett (1992, p. 16) summarized the distinction:

> Abstracting a bit, we could say that in the structural equivalence approach, the network or labeled graph represents the underlying structure of a group; hence an actor's location in that structure represents his or her position in the group. In contrast, in the [automorphic equivalence] approach, the structure of interest is not the labeled graph

> itself, which is seen as the observed or "surface structure," but the structure of the surface structure, which is the unlabeled graph that underlies the labeled graph. It is the actor's location in this "deep structure," then, that represents his or her position in the group.

By relaxing the structural equivalence requirements, automorphic equivalence becomes very useful in facilitating empirical research corresponding to many social theories. Borgatti and Everett (1992) argued that several studies operationalizing theories using structural equivalence would be better analyzed using automorphic equivalence. For example, they addressed Burt's (1979) proposal to define the industries and sectors of an economy as sets of firms that produce similar types of goods and occupy a single

Figure 4.20 Structural Equivalence Versus Automorphic Equivalence



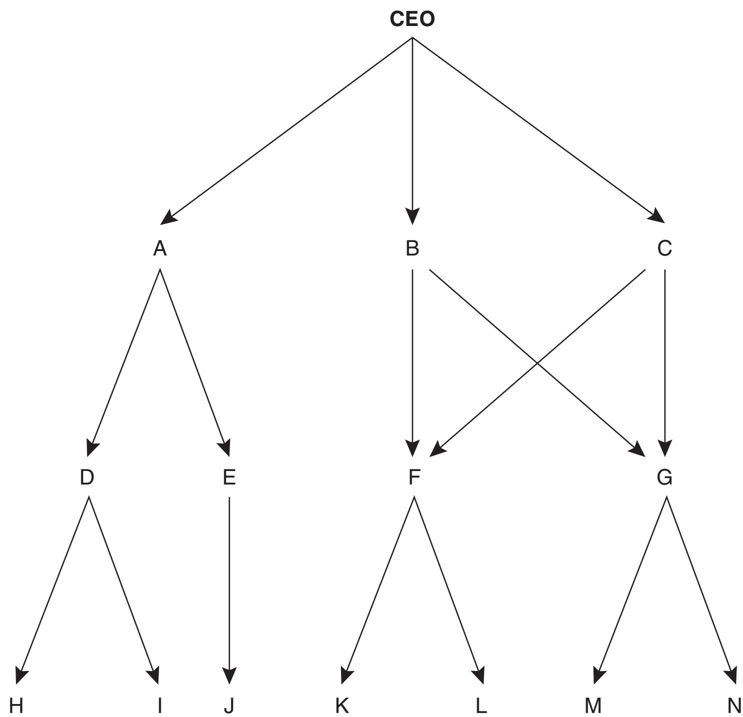
position within an interorganizational network. Borgatti and Everett (1992, p. 21) asserted that structurally equivalent firms, which by definition must buy from the same providers and sell to the same clients, hardly constitute recognizable sectors. But automorphically equivalent firms, which buy from similar vendors and sell to similar customers, may very well comprise meaningful industries and sectors.

Regular Equivalence. The least restrictive of the most commonly used forms of equivalence, regular equivalence requires neither structural equivalence's ties to identical actors nor automorphic equivalence's indistinguishable positions. Actors are regularly equivalent if they have the same kinds of relations with actors that are also regularly equivalent. Another way to conceptualize the idea is that, if a first actor occupying a position is tied to someone in a second position, then a regularly equivalent second actor must have an identical tie to someone else in a second position (White & Reitz, 1983, p. 214). All mothers with children are regularly equivalent, regardless of their numbers of offspring, as are all children who have mothers. In a hospital, the doctors are regularly equivalent in relation to their patients and nurses, even when the numbers of patients and nurses connected to the doctors vary. The generality of regular equivalence makes it perhaps the most important measure for sociologists attempting to capture social roles and positions.

Both automorphic equivalence and regular equivalence require that a pair of actors connect with the other actors who are structurally equivalent on the same relation. However, the distinction between automorphic and regular equivalence is sometimes ambiguous. Automorphic equivalence requires that unlabeled graphs be strictly substitutable for one another, but regular equivalence does not require a complete substitutability between subgraphs.

To demonstrate the difference, Figure 4.21 depicts a hierarchical organizational chart consisting of four vertical levels linked by supervisory relations. The CEO supervises three executive managers (A, B, C), who supervise four middle managers (D, E, F, G), who in turn supervise seven front-line employees (H through N). If we ignore the employees, then executive managers B and C are structurally equivalent because both have identical supervisory ties to the same middle managers (F and G). But A is not structurally equivalent to B and C because A supervises different middle managers. However, the three executives are regularly equivalent because each supervises the same number of middle managers (two apiece). If we consider all hierarchical levels, B and C are also automorphically equivalent because their subgraphs are substitutable for one another if the labels were removed. But A's subgraph cannot be substituted for B's or C's subgraphs because A's two middle managers supervise only three front-line employees, whereas both B's and C's subgraphs each have four employees.

Figure 4.21 Structural Equivalence, Automorphic Equivalence, and Regular Equivalence in an Organizational Hierarchy



Considering only the two lowest levels, none of the four middle managers are structurally equivalent because they all supervise different frontline employees. Instead, three of the middle managers (D, F, and G) are automorphically equivalent because their two-employee subgraphs are completely substitutable once the labels are removed (unlike E, who supervises only one employee). But all four middle managers meet the regular equivalence criterion by supervising at least one employee. Figure 4.21 demonstrates that structural equivalence is the most restrictive form, regular equivalence is the least restricted, and isomorphic and automorphic equivalence lie in between. Regular equivalence seems a very flexible method for identifying generalized social roles in networks, broadly defined as aggregate classes or categories of actors having similar structural relations with other positions in a social system (Faust, 1988, p. 315).

Chapter 5

ADVANCED METHODS FOR ANALYZING NETWORKS

This chapter discusses several advanced methods for analyzing social networks. In particular, we provide overviews of ego-nets, network visualizations, multimode network analysis, community detection algorithms, and exponential random graph models. In the concluding section, we offer speculations about future directions in social network analysis.

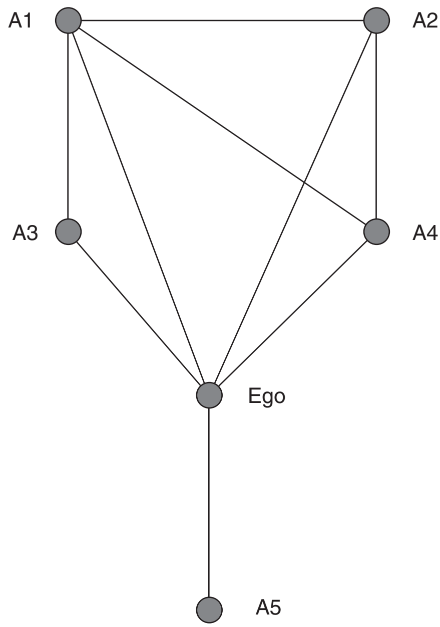
5.1. Ego-Nets

This section discusses the analysis of ego-net data, whose collection we discussed in Chapter 3.2. We also describe the connections and differences between ego-nets and whole network studies.

An ego-net is the network that forms around a specific social actor, which can be a person or a collective entity such as a corporation, voluntary association, or nation-state (Crossley, Bellotti Edwards, Everett, Koskinen, & Tranmer, 2015; Marsden, 2005; Marsden, 2011). The social actor is the ego, and its direct contacts are the alters. Linking ego and alters could be many types of relations, from emotional support between individuals to goods/information exchanges between companies to economic ties or wars between nations to strategic alliances among for-profit firms. Figure 5.1 presents a visualization of an ego-net, in which ego has five alters. The lines connecting ego to each alter and the lines among alters indicate some type of relation defined and measured in the study. Assuming that this relation is "discuss important matters" and the actors are persons, the diagram shows that ego named five alters with whom she discusses important matters. In addition, important matter discussion also occurs among some alter dyads; e.g., between Alters 1 and 2 and between Alters 1 and 3.

How does one identify egos, the particular social actors? The traditional sampling of random cases from a population applies: egos are those cases that are sampled to be representative of the population. In other words, egos are simply respondents in conventional sample surveys. In this regard, the ego-nets have clear advantage over whole network analyses by tapping into the general population. Conversely, although whole network analysis is ideal for dissecting a particular type of network structure among a finite set of people/actors (such as a friendship network among students in a classroom, information exchange, or strategic partnership among IT

Figure 5.1 An Ego-Net of Important Matters Network



companies), the sampling in whole network analyses is rarely random because it either encompasses an entire population or the sampling distribution is unknown (Yang, Keller, & Zheng, 2016, Chapter 4). Although conventional inferential statistics such as *t*-tests and *F*-ratio tests are readily applicable to analyzing the representativeness of egos, they are not appropriate to analyzing whole network data. Scholars have developed new and innovative inferential methods for whole network analyses (Lusher, Koskinen, & Robins, 2012), which we discuss subsequently.

After ego-net data are collected, the remaining question is how to analyze them. Relatively few methods are readily available. The simplest and most straightforward measure is ego-net size. For example, the average size for egos/respondents in the 1985 General Social Survey (GSS) core discussion network was 2.94, but that mean dropped to 2.08 in the 2004 GSS (McPherson et al., 2006). In another study, Fischer (1982) used nine name generators to elicit a wide range of alters, from 2 to 65 from the respondents, with a mean 18.5 alters. His name generator items include loan borrowing, socializing, information sharing related to jobs, hobbies, and discussing personal issues.

Another measure of network features for an ego-net is density, which reveals the extent to which the alters are directly connected to one another. Importantly, the ego-net attached to an ego is conceptually a whole network. The only difference between ego's whole network and the other whole networks is that ego has direct connections to everyone by design. So, the calculation of ego-net density requires ignoring ego's ties to alters and considering only the ties among alters. Figure 5.2 duplicates the graph in Figure 5.1 but omits ego's ties to each alter. Applying Formula 4.8 from Chapter 4 to this graph produces a density of 0.40 (4 ties that are present/ 10 maximum possible ties among the five nodes). Density is an important parameter for datasets with ego-net information. For example, in the 1985 GSS, the mean density of ego-nets identified by the name generator "discuss important matters" was 0.60. In the 2004 GSS, the mean density of ego-nets increased to 0.66 (McPherson et al., 2006).

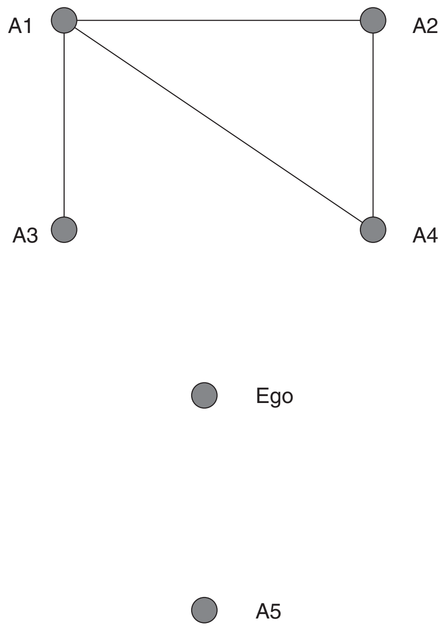
The third important measure of ego-net data is heterogeneity. Calculation of heterogeneity differs depending on the type of variable. For example, continuous variables such as age, income, or education can use standard deviations as heterogeneity measures. However, categoric variables, such as race and gender, must rely on the Index of Qualitative Variation (IQV) to divulge their heterogeneity. The following two formulas are for standard deviation and IQV, respectively:

$$s = \frac{\Sigma\left(X_i - \bar{X}\right)^2}{N-1} \tag{5.1}$$

$$IQV = \frac{1-\sum_{i=1}^{k} P_i^2}{\left(K-1\right)/K} \tag{5.2}$$

In Formula 5.1, s is the standard deviation, X_i is an alter's value on the variable, $\bar{X}$ is the mean score of the alters, and N is the total number of alters. In Formula 5.2, K is the total number of categories in the nominal variable and p_i is the proportion of the category i in the entire sample. To illustrate, if ego names three friends with ages of 79, 25, and 47, the standard deviation of its age heterogeneity would be 27.15. If one of the alters is white, a second is African American, and a third is Hispanic, the IQV for its racial composition is 1 $\frac{1-\left(1/3^2+1/3^2+1/3^2\right)}{2/3}=1.$ Interpretation of standard deviation and IQV is different: the larger the standard deviation, the larger the heterogeneity; whereas the IQV ranges from 0 to 1, with values close to 1 suggesting even distribution cases of all the categories

Figure 5.2 Calculating Density for an Ego-Net



(hence, great heterogeneity) and values close to 0 indicating uneven distribution (hence, low heterogeneity or high homogeneity). Again, each ego would have a heterogeneity measure of its alters' attributes (age, education, income, race, and gender). Aggregating the heterogeneity across the entire sample would produce important indicators such as means and standard deviations for each variable. For example, the mean age heterogeneity of the 1985 and 2004 GSS was 10.35 and 10.34, respectively, unchanged across 2 decades. The IQVs for race in 1985 and 2004 were 0.05 and 0.09 respectively, indicating low heterogeneity (high homogeneity) in the racial composition of Americans' core discussion networks. In contrast, the IQVs for gender composition were quite high (0.67 for 1985 and 0.68 for 2004), suggesting persistently gender-balanced core discussion groups (McPherson et al., 2006).

Many texts on social networks separate ego-nets or egocentric networks from the whole network, as if the ego-net is a very special type, subsumed under whole networks. In reality, ego-nets are important topics in and of themselves (Crossley et al., 2015), and comparisons and contrasts with whole networks are very important but, in our view, too often receive insufficient attention and discussion.

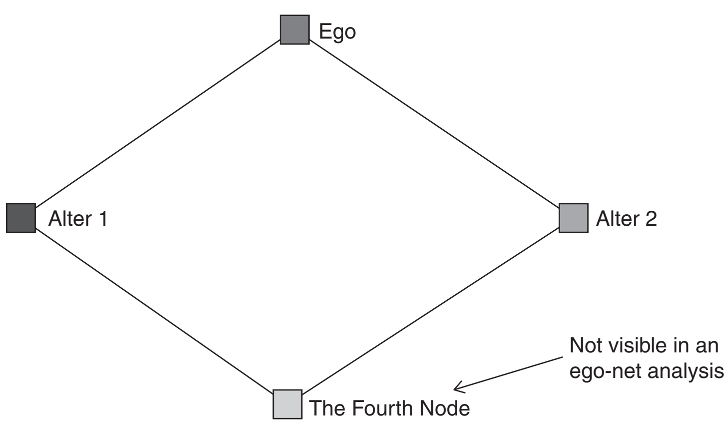
A first point of comparison is that each ego-net in egocentric network studies is basically a whole network, with the main difference between the two being the networks' sizes (apart from the obvious fact that an ego-net by design has an ego directly connected to a set of alters). Although the ego-net from each ego/respondent is commonly small, ranging from a couple of alters to five or six at most (the GSS allowed respondents to name up to 6 alters), the whole network can have tens, hundreds, or even millions of nodes in Big Data contexts. However, other than the size, the basic structure between the two is the same, and the set of methods used to analyze whole networks is also available to analyze each ego-net, and vice versa. An important advantage of ego-nets is that they are randomly sampled, so their network features can be generalized to the whole population. For example, the average IQV of racial composition for American's core discussion group remains small, implying persistent high homogeneity in American discussion networks. However, the same cannot be said for the whole network data, many of which derive from purposive sampling, that is, from nonrandom samples.

A second point of comparison is that, whereas ego-nets capture ego's diverse networks, whole network analysts typically seize one slice of ego's many networks and magnify it for whole analyses. An ego in modern societies interacts and forms social ties across many distinct social circles or domains, whose members (excluding ego) rarely encounter one another (White, 2008). For example, ego interacts with differing sets of alters in families, workplaces, neighborhoods, special interest groups (e.g., Bible studies, birdwatching), gyms, and local bars among numerous settings (Crossley et al., 2015). Although all those networks obviously involve ego, they generally do not overlap with one another, creating distinct social circles or domains. Ego's gym buddies may not know anything about his Bible study group. Although ego-nets would be able to capture variation across diverse domains, whole networks would be unable to accomplish that task. However, researchers interested in a particular type of network, such as Bible study groups, can extract them from various ego-nets, thus magnifying those networks for further investigation. Clearly, ego-nets and whole networks are two distinct research tools at a researcher's disposal. If a researcher's objective is to analyze diverse networks centered around individual actors, then using ego-nets is the best way to approach the matter. But, if the aim is to dissect a particular type of network, then whole networks are a better approach.

However, ego-nets can partially reveal a ego's structural positions in relation to its alters. Such partiality may result in inaccurate interpretation of ego's structural leverage. Here, the whole network can compensate for the deficiency of ego-nets, depicting a more accurate network structure between egos and their alters. Figure 5.3 shows such a scenario: ego is connected to two alters that are also both connected to a fourth node. If information is only available about the ego-net, the fourth node will not be known to the researcher, leading to a conclusion that ego occupies a "strategically advantaged" brokerage position between the two alters (Burt, 1992). Such "hidden" fourth nodes can be uncovered using a whole network design, showing the flawed inference about ego's alleged brokerage position. Such revelation underscores that researchers must be careful when studying ego-net data, not to distort ego's structural advantage by using incomplete network structures.

The last comparison between ego-net and whole network data is about the density of ego-nets. Although a random sample of egos from a large population ensures representativeness, it assumes that the ego-net's density is randomly distributed, which may not be true. In some special types of relations (teenager sexual activity networks, for example), power-law distribution or bandwagon effects are very pronounced; a few nodes may have very high numbers of alters, whereas most others in the population have only small numbers of partners. Such networks have distinctive structures centralized around a few dominant nodes, which may elude random sample surveys. Random sampling would be unlikely to select those hub nodes in the population. In contrast, whole network research designs would be able to capture and show such bandwagon structures with clear demarcations between central and peripheral players.

Figure 5.3 The "Hidden" Fourth Node



Source: From Crossley, N., Bellotti, E., Edwards, G., Everett, M. G., Koskinen, J., & Tranmer, T., 2015, *Social Network Analysis for Ego-Nets*, p. 22. London, England: Sage.

5.2. Visualizations: Clustering, MDS, Blockmodels

Images of networks are commonly created in social network studies to develop structural insights and to communicate those insights to others. Social network analysis experienced three distinctive periods of visual display innovation (Freeman, 2000, 2005). The initial stage began in the 1930s, when Jacob Moreno created hand-drawn, ad hoc sociograms to depict relations among actors such as schoolchildren (Moreno, 1934). This freestyle approach gave way to standard computational procedures for plotting the points and lines of a graph. A basic principle from the initial era of visualization is that spatial representations should preserve the underlying pattern of actor ties by depicting pairs that are socially closest in a data matrix as closest in a graphic image (Freeman, 2005). However, preserving precise proximities and distances among numerous actors in two- or three-dimensional visual displays is usually impossible, so researchers eventually developed methods for systematically simplifying and reducing the number of dimensions while still reflecting the original data patterns. The second phase of visualization, beginning around the 1960s, used mainframe computers and software to produce graphics automatically. In particular, network analysts made increasing use of the hierarchical clustering and multidimensional scaling (MDS) methods described in this section. The most recent phase, starting in the mid-1990s with the advent of the World Wide Web, high-speed computer networks and browsers, and widespread personal computers, opened new opportunities for large-scale visual displays of relational data, including animation of longitudinal network changes (for example, Christakis & Fowler, 2008; Healy & Moody, 2014; van den Elzen, Holten, Blaas, & Van Wijk, 2016).

Lacking space to cover all these developments in depth, we concentrate on explicating the fundamental ideas of three tools for displaying relations in network data: hierarchical clustering, MDS, and blockmodeling. Visual displays are useful for exploring social network data to uncover cohesive subgroups and to reveal how they relate to one another. In this process, complex network structures can be simplified by reducing their representation from many actors to a smaller number of *jointly occupied positions.* Structural equivalence methods in blockmodeling analyze a matrix of dyadic dissimilarities (Euclidean distances) or similarities (correlation coefficients) to identify blocks that are jointly occupied by sets of actors with either identical or very similar patterns of ties to others. Clustering and MDS, when applied to the same matrix, produce two- or three-dimensional diagrams. Blocks can then be visualized by drawing contiguity lines around members of each jointly occupied position.

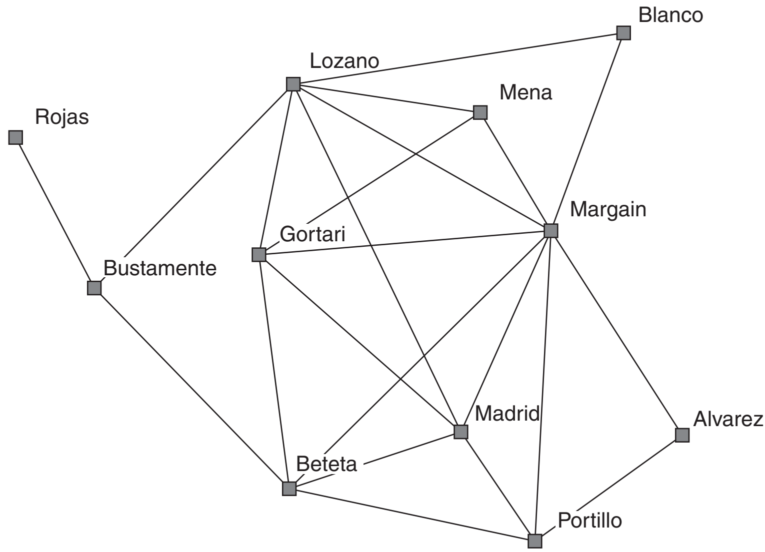
We illustrate these methods on a small dataset, the 1990 Mexican political elite network reported in Mendieta, Schmidt, Aastro, and Ruiz (1997). From 5400 members of the national government between 1920 and 1990, the authors identified "37 core actors, who have played a central role in Mexican politics after the 1910 revolution" (p. 35). A connection between a pair of actors represents any formal, informal, and organizational relation, for example, "common belonging (school, sports, business, political participation), or a common interest (political power)" (p. 37). The article displayed six diagrams showing the ties among core actors active at each decade from 1940 to 1990. Figure 5.4 displays an undirected graph of the 1990 network, when a third generation of 11 politicians had ascended to power. Three men successively held the presidency: José López Portillo (1976–1982), Miguel de la Madrid (1982–1988), and Carlos Salinas de Gortari (1988–1994), and the latter was the son of a fourth core member, Raúl Salinas Lozano. Of 55 possible ties between dyads, 22 occurred, for a network density of 0.40.

Clustering. *Hierarchical agglomerative cluster analysis,* or cluster analysis for short, partitions actors into subgroups (jointly occupied positions) whose members are perfectly or approximately structurally equivalent. Each actor is treated initially as a singleton cluster, and then clusters are successively joined until all actors are merged into a single remaining cluster. A *dendrogram*, also called a *tree diagram*, visually depicts this hierarchical sequence of merging clusters.

Clustering algorithms typically process a square $g \times g$ matrix of either Euclidean distances (d_{ij}) or correlation coefficients (r_{ij}), where g is the number of actors in a specific relation. (We discuss Euclidean distances in Chapter 4.6.) One comparative study of cluster analyses suggested that both measures produce very similar results, although cluster analyses using correlations may be somewhat easier to interpret (Aldenderfer & Blashfield, 1984, pp. 24–28). Multiple network relations *(R)* can be clustered simultaneously by constructing a data array with dimensions $g \times g \times R$ then computing Euclidean distances or correlations for every dyad across all R matrices before conducting the cluster analysis. Although hierarchical agglomerative clustering can be performed on both binary and valued data, for simplicity we focus on a symmetric binary matrix.

After computing a matrix by correlating all pairs of rows and columns, a cluster analysis proceeds to combine actors according to a threshold value, α, which serves as a ceiling, or upper bound, for the analyst to decide which actors belong in the same jointly occupied position (cluster) at a particular level of structural equivalence. Actors i and j jointly occupy a position only if $d_{ij} \leq \alpha$. Actors within one cluster have smaller social distances from one another (i.e., are more structurally equivalent) than

Figure 5.4 The 1990 Mexican Political Elite Network



from the actors occupying other clusters. The clustering algorithm proceeds incrementally, applying successively less-restrictive levels of α (i.e., higher values of α) to aggregate actors into positions until the entire network merges into a single all-inclusive cluster. Although hierarchical agglomerative clustering produces nonoverlapping clusters, the clusters are nested; that is, smaller clusters are subsumed within successively larger clusters at higher values of α (i.e., lower structural equivalence, less similarity, greater within-cluster distances). Ultimately, the researcher must decide which level of agglomeration (i.e., which value of α) provides the best substantive representation of the number of structurally equivalent positions in the network.

Researchers may choose from among three basic criteria for forming clusters: single link, average link, or complete link. At a given level of α, the single-link criterion merges two clusters into one cluster when their two closest actors have a distance less than α. Under a complete link, two clusters merge when the distance between every pair of actors is less than α. The average linkage option is a compromise, requiring that two merging clusters have an average distance among both sets of actors that is less than α. Empirical analysts reported advantages and disadvantages from using each option, which should prompt researchers planning to use cluster analysis to carefully consider these alternatives (Aldenderfer &

Blashfield, 1984, pp. 53–62). Complete link clustering appears to produce large numbers of homogeneous and tightly bound clusters, with a lower probability of "chaining," the formation of a single large cluster by successively adding one actor at a time (Burgin, 1995; Wasserman & Faust, 1994, p. 381).

Multidimensional Scaling. MDS is a method for estimating similarities of dyads in whole networks and visualizing the network's underlying relational structures (Borg et al., 2017; Hout et al., 2013; Kruskal & Wish, 1978; Young, 1987). MDS has facilitated research on such diverse topics as examining the relation of friendship ties and creativity (McKay et al., 2017), profiling and visualizing criminal networks (Park et al., 2012), visualizing political networks (Pfeffer, 2017), comparing interest group networks over time (Box-Steffensmeier & Christenson, 2015), mapping the flow of extreme financial episodes among global stock exchanges (Fernández-Avilés & Montero, 2016), and visualizing networks of cruise ship destinations in the Baltic Sea (Marcussen, 2017). The primary purpose of MDS is to detect meaningful underlying dimensions that reflect the similarities (proximities) or dissimilarities (distances) among network actors. As in cluster analysis, the typical input to MDS is a $g \times g$ matrix of either Euclidean distances (d_{ij}) or correlation coefficients (r_{ij}), where g is the number of actors in a particular relation. For most social data, a nonmetric solution is preferable because it assumes that only the ranks of the distances are known (in contrast to metric distances such as geographic mileage between cities). An MDS visual output is a plot, or social map, in which actors with smaller distances (greater similarities) between them are located closer in space than are actors with larger distances (greater dissimilarities). Although the MDS diagram coordinates can be estimated for multiple dimensions, most analysts display two- or three-dimensional maps.

The distances between actors displayed in an MDS map are related to, but not identical to, the similarity or dissimilarity values in the input matrix. Rather, they reflect the pairwise distances estimated from those data by the MDS program. A *stress* indicator measures the discrepancies across all pairs between the observed matrix and the computed matrix (Kruskal & Wish, 1978, pp. 23–30):

$$Stress = \sqrt{\frac{\Sigma\Sigma\left(f\left(x_{ij}\right)-d_{ij}\right)}{Scale}} \tag{5.3}$$

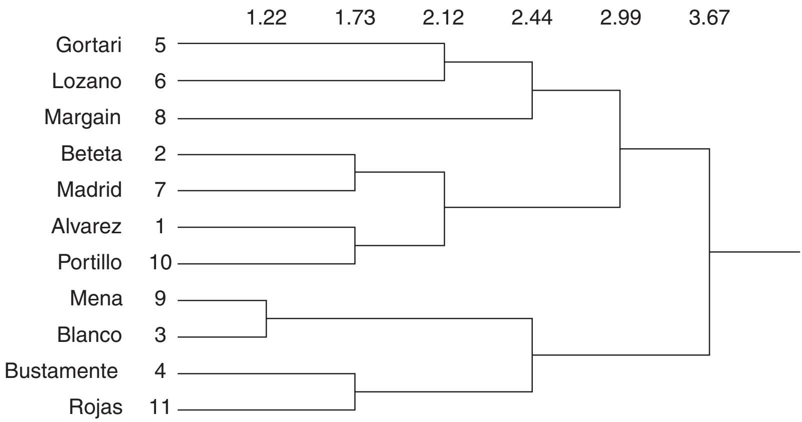
where $f\ (x_{ij})$ is a nonmetric, monotonic function of the input values (Kruskal & Wish, 1978, p. 29) and d_{ij} is the Euclidean distance between actors i and j as displayed in the map coordinates. *Scale* is a scaling factor that constrains the stress values between 0.0 and 1.0. If an MDS map perfectly reproduces the input data, then $f(x_{ij}) = d_{ij}$ for all i and j, and thus its stress would equal zero. Hence, the lower the stress value, the more closely an MDS spatial diagram represents the observed social distances among the network actors. For this reason, stress is also dubbed a badness-of-fit measure of MDS, in that stress values below 0.1 are considered an excellent fit, values between 0.1 and 0.2 are an adequate fit, and values above 0.2 are a poor fit (Kruskal & Wish, 1978, p. 52; Slez & Martin, 2007).

To illustrate how clustering and MDS can work together to reveal network positions, the hierarchical cluster dendrogram in Figure 5.5 and the MDS map in Figure 5.6 are both based on a matrix of Euclidean distances computed on the binary graph in Figure 5.4. The cluster analysis used the complete link option: Clusters merge when the distance among all their actors was less than α (these values appear on the horizontal axis of Figure 5.5). For example, Mena and Blanco formed the first cluster at $\alpha = 1.22$, followed by three more clusters at $\alpha = 1.73$. At $\alpha = 2.44$, the three clusters included all 11 politicians in the network. Eventually, all merged into a single cluster on the right side of the dendrogram.

On the MDS map in Figure 5.6, which has a stress value (0.087), indicating an excellent fit, we drew contiguity lines around subgroups of actors jointly occupying three clusters that emerge when $\alpha = 2.44$ in the clustering dendrogram. Although the choice of an a for identifying clusters is somewhat arbitrary, we recommend choosing a value that balances between cohesion and divisiveness among network actors. For example, at $\alpha = 1.73$, the result is five two-person clusters and a singleton. In contrast, had we chosen $\alpha = 2.99$, the result would be two large clusters. The first choice indicates greater fragmentation due to a low threshold for "being close," whereas the second implies great cohesiveness due to larger distances among actors in a cluster. Neither approach strikes us as informative as the middling criteria ($\alpha = 2.44$), which reveals a more balanced picture of the 1990 Mexican national elite's division and cohesion.

An extension is weighted multidimensional scaling (WMDS), which adds a component (weight) to the conventional MDS approach, to represent information about variation between matrices. Thus, WMDS generalizes the distance model by allowing multiple matrices to be systematically differentiated. For example, if each matrix corresponds to a different individual, WMDS uses the weight to portray differences in how those individuals think about or perceive relations. For this reason, WMDS is also called individual differences scaling (INDSCAL). Although we cannot elaborate here on WMDS, we encourage interested readers to consult Schiffman, Reynolds, and Young (1981, pp. 55–85) and applications to evaluate test score validity for English language

Figure 5.5 Hierarchical Cluster Dendrogram of Figure 5.4 Network

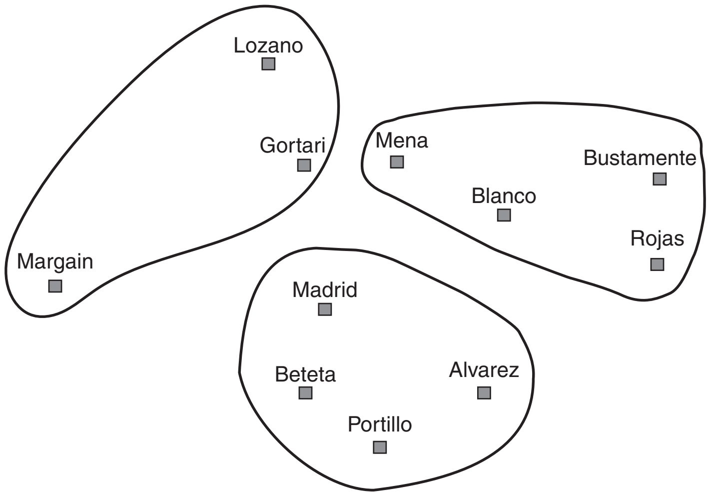


learners (Sireci, Han, & Wells, 2008) and to assess structural equivalence between German and English versions of a networking scale (Wolff, Schneider-Rahm, & Forret, 2011).

Blockmodels. Blockmodeling is a matrix algebraic method for sorting network actors into jointly occupied, structurally equivalent positions. Blockmodel methods were initially developed by Harrison White and his associates (Boorman & White, 1976; Schwartz, 1977; White et al., 1976). Since that groundbreaking work, researchers have fruitfully applied blockmodeling methods to various topics ranging from interorganizational networks (Knoke & Rogers, 1979) to diffusion of new technology (Anderson & Jay, 1985) to the roles of cities in the world system (Alderson & Beckfield, 2004) to international trade and diplomacy networks (NasehiMoghaddam & Ghazanfari, 2016). Methodological advances have extended blockmodel algorithms to dynamic stochastic blockmodels that can investigate change in social network structures over time (Ludkin, Eckley, & Neal, 2017; Olivella, Pratt, & Imai, 2018; Xu & Hero, 2014). Because space limitation prohibits an extended discussion on this large literature, we focus on fundamental issues of blockmodeling methods, implementation, and interpretation of outputs.

A *blockmodel* is the partition of a sociomatrix of *g* actors, in one or more relational networks, into two or more mutually exclusive subgroups called *blocks*. The term *block* refers to a square submatrix of structurally equivalent actors that have very similar, if not identical, relations with the actors occupying the other blocks. Blockmodeling is a data reduction technique that systematically searches for relational patterns in network data by

Figure 5.6 Multidimensional Scaling of Euclidean Distances in Figure 5.4 Network



regrouping actors and presenting condensed aggregate-level information. The outputs are permuted density and image matrices displaying the pattern of ties within and between the blocks for each type of relation. Blockmodeling can be applied to single or multiple relations, directed or undirected ties, and binary or valued graphs. We examine only a binary matrix and refer readers to an extended discussion of blockmodeling for both binary and valued graphs by Doreian, Batagelj, and Ferligoj (2005, pp. 347–360).

A blockmodel could be constructed a priori using theoretical principles, for example, by sorting the employees of different bureaucratic departments into separate blocks. However, the most common applications of the method are exploratory searches for empirical patterns in a relational dataset. Blockmodeling is often implemented using CONCOR (**Con**vergence of Iterated **Cor**relations), which is available in several social network analysis programs (Schwartz, 1977). CONCOR can operate by correlating pairs of rows, pairs of columns, or both paired matrix vectors simultaneously. For undirected matrices, the rows and columns are identical, so only one needs to be correlated. For the initial step, CONCOR calculates Pearson correlation coefficients for every pair of columns in a binary $g \times g$ sociomatrix. (To blockmodel R multiple relations, the separate $g \times g$ matrices are "stacked" [arranged as a single two-dimensional data array] for input to CONCOR.) These computations exclude the direct ties between each dyad, as their structural equivalence depends only on the pair's ties to the other $g - 2$ network actors. If actors i and j have exactly identical connections with all other actors, their correlation coefficient, r_{ij}, will equal 1.0. In contrast, if two actors have exactly opposite patterns of connections, their correlation will be –1.0. Almost always, empirical correlation coefficients fall somewhere between these extreme values. The result from the initial CONCOR step is a symmetric $g \times g$ matrix of correlations for every pair of actors, showing the extent of each dyad's structural equivalence. The second step, and all subsequent iterations, repeats this process of correlating pairs of columns in the correlation matrix produced by the preceding step. At some point, the correlation coefficients in every cell converge to either 1.0 or –1.0, at which point the iterations cease.

Next, CONCOR permutes the final correlation matrix into two homogeneous blocks. *Permutation* of a sociomatrix involves simultaneously rearranging both the rows and columns to bring together in adjacent portions those actors jointly occupying the same block. CONCOR's initial partition and permutation of the sociomatrix always yields two submatrices, not necessarily having equal numbers of actors, in which all the correlation coefficients among pairs of actors within each block equal 1.0 but all the correlations between the two blocks equal –1.0. Repeating these procedures, CONCOR can subdivide each of the two initial blocks into two more blocks, and so on. The network researcher must decide where to stop the division process, thus determining the ultimate number of blocks obtained.

Blockmodel analysis results in two forms of output: a density matrix and a corresponding image matrix. A *density matrix* is a $b \times b$ matrix whose cell values are the densities within and between the blocks, where b is the number of blocks. (Density is a proportion calculated by dividing the number of observed ties in a permuted submatrix by the number of possible ties.) An *image matrix* is also a $b \times b$ matrix, obtained from the density matrix by recoding each cell density to either 0 or 1. Two alternative criteria may be used to determine the image values: (1) any cell with no ties among its actors *(zero-block)* is recoded as 0 and any cell with at least one tie among its members *(one-block)* is recoded as 1 or (2) the researcher chooses a *density cutoff,* α (alpha), recoding all densities below this cutoff to 0 and all densities of α or higher to 1. The first option is an unrealistic standard because empirical densities of 0.0 seldom occur unless a network has very few relations, so almost all image matrices would consist only of 1s in every cell. The second option, using an α density cutoff to dichotomize the image values, is the most common practice. Researchers typically choose the density of the entire matrix as the cutoff value. However, because choosing an α value inevitably involves the researcher's judgment, selecting a particular value is vulnerable to the criticism of arbitrariness. In response, researchers should try to justify their choices on theoretical and empirical grounds rather than appealing solely to expediency (Scott, 1991, p. 136).

To illustrate blockmodeling, we again analyzed the 1990 Mexican political elite network. We entered the binary matrix into the CONCOR program, requesting two splits that resulted in the four-block partition, density, and image matrices in Table 5.1. For determining the 1-blocks in the image, a density cutoff $\alpha = 0.40$ or higher was used. The first block is a clique (Alvarez, Margain, and Portillo all have ties to one another), but none of the other three blocks has a sufficiently high density to warrant a "1" on the diagonal entries of the image matrix nor does the large second block have a link to the first block. The two tiny third and fourth blocks, one of which is a singleton, have some higher-density connections with members of the two larger blocks. Figure 5.7 displays an MDS analysis of the correlations in which the block members are inside contiguity lines. The MDS stress (0.153) indicates a barely adequate fit. Not surprisingly, the hierarchical cluster and blockmodel analyses produced different results, as the former is based on social distance measures and the latter of structural equivalence. The two visualization methods provide alternative interpretations of network positions, and researchers must decide which approach better answers their theoretical concerns.

5.3. Two-Mode and 3-Mode Networks

Two-mode and 3-mode networks consist of multiple types of entities and relations between them. Entities might be individuals, groups, organizations, nations, events, documents, and websites. In this section, we discuss both 2- and 3-mode data structures and illustrate the application of analytic methods with a simple toy dataset drawn from a real-world example. These principles could be generalized to *N* types of entities, but the analytic complexities render multimodal networks with many types of nodes empirically impractical.

Two-Mode Networks

Two-mode networks, sometimes called *affiliation networks* or *membership networks*, consist of two distinct sets of nodes, such as actors and events, where the relations among the entities in one mode are based only

Table 5.1 Blocked, Density, and Image Matrices of the 1990 Mexican Political Elite Network

Blocked Matrix

```
                      1                     1
                  1 8 0   7 4 3 9 5   2   6 1
                -------------------------------
   1    Alvarez |   1 1 |           |   |     |
   8    Margain | 1   1 | 1   1 1 1 | 1 | 1   |
  10   Portillo | 1 1   | 1         | 1 |     |
                -------------------------------
   7     Madrid |   1 1 |         1 | 1 | 1   |
   4 Bustamente |       |           | 1 | 1 1 |
   3     Blanco |   1   |           |   | 1   |
   9       Mena |   1   |         1 |   | 1   |
   5    Gortari |   1   | 1     1   | 1 | 1   |
                -------------------------------
   2     Beteta |   1 1 | 1 1     1 |   |     |
                -------------------------------
   6     Lozano |   1   | 1 1 1 1 1 |   |     |
  11      Rojas |       |   1       |   |     |
                -------------------------------
```

Density Matrix

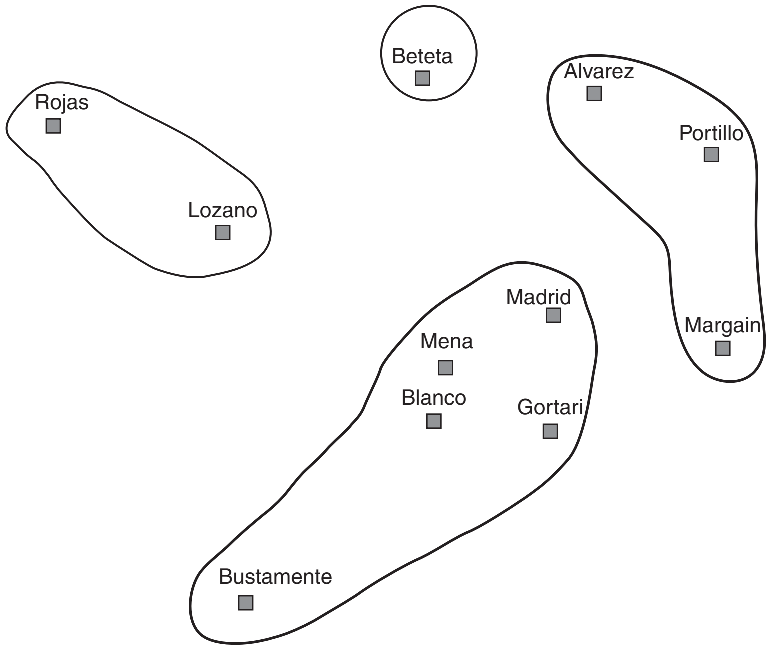
	1	2	3	4
Block 1	1.000	0.333	0.667	0.167
Block 2	0.333	0.200	0.600	0.600
Block 3	0.667	0.600	0.000	0.000
Block 4	0.167	0.600	0.000	0.000

Image Matrix

	1	2	3	4
Block 1	1	0	1	0
Block 2	0	0	1	1
Block 3	1	1	0	0
Block 4	0	1	0	0

on their connections to the second mode (Wasserman & Faust, 1994, pp. 291–343). Direct relations among the entities within a mode may be unavailable, meaningless, or ignored. A classic example is the Southern Women network, consisting of 18 women and their attendance at 14

Figure 5.7 Multidimensional Scaling of Correlations in Figure 5.4 Network



informal gatherings and civic events spanning 9 months in Natchez, Mississippi, during the 1930s (Davis, Gardner, & Gardner, 1941). Anthropologists collected data through interviews, participant observations, guest lists, and newspaper reports. Information about the direct ties among the women, such as their kinships and friendships, was not reported in a published binary matrix showing which women were present or absent at which events. Social network analysts subsequently conducted at least 21 analyses to identify the structural positions occupied by the Southern Women according to their presence and absence at events (Freeman, 2003). Other examples of 2-mode networks include protestors attending social movement protests, readers commenting on web blogs, legislators cosponsoring bills, political action committees (PACs) contributing funds to candidates' election campaigns, firms joining strategic alliances, and nations signing military treaties.

Two-Mode Matrix and Bipartite Graph. A binary 2-mode network can be formally represented by a *2-mode matrix* that records the presence and absence of g actors at h events. Thus, its dimensions are g rows and h

columns, respectively. If actor i attends event j, the entry in the i, jth cell in the matrix equals 1; otherwise, the entry is 0. Denoting a binary bipartite matrix as **A**, its x_{ij} values meet these conditions:

$$x_{i,j} \begin{cases} 1 & \text{if actor } i \text{ is affiliated with event } j \\ 0 & \text{otherwise} \end{cases}$$

The row totals, also called row marginals, of matrix **A** sum to the number of events that each actor attended $\left(\sum_{j=1}^{h} x_{i,j}\right)$. The column marginals $\left(\sum_{i=1}^{g} x_{i,j}\right)$ indicate the number of actors who attended each event.

A 2-mode network may also be displayed as a *bipartite graph,* in which undirected lines connect actors aligned on one side of the diagram to the events aligned on the other side. Importantly, a bipartite graph does not permit lines among the actors nor among the events. A *bipartite matrix* contains both sets of actors and events in the rows and columns. Assuming that a 2-mode network has g actors and h events, the bipartite matrix has dimensions $(g + h) \times (g + h)$.

Figure 5.8 displays a 2-mode network consisting of five PACs that made donations to the campaigns of four U.S. senators during the 2008–2012 election cycle. (UPS is United Parcel Service, MS is Microsoft, HD is Home Depot, SEU is Service Employees Union, and ANA is American Nurses Association.) Republican and Democratic party membership of each senator is indicated, respectively, by -R or -D appended to their names. In this example, the PACs are the actors and the senatorial election campaigns are the events. Although the bipartite graph contains no lines directly connecting PACs to one another, they are indirectly linked through donations made to the same senators. UPS and MS gave only to Republicans, the SEU and ANA donated only to Democrats, whereas HD funded members of both parties. Table 5.2 is the bipartite matrix corresponding to the graph in Figure 5.8, where a cell value of 1 indicates a PAC donation to a senatorial campaign.

The 5×4 submatrix in the upper right quadrant, consisting of the five PACs in the rows and the four senators in the columns, is the 2-mode affiliation matrix, denoted as **A**. The lower left quadrant is the transpose of **A**, denoted as **A′** (with dimensions 4×5 and $x_{ij} = x_{ji}$). The transposed matrix shows which senators in the rows received funding from the PACs in the columns. The other two quadrants of a bipartite matrix always have only 0s in their cells because the actors are not directly linked to one another and neither are the events. A bipartite matrix can be schematically represented as the following:

$$\mathbf{X}^{\mathbf{A,E}} = \begin{bmatrix} 0 & \mathbf{A} \\ \mathbf{A}' & 0 \end{bmatrix} \tag{5.4}$$

At the margins of the table, but not properly part of the matrix, the row totals equal the column totals. The total for a PAC shows the number of senators to whom it donated, whereas the total for a senator indicates the number of PACs from which he received funding.

Multiplying the two submatrices (**A** and **A′**) in two different orders yields additional information about relations among the actors and among the events that is not available in a bipartite matrix. This process, called *projection*, reduces a 2-mode matrix to a pair of 1-mode matrices. The first result, $\mathbf{X}^{\mathbf{A}}$, is a symmetric, valued matrix of coattendances for pairs of actors, obtained as the product of matrix multiplication:

$$\boldsymbol{X}^{\boldsymbol{A}} = \boldsymbol{AA}' \tag{5.5}$$

In general, a 2-mode network **A** is a $g \times h$ matrix, and its transpose **A′** is an $h \times g$ matrix. Thus, $\boldsymbol{X}^{\mathbf{A}}$ is always a $g \times g$ matrix, whose nondiagonal cell

Figure 5.8 Bipartite Graph of PAC Donations to Senatorial Election Campaigns

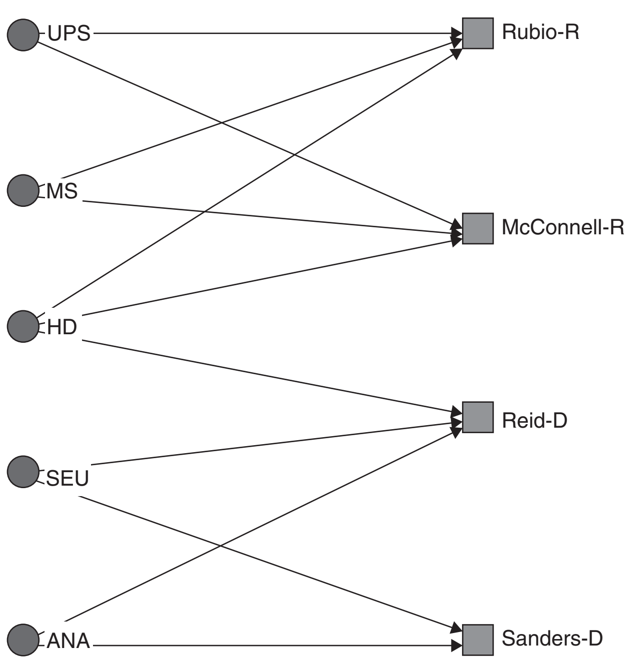


Table 5.2 Bipartite Matrix of the Graph in Figure 5.8

	UPS	*MS*	*HD*	*SEU*	*ANA*	*Rubio-R*	*McConnell-R*	*Reid-D*	*Sanders-D*	*Row Total*
UPS	—	0	0	0	0	1	1	0	0	2
MS	0	—	0	0	0	1	1	0	0	2
HD	0	0	—	0	0	1	1	1	0	3
SEU	0	0	0	—	0	0	0	1	1	2
ANA	0	0	0	0	—	0	0	1	1	2
Rubio-R	1	1	1	0	0	—	0	0	0	3
McConnell-R	1	1	1	0	0	0	—	0	0	3
Reid-D	0	0	1	1	1	0	0	—	0	3
Sanders-D	0	0	0	1	1	0	0	0	—	2
Column Total	2	2	3	2	2	3	3	3	2	—

values are the numbers of events attended by both actor i and actor j. The diagonal entries of $\mathbf{X^A}$ show the number of events each actor attended.

The second result, $\mathbf{X^E}$, is a symmetric, valued matrix of coparticipants at pairs of events, obtained by multiplying the matrices in reverse order:

$$\boldsymbol{X^E} = \boldsymbol{A'A} \tag{5.6}$$

The values in the nondiagonal cells of this $h \times h$ matrix are the number of actors participating in both event i and event j, whereas the diagonal entries of $\mathbf{X^E}$ are the number of actors attending each event.

The 2-mode PAC-senator matrix and its transpose are the following:

$$\mathbf{A} = \begin{pmatrix} 1100 \\ 1100 \\ 1110 \\ 0011 \\ 0011 \end{pmatrix} \quad \mathbf{A'} = \begin{pmatrix} 11100 \\ 11100 \\ 00111 \\ 00011 \end{pmatrix}$$

The first projection, from the 2-mode matrix to a 1-mode PAC matrix $\mathbf{X^A}$, multiplies $\mathbf{A}$ by $\mathbf{A'}$:

$$\mathbf{X^A} = \mathbf{A}A' = \begin{pmatrix} 22200 \\ 22200 \\ 22311 \\ 00122 \\ 00122 \end{pmatrix}$$

$\mathbf{X^A}$ is a 5 × 5 symmetric matrix whose diagonal cell entries show the number of senators to whom each PAC donated. HD gave to three senators; all others donated to two. The off-diagonal values show how many senators received donations from each pair of PACs. For example, the third row shows that HD and UPS donated to two senators, as did HD and MS. But, two dyads (HD and SEU, HD and ANS) had only one senatorial recipient in common. Although the projection matrix shows the total numbers of senators receiving donations from dyads, information about which specific senators received those funds is no longer available.

The second projection, from the 2-mode matrix to a 1-mode senator matrix $\mathbf{X^E}$, multiplies $\mathbf{A'}$ by $\mathbf{A'}$:

$$\mathbf{X^E} = \mathbf{A'A} = \begin{pmatrix} 3310 \\ 3310 \\ 1132 \\ 0022 \end{pmatrix}$$

The diagonal cells of this 4 × 4 symmetric matrix show the number of PACs jointly donating funds to each senatorial campaign. Sanders was funded by two PACs; all other senators received donations from three PACs. The off-diagonal entries are the number of PACs jointly funding senatorial dyads. The first row and column show that Rubio and McConnell received money from the same three PACs, Rubio and Reid were jointly funded by one PAC, but Rubio and Sanders had no common sources of campaign donations. Any projection involves a loss of information, in this case, the identities of the PACs funding the senators.

At the network level of analysis, mean rates of activity are readily computed from values in the two projection matrices. Because the diagonal entries of $\mathbf{X^A}$ are the numbers of actors participating in each event, summing them and dividing by the number of events yields the mean, $\bar{\mathbf{X}}^A = \frac{\sum_{i=1}^{g} x_{i,i}}{g}$. In the example, PACs donated to a mean of (11/5) = 2.20

senatorial election campaigns. Likewise, summing the diagonal values of $\mathbf{X^E}$ and dividing by the total number of events results in the mean number of actors at an event, $\bar{\mathbf{X}}^{\mathbf{E}} = \frac{\sum_{i=1}^{h} x_{i,i}}{h}$. In the example, mean number of PAC donations per senator was (11/4) = 2.75.

Density and Centrality in 2-Mode Networks. Density and centrality are important basic network properties that also apply to 2-mode networks. As discussed in Chapter 4, density measures reveal either the proportion of ties present in a binary graph or the mean value of the observed lines in a valued graph. Similarly, the interpretation of density for a 2-mode network depends on whether it is a binary or valued graph (Wasserman & Faust, 1994, p. 316).

For a symmetric $g \times g$ coattendance matrix, $\mathbf{X^A}$, whose nondiagonal values are the number of events attended by each pair of actors, the density measure is the following:

$$D^A = \frac{\sum_{i=1}^{g} \sum_{j=1}^{g} X_{ij}^{A}}{\frac{(g-1)}{2}} (i < j) \tag{5.7}$$

The numerator sums all the values in the upper triangle of the coattendance matrix (i.e., above the diagonal because the lower triangle values are identical). The diagonal values are excluded because we cannot consider an actor as attending an event with itself. The denominator is the total number of nonordered dyads, again excluding the diagonal. For a symmetric $h \times h$ coparticipation matrix, $\mathbf{X^E}$, whose nondiagonal values are the number of actors participating in each event, the density measure is the following:

$$D^E = \frac{\sum_{i=1}^{h} \sum_{j}^{h} X_{i,j}^{E}}{\frac{h(h-1)}{2}} (i < j) \tag{5.8}$$

In the example, $D^A = 11/10 = 1.10$, meaning that pairs of PACs jointly donated to a mean of 1.10 senatorial election campaigns. From senatorial perspective, $D^E = 7/6 = 1.17$, indicating that pairs of senators attracted contributions from mean of 1.17 PACs. Because projection matrices may have cell values greater than 1, densities could be larger than 1.0. Consequently, each density value should be interpreted not as a proportion but as either the average number of events two actors attended (D^A) or the average number of actors present at two events (D^E), respectively.

Social network analysts have studied centrality at the actor level and centralization at the graph level of analysis for decades (Freeman, 1979; Wasserman & Faust, 1994, chap. 3). We commented in Chapter 4 that actor centrality measures the importance or visibility of actors within a network. Analysts describe four major types of centrality: degree, closeness, betweenness, and eigenvector centrality. Degree centrality reflects the extent to which an actor is active in a network, closeness centrality measures the extent to which an actor is connected to other actors in a network via shortest paths, betweenness centrality captures the extent to which an actor mediates flows of information or resources between other actors in a network, and eigenvector centrality reflects the extent to which an actor is connected to other central actors in a network. Faust (1997) discussed application of these four centrality measures to affiliation networks, but space constraints allow us to cover only degree centrality in 2-mode networks.

Drawing on the general idea that degree centrality involves the total number of direct ties, actor degree centrality in an affiliation network is the total number of actor contacts that the *i*th actor has through its attendance at all events, obtained by summing the *i*th row of the coattendance matrix, $\mathbf{X^A}$:

$$C_D^A\left(a_i\right)=\sum_{j=1}^{g} x_{ij}^A\left(i \neq j\right) \tag{5.9}$$

In the example, HD donated to the same senators as UPS, MS, SEU, and ANA, so its degree centrality is 4. The degree centralities of the other four PACs are 2 because they each funded the same campaigns as two other PACs. Verify these dyadic commonalities by visually inspecting Figure 5.8.

Likewise, an event's degree centrality in an affiliation network is the total number of event contacts that the *j*th event has through the participation of all actors, obtained by summing the *j*th column of the coparticipation matrix, $\mathbf{X^E}$:

$$C_D^E\left(e_i\right)=\sum_{j=1}^{h} x_{ij}^E\left(i \neq j\right) \tag{5.10}$$

The degree centralities for the four senatorial campaigns are 2, 2, 3, and 1, respectively. Reid had the highest degree centrality because he received contributions from HD, which also funded Rubio and McConnell, and from two PACs (SEU and ANA) that also donated to Sanders. In contrast, the Sanders campaign shared funding sources only with Reid.

Three-Mode Networks

The principles and procedures for analyzing 2-mode networks can be readily applied to 3-mode networks. A 3-mode network consists of relations among three distinct sets of entities. In a restricted 3-mode network, the nodes in two sets of entities may be connected only to one or more nodes in the third set, and none of the nodes within a set have links to one another. Entities may all be the same type (e.g., three types of organizations, such as political parties, electoral campaigns, and PACs) or each type of entity may differ (e.g., persons, groups, and events). Similarly, the relational contents connecting the entities may be the same or different. Ties could be directed from one entity to another; for example, donors give money to candidates. Or ties may be undirected; for instance, candidates discuss election strategies with their advisors. The ties connecting different types of entities may constitute a hierarchy of authority (such as a military or corporate command structure) or the relations between entities may be nonhierarchical.

To describe a restricted 3-mode network, we'll assume that actors participate in events and events result in products but no direct relations exist between actors and products. A restricted 3-mode network can be conceptualized as a pair of 2-mode matrices: the first matrix **A** has actors connected to events (with order $g \times h$) and the second matrix **P** has the same events linked to a set of products (with order $h \times i$), where g, h, and i are integer numbers of actors, events, and products, respectively. Both 2-mode matrices can be transposed: the order of $\mathbf{A'}$ is $h \times g$ and the order of $\mathbf{P'}$ is i x h.

A 3-mode matrix can be constructed by combining the pair of 2-mode matrices and their transposes so that all entities appear in the rows and columns. As shown schematically in Table 5.3, the 2-mode matrices and their transposes create a square matrix **T** whose order is $(g + h + i) \times (g + h + i)$. Matrix **T** has five submatrices of structural zeros, reflecting restrictions on the types of relations that cannot be ascertained (no intraentity relations and no direct relations between actors and products).

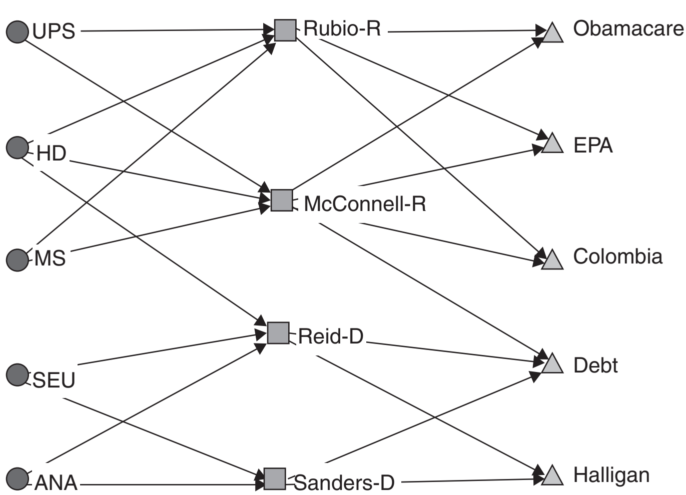
We expand the example by adding five products, congressional bills for which the senators cast votes in favor of or against becoming a law. The bills are repeal of Obamacare, Environmental Protection Agency regulation prohibition, Colombia trade agreement, debt limit increase, and nomination of Caitlin Halligan to the DC Court of Appeals. In the tripartite graph in Figure 5.9, a directed line from a senator to bill indicates a vote in favor of passage, whereas the absence of a line means a vote against the bill. The first three bills were backed by both Republican senators, the last bill was supported by both Democrats, and only the bill to raise the national debt ceiling received votes from senators of both parties.

As discussed in Chapter 4, an important objective of many empirical network analyses is to identify subgroups of entities that jointly occupy structural positions in a network. Some of the methods for identifying subgroups in 1-mode networks can also be applied to bipartite and tripartite matrices, usually resulting in subgroups occupied by all types of entities. To illustrate, we conducted a faction analysis of the 3-mode example. Given the small number of entities, we decided to partition the data into two factions (Borgatti et al., 2013, pp. 191–195). A combinatorial optimization algorithm called Tabu Search begins by arbitrarily assigning network

Table 5.3 Schematic of Tripartite Matrix **T**

	Actors	*Events*	*Products*
Actors	0	A	0
Events	A′	0	E
Products	0	E′	0

Figure 5.9 Tripartite Graph of PAC Donations to Senatorial Election Campaigns and Senators' Votes for Bills



entities to one of the hypothesized factions to maximize a fit criterion (Glover, 1989, 1990). That criterion is an empirical solution with an ideal clique structure where all within-group tie densities = 1.00 and all between-group densities = 0.00. The algorithm then moves some entities to other factions, recalculates the fit, and continues until no further improvement in fit is possible. Unfortunately, combinatorial optimization procedures always generate a solution, even if the purported factions aren't really cohesive subgroups as indicated by poor fit values. Borgatti et al. (2013, p. 192) advocated repeating the analysis "a number of times to see whether the final factions are the same or similar."

In our example, the same factions emerged time after time because the cleavage is so clearly polarized. In Table 5.4, the permuted matrix shows that one faction has the two Democratic senators, the two PACs donating only to them, and both bills they voted for. The other faction consists of the two Republican senators, the three PACs funding them, and the three bills they supported. The two within-faction densities are both 1.00 (connections that are not permissible, indicated by dashes, are ignored in calculating the densities). The between-faction density is not zero but is very low (0.10) because HD gave funds to members of both factions and

Table 5.4 Factions in Tripartite Network of PAC Donations to Senatorial Election Campaigns and Senators' Votes for Bills.

1 1 1 1 1

		8 R	9 S	3 D	4 S	5 A	4 H	3 H	1 U	2 M	0 O	1 C	2 E	6 R	7 M
8	Reid-D	–	–	1	1	1	1	1	0	0	0	0	0	–	–
9	Sanders-D	–	–	1	1	1	1	0	0	0	0	0	0	–	–
13	Debt	1	1	–	–	–	–	–	–	–	–	–	–	0	1
4	SEU	1	1	–	–	–	–	–	–	–	–	–	–	0	0
5	ANA	1	1	–	–	–	–	–	–	–	–	–	–	0	0
14	Halligan	1	1	–	–	–	–	–	–	–	–	–	–	0	0
3	HD	1	0	–	–	–	–	–	–	–	–	–	–	1	1
1	UPS	0	0	–	–	–	–	–	–	–	–	–	–	1	1
2	MS	0	0	–	–	–	–	–	–	–	–	–	–	1	1
10	Obamacare	0	0	–	–	–	–	–	–	–	–	–	–	1	1
11	Colombia	0	0	–	–	–	–	–	–	–	–	–	–	1	1
12	EPA	0	0	–	–	–	–	–	–	–	–	–	–	1	1
6	Rubio-R	–	–	0	0	0	0	1	1	1	1	1	1	–	–
7	McConnell-R	–	–	1	0	0	0	1	1	1	1	1	1	–	–

Republican senator McConnell voted for the debt ceiling bill supported by the Democratic faction.

5.4. Community Detection

In the early 21st century, physicists, biologists, computer scientists, and mathematicians grew increasingly involved in network analysis and developed many techniques for finding positions under the general rubric of community detection (Freeman, 2011). Less "well-posed" than graph partitioning methods, community detection seeks to find the "natural division" of a network into subsets of nodes, regardless of the number or size of groups, having many lines within groups and few lines between groups (Newman, 2010). Mark Newman (2006a) argued that community structures correspond to a statistical arrangement of edges (lines), as measured by the modularity of a network partition (see also Girvan & Newman, 2002; Newman, 2003; Newman, 2006b). "The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random" (Newman, 2006a, p. 2). If the observed number of lines is no greater than random, modularity is zero, and thus network partitioning into meaningful subgraphs is not possible. As modularity approaches a maximum of one, a network is characterized by a strong community structure with higher-than-random intragroup ties and sparse intergroup connections. Newman reformulated the optimal modularity method in terms of the "eigenvectors of a new characteristic matrix for the network, which we call the modularity matrix, and . . . this reformulation leads to a spectral algorithm for community detection" (2006a, p.1).

During the past decade, numerous community detection algorithms based on diverse assumptions proliferated, a plethora that threatened to overwhelm the capacity of empirical researchers to select practical tools appropriate to their tasks (Fortunato, 2010; Lancichinetti & Fortunato, 2009; Leskovec, Lang, & Mahoney, 2010). Orman, Labatut, and Cherifi (2011) compared 11 community detection algorithms applied to artificial network datasets. They used a normalized information measure to assess the extent of similarity between observed and estimated community structures. They concluded that network size and average proportion of intracommunity to intercommunity ties had the greatest impacts on algorithm performances. The most consistent method was by Infomap (a compression-based algorithm), followed by Walktrap (node-similarity based on random walks), MarkovCluster (diffusion), SpinGlass (node-similarity), and Louvain (modularity). In a subsequent article, Orman, Labatut, and Cherifi (2012) evaluated a representative set of eight community detection algorithms by applying both traditional measures of community structure as a partition (sets of nodes) and measures of community topological properties (e.g., density, distance, transitivity) to artificially generated realistic networks. Finding no equivalence between the two approaches, they concluded that "high performance does not necessarily correspond to correct topological properties, and vice-versa" (p. 1). The analysts recommended applying both complementary approaches to perform a thorough assessment.

The Louvain method rapidly became "one of the most popular algorithms for maximizing modularity" (Bhowmick & Srinivasan, 2013, p. 111) due to its ability to detect community partitions in networks with millions of nodes and billions of links in a fast and efficient manner. Computer scientists at the Université Catholique de Louvain in Belgium developed an algorithm to handle the resolution limit problem: the inability to detect smaller communities in large networks (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008, 2011). Modularity optimization in the Louvain method proceeds in two iterative steps. First, the algorithm finds small communities by optimizing local modularity. Second, it aggregates nodes belonging to the same community and constructs a new network with communities as the nodes. These steps repeat until maximum modularity is achieved and a hierarchy of communities emerges. Empirical applications of the Louvain method included mobile phone networks (Blondel, Krings, & Thomas, 2010; Walsh & Pozdnoukhov, 2011), airline transportation networks (Chopade & Bikdah, 2015; Gegov, Postorino, Atherton, & Gobet, 2013), political networks (Porter et al., 2005), and innumerable analyses of Twitter networks (Grabowicz, Asco, Moro, Pujol, & Eguiluz, 2012; Labatut, Dugué, & Perez, 2014).

For 2-mode networks, Larremore et al. (2014) formulated an approach to community detection without resorting to collapsing the data into 1-mode projections. Projections discard information and create networks composed of overlapping cliques that violate assumptions underlying community detection methods, which could result in finding strong community structures where none exist. They proposed a bipartite stochastic blockmodel method that can "efficiently and accurately find community structure in synthetic bipartite networks with known structure and in real-world bipartite networks with unknown structure."

Core/Periphery Models. A core/periphery structure consists of "two classes of nodes, namely a cohesive subgraph (the core) in which actors are connected to each other in some maximal sense and a class of actors that are more loosely connected to the cohesive subgraph but lack maximal cohesion with the core" (Borgatti & Everett, 1999). This idealized pattern generalizes Freeman's (1979) maximally centralized graph, a star-shaped diagram in which one node (the center) has direct ties to all $g - 1$ other nodes, which are unconnected to one another. A core/periphery model simply adds other entities to the graph's center and connects them to one another and to the periphery entities. However, such a pure core/periphery structure is very unlikely in an empirical network, so Borgatti and Everett proposed approximations to the ideal pattern, where the members of the core 1-block may have less than complete connections among themselves and the 0-blocks may contain a few links. Borgatti, Everett, and Johnson (2013, pp. 223–229) described an algorithm for detecting core/periphery structures in empirical networks by partitioning a binary matrix into two positions that maximize a fit statistic.

Two core/periphery options are categorical (discrete) models and continuous models, with measures of model fit based on the correlation between the data matrix and an ideal block model. The continuous model conceptualizes the probability of a tie between two entities as a function of product of each node's "coreness," that is, the closeness of the core to each entity. Its algorithm calculates the extent to which a network has a core/periphery structure for different sizes of the core. The results of an empirical core/periphery analysis may be unstable. Analysts should take care when using the routines because alternative partitions may produce equally good fits. To test for solution robustness, Borgatti and Everett advocated rerunning the analyses several times from different starting configurations. A good agreement among those alternative outcomes indicates a clear split of the network into core and peripheral positions.

To fit categorical core/periphery models to 2-mode networks, Everett and Borgatti (2013) advocated a dual-projection method. First, each of the two 1-mode projections—created by multiplying a 2-mode matrix and its transpose—is separately partitioned into core and periphery subsets. Then, the dual partition assignments are applied simultaneously to the corresponding rows and columns of the original 2-mode matrix. The result is a 2-mode core/periphery model and its associated 2 × 2 density and image matrices. Borgatti et al. (2013, pp. 243–244) discussed how to analyze a 2-mode dataset by constructing two 1-mode projections for each type of entity, finding the continuous core/periphery model separately for each projection, and then applying the combined results to the original affiliation matrix.

Affiliated Graph Model. The AGM approach to community detection conceptualizes communities as overlapping tiles, analogous to shingles on the roof of a house.

> Thus, just as the overlap of two tiles leads to a higher tile height in the overlapping area, the overlap of two communities leads to higher density of edges in the overlap. . . . The composition of many overlapping communities then gives rise to the global structure of the network. (Yang & Leskovec, 2014, p. 1894)

AGM measures the likelihood of a community affiliation graph and identifies the most likely community memberships of every entity by fitting the AGM to an observed network. It parameterizes each community A with a single probability (pA). Two nodes that belong to community A form a connection (edge) with probability A in the underlying network. Through an iterative search, each community generates links among its members independently, with the proviso that duplicate connections are not included. Pairs of nodes belonging to multiple communities "become connected in the underlying network with a higher probability, since for each shared community the nodes are given an independent chance of forming an edge" (p. 1895). The algorithm combines "a maximum-likelihood approach with convex optimization and a Monte Carlo sampling algorithm on the space of community affiliation graphs." The process is an efficient search that identifies the community affiliations of nodes, giving the observed community affiliation graph the highest likelihood. The program allows analysts to specify the number of communities to identify and to control the probability of links between nodes not sharing any communities. Alternatively, the program can derive those parameters empirically.

AGM was developed by Jure Leskovec and colleagues and is available on the Stanford Network Analysis Platform (Leskovec & Sosič, 2018). Applied to diverse datasets, the algorithm revealed how overlapping communities unify the modularity and core-periphery organizing principles in complex social and biological networks (Yang & Leskovec, 2014; Yang, McAuley, & Leskovec, 2014). The analysts demonstrated that dense network cores form as an intersection of multiple overlapping communities, and all nodes belonging to multiple communities also reside in the network's core community.

5.5. Exponential Random Graph Models (ERGMs)

Most social network analytic methods are descriptive, depicting structures and processes in relational data. Indeed, those descriptive measures, along with many powerful visualization tools, such as MDS and hierarchical clustering, give social network analysis its unique advantages for communicating complex social relations to general audiences through simple and vivid graphs. However, to social network researchers, mere description of social networks is only a first step. They are keenly interested in what lies beneath network patterns, what causes structures to emerge and to sustain or change. In parallel to descriptive and inferential statistics, social network analysis also encompasses descriptive and explanatory components. A major development in the explanatory branch is exponential random graph modeling (ERGM), a family of probability distributions on graphs.

We start with the very basic question: why can't we directly apply some basic statistical procedures to social network data? After all, before the advent of ERGM, many scholars estimated multivariate regression equations with network data. Especially at the dyadic level, logistic regression is used to explain the presence or absence of dichotomous network ties between pairs of actors, using such predictors as similarity/dissimilarity of a dyad's genders, races, nationalities, and other variables. Unfortunately, such applications confront two issues. (1) It violates the necessary assumption in multivariate regression that cases in the sample are selected independently of one another. In network data, cases and dyads are interdependent, not independent. For example, Ivan and Katya are more likely to be friends than any random dyad if they are both friends with Vlad. (2) Conventional inferential statistics assume that a sample is selected from the population with a known a priori probability. With appropriate weights, if needed, such a sample is representative of the population from which it was drawn. Further, inference about cause and effect is not disrupted by processes related to sample selection, as it was random. However, other than for ego-centered networks, network data are rarely random sampled; rather, network data comprise the entire population or the sampling distribution is unknown (Yang, Keller, & Zheng, 2016). Although one can still conduct conventional inferential statistical tests with network data, interpretations of the results is not equivalent to analyses of randomly sampled data. Simply put, statistically, inferences from network data do not give us much assurance due to their nonrandom sampling procedures. Alternative innovative methods of inferential statistical methods were developed to deal with the situation.

One way to circumvent the issues of interdependency and nonrandom sampling for inferential statistics is through computer simulation. David Krackhardt (1987) used simulation to imitate the network sampling distributions. The quadratic assignment procedure (QAP) calculates correlations or regressions using two or more matrices of network data. Let's assume that two observed matrices—for example, friendship and advising—have an observed correlation coefficient of $r = 0.35$. QAP randomly permutes the rows and columns of the two matrices, using the same permutation for both, then recalculates the correlation. The process is repeated thousands of times (or more, depending on the analyst's demand), generating an empirical sampling distribution for the correlation coefficient. A probability (p) is estimated based on the proportion of coefficients in the sampling distribution that is equal to or larger than 0.35, the observed correlation. If $p < .05$, then the correlation is unlikely to have occurred by chance. Conversely, if $p > .05$, then the observed coefficient is likely to occur by chance. The computed *p* value differs from the *p* value in parametric statistics. It doesn't correct or change the fact that the sample is not randomly selected. It just implies that the likelihood that the two matrices are independent of each other is fairly low or that the two matrices are likely correlated.

The ERGM follows the similar logic of computer simulation and inferential statistics as QAP. But, ERGM simulation is much more complex and more computationally intensive and it had a developmental path that began in the 1950s. We describe that path in the following subsections about the origins of the model, the model itself, and the explanatory framework of the model. We skip such topics as algorithms, simulations, and the estimation process, referring interested readers to more focused books on ERGM (Harris, 2014; Lusher, Koskinen, & Robins, 2013).

Origins of ERGM. Erdős and Rényi (1959) developed a simple random graph model to describe observations of certain network graphs with probability terms. The simple random graph model did a poor job of capturing the observed network structure because it assumes that network ties are formed randomly and independently of one another (Harris, 2014). However, the simple random graph model served as a foundation leading to more complex models.

Random graph models witnessed major developments in the 1980s. Holland and Leinhardt (1981) developed the P_1 model to account for two characteristics commonly observed in directed binary graphs: (1) the network has large variation in nodal indegrees and (2) reciprocity occurs more commonly than expected. Their model explicitly accounts for reciprocation and differential attractiveness, using those features to explain the probability of an observed network. Frank and Strauss (1986) introduced the dyadic dependence model (dubbed the P_2 model), which applies Markov dependence to examine dyadic dependence. For example, the model accounts for the tendency for A to know C if both are tied to B. However, although the P_2 model used structural characteristics, it did not include the individual network actor's characteristics. Such restriction was later relaxed with Wasserman and Pattison's (1996) work on a model called p*. Compared with previous models, p* accounts for a variety of structural features such as reciprocity, homophily, transitivity, and nonuniform degree distributions. In addition, it also allows researchers to examine the effects of network actor attributes, such as gender, race, education, and age.

The ERGM arose from developments on three fronts: (1) social scientists struggling with ways to account for observed network data (they thus identified problems to be solved), (2) research advances in graph theory, which produced mathematical solutions to explanatory models of social network data, and (3) advances in computer algorithms that allowed researchers to implement mathematical solutions with suitable algorithms (in particular, the Markov chain Monte Carlo algorithm). Three well-known softwares emerged to help researchers estimate ERGM models: R, Siena, and PNet.

The ERGM. Several distinguishable features mark ERGM as superior to its predecessors. First, the model provides definitive answers to the importance of each explanatory variable with a parameterized model. Second, the model simultaneously accounts for three types of explanatory factors affecting the creation or change of network ties: the endogenous network configurations (density, dependency, reciprocity, and transitivity, etc.), the actor attributes (age, gender, race, etc.), and other environmental covariates that are exogenous. Third, the model overcomes the problem of degeneracy by achieving convergence on a solution. Simply put, the model ensures that parameters reach their final values in a convergent model. The following formula shows the ERGM in a generalized mathematical formulation (Lusher et al., 2013, p. 9):

$$P_\theta\left(G\right) = ce^{\theta_1 Z_1(G) + \theta_2 Z_2(G) + \ldots + \theta_p Z_p(G)} \tag{5.11}$$

ERGM tries to explain the emergence of network ties with a sum of network statistics weighted by parameters inside an exponentiated term. Among those network statistics, the local endogenous network configurations are important factors to include in accounting for the presence/absence of binary ties in an observed network. The dependent matrix in many ERGMs is binary, directed, or undirected, with valued graphs only entering exogenous explanatory variables. But, recently developed generalized exponential random graph models (GERGMs), which can accommodate continuous-valued relations, greatly expand the types of social network data subjected to statistical analysis (e.g., Desmarais & Cranmer, 2012).

ERGM usually applies to a whole network with a finite number of nodes. Looking forward, several research programs aim to extend ERGM, such as longitudinal ERGM, bipartite ERGM, and multilevel ERGM. Regarding Big Data, although ERGM can handle networks with 1000 to 2000 nodes, going beyond a couple of thousands can be challenging. This constraint occurs because an ERGM sampling distribution is very computationally expensive. For example, the number of graphs that ERGM needs to create for simulating the sampling distribution of an observed graph with *n* nodes is $2^{n(n-1)/2}$ (Newman, 2010, p. 567). With a mere 6 nodes, the simulations would require

32,768 ($2^{15} = 32,768$) simple graphs to be generated. Perhaps with advances in fast computer processors, fast algorithms, and parallel computing, future ERGM can handle Big Data graphs with millions of nodes.

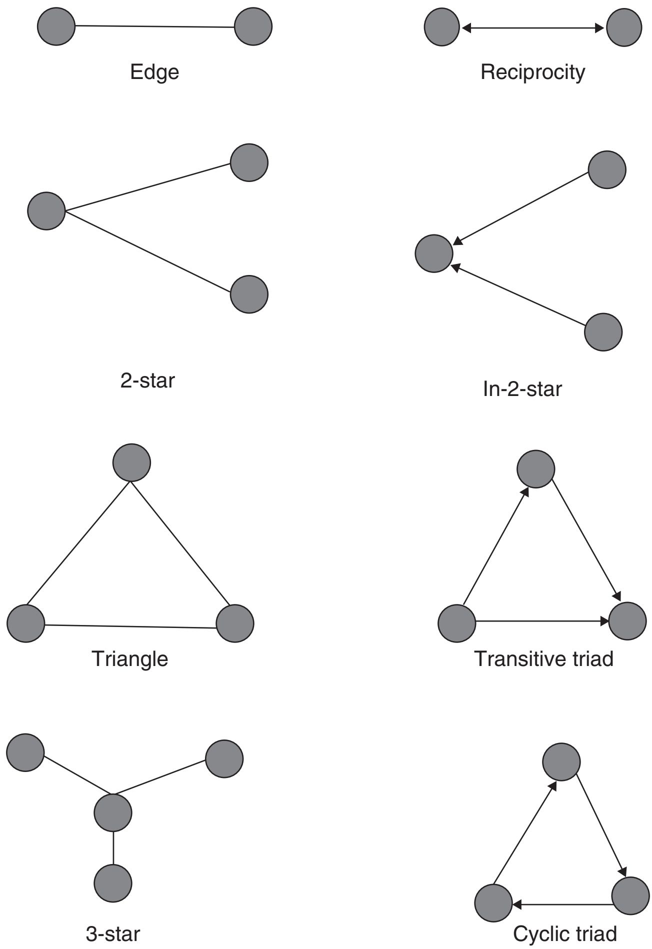
The Exploratory Framework of ERGM. One question researchers often ask is what network configurations should be included in a model? The answer lies in the researchers' objective/research topic, which means researchers must decide for themselves about what to include and what to exclude. However, Figure 5.10 shows the basic network configurations (undirected graphs on the left; directed networks on the right) that many researchers commonly consider for inclusion in their ERGMs. Including those endogenous network configurations allows researchers to assess the effects of those network characteristics on the emergence of network ties in the observed network.

A second dimension of explanatory factors in an ERGM is the individual attributes, which are familiar variables, such as a person's race, gender, age, occupation, attitudes, and beliefs. Those traditional independent variables in conventional social statistics are also present in ERGMs and permit many important social relations to be examined. For example, the homophily principle stipulates that people are attracted to similar others. Gender, race, age, education, religion, or sexual orientation can be used to measure various homophilies that may influence network tie formation. Analysts should distinguish between individual attributes that are exogenous and local network configurations that are endogenous. Individual attributes may affect tie formation, along with some purely structural effects. To look at only the individual effects without examining the structural influences will likely cause overestimation of individual effects. For example, a friendship between two female colleagues may be affected by a structural tendency toward reciprocity. Without accounting for such structural proclivities, the effects of their gender similarity are likely to be overestimated, attributing their connection to having the same gender.

A third dimension of ERGM explanatory variables is environmental. Environmental covariates capture contextual characteristics, operating much like independent control variables in conventional multiple regression models. For example, when examining tie formation in a formal organization, the formal job authority relations among employees in the organizational hierarchy must be considered when a researcher investigates the formation of informal ties. That way, the researcher can distinguish between ties that are mandated by job descriptions (subordinates must report to supervisors) and those which form spontaneously (through mutual attraction or shared interests).

A final dimension of ERGM variables is temporal. As we remarked in Chapter 2.1, the proliferation of longitudinal network datasets fosters the

Figure 5.10 Endogenous Network Configurations



invention of new statistical packages for analyzing them. Important advances include stochastic actor-oriented models, such as SIENA and RSiena (e.g., Niezink, Snijders, & van Duijn, 2019; Stadtfeld, Snijders, Steglich, & van Duijn, 2018), and temporal exponential random graph models (TERGMs) (e.g., Leifeld & Cranmer, 2019; Leifeld, Cranmer, & Desmarais, 2019). These models are sufficiently advanced that we do not attempt to describe them here.

5.6. Future Directions in Network Analysis

Social network analysis is an expanding multidisciplinary field that continually inspires new projects, generates methodological innovations, and poses challenges for collaborative efforts. In this concluding section, we speculate about some future directions for social network analysis.

Social network analysts model networking activities around recurring social events, many of which take place within certain geospatial confines. Routine social interactions are extensively location-based. For example, neighbors living close by often pay social visits to one another, discuss neighborhood issues, or exchange information and gifts. But neighbors residing on more distant blocks rarely interact. Colleagues whose offices are right next door commonly engage in daily greetings and casual conversations, which sometimes evolve into more serious topics such as office politics. But coworkers located on different floors seldom visit. More importantly, the geospatial sphere can be confounded with social causation. For example, the chances of becoming obese are probabilistically linked to the weights of one's social contacts (Christakis & Fowler, 2007), whereas the odds of obesity are also linked to geospatial access to pedestrian infrastructure or fitness facilities (Andris, 2016). In the past, social network analysts rarely paid much attention to geospatial characteristics, which they treated as an implicit factor under overarching social forces. Now, with both the social network analysis and Geographic Information System (GIS) in their respective mature stages, a communication or collaboration between the two sides could be groundbreaking—blazing new trails, enriching and extending both domains of existing knowledge. For example, sociologists and geographers could jointly leverage the constraints of social distance and geospatial distance to explain the formation, evolution, decay, and dissolution of dyadic ties.

A second area where social network researchers still see great growth potential is social media Big Data. Computer engineers have already made much progress with scraping and mining online social media network data, such as Facebook, Twitter, and LinkedIn. Currently, the task of data acquisition of massive online social media network data is less daunting: researchers with necessary skillsets can develop their own crawlers, but, for those who are not technologically savvy, commercial packages (often for a fee) are readily available to help with data scraping from popular social networking sites. Between building one's own crawlers and paying to use commercial data mining tools, data scientists developed a third option: application programming interface (API) for data scraping. With a few lines of code, API can scrape large amount of online social media data within a short time (adams, 2020). One crucial question is what is the quality of data scraped online? For example, one study reported that data available via API scraping may be biased in systematic and unknown ways from the potential population of users (Hargittai, 2018). Another showed some ways that the API process itself makes data available to researchers to produce additional unknown biases (González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014). Traditional inferential statistical methods were not developed to deal with assessing Big Data collected from online social media. We should anticipate future collaborations among social scientists, computer and information scientists, and mathematicians and statisticians to develop innovative methods to assess and improve data quality.

Another fruitful field for future development is the application of inferential statistics to social network analysis. We refer to statistical analyses of whole network data because conventional *t*-tests and *F*-tests apply to ego-net data (e.g., Marsden, 2011). The past decade saw substantial progress in explanatory models of ERGM. But, for many network methods that are descriptive, few inferential statistics are available to understand beyond simply describing a specific network. In other words, we have a poor understanding of how well a particular whole network dataset is related to other similar networks or the general population. Although some network analysts investigated sampling of nodes and ties (Frank, 2011), much more research is needed to improve knowledge of how whole network data relate to other network structures or the general population. Fortunately, some statisticians appear to be on the case (e.g., Crane & Dempsey, 2016; Dempsey, Oselio, & Hero, 2019).

We believe the network topics described previously have great growth potential or urgently need more research effort. They are certainly not an exhaustive list of future frontiers. For example, new methods of multiplex network analysis deserve greater attention. Teasing out the complex interconnections between multiple layers (node, dyads, triads, cliques, and whole network) and multiple relational contents (friendship, advice-seeking, information exchange, etc.) presents numerous challenges and opportunities, for example, how to visualize network structures in three dimensions. Longitudinal network analysis adds a fourth dimension: how to track changes in a 3-D network structure over time. With the proliferation of longitudinal network methods (Snijders, van de Bunt, & Steglich, 2010), many researchers will obtain fruitful results by applying these innovations to longitudinal social network datasets that are increasingly available. By combining longitudinal data and ERGM, social network scholars will be in a better position to address the issue of causality that has challenged network research for years. Indeed, although similarity breeds connections, connectivity also induces similarity. Cause-and-effect ambiguity is omnipresent in most social network studies. Thus, smokers hang together outdoors due to their shared vice—network as dependent variable. Conversely, teens start smoking when their friends do—network as independent variable. With longitudinal whole network data, we can establish the time order of changes among interdependent variables. ERGM can identify and control for confounding factors, covariance, and endogenous network processes. Working in tandem, causality in dynamic networks can be inferred from time order, probability values, and elimination of spuriousness.

We concur with adams (2020) that social network analysts should pay greater attention to the substances flowing through network ties and their impacts on social actors. We envision fruitful collaborations among social network, social capital, and social resource scholars. After all, social network analysis is not only about formal structures and methodologies but also about the effects and consequences of social networks for social actors. More importantly, between structure/methodology and substantive contents, the relationship is bilateral: we easily envision how the advancement of network methodologies improves knowledge of substantive issues. In turn, substantive investigations also present challenges, inspiring innovative ways to improve social network methodology. We look forward to social network analysis continuing to grow and mature in the years to come.

APPENDIX

Social Network Analysis Software Packages

In this book, we illustrate various social network analysis methods and visualizations with UCINET, developed by Borgatti, Everett, and Freeman (2002). The biggest advantage of UCINET is its user-friendly interface, which enables point-and-click menus to command a wide variety of basic and advanced network analytics. New users do not need a long time to learn how to command the software. However, UCINET is not created to deal with very large social network datasets, which is increasingly crucial in today's data science and data driven era. The majority of UCINET's analytical methods are descriptive, leaving out important statistical methods such as exponential random graph modeling (ERGM). Here, we summarize a half-dozen social network analysis softwares that users may explore as alternatives or supplements to UCINET.

1. **R** is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing (www.r-project.org/about.html). It is written primarily in programming languages such as C, C++, Fortran, Java, and R itself. R capabilities are extended through user-created packages, which allow specialized statistical techniques, graphical devices, import/export capabilities, reporting tools, etc. The biggest advantages of R for network analysts are (1) it is free of charge and (2) it is quite powerful, capable of handling very large datasets. However, R has a rather steep learning curve, particularly for new users who are not computer command savvy. Its open-source approach also allows developers to add or to delete/modify routines and methods, which may frustrate users trying to stay abreast of the software. Social network analysis functions and methods available in R, in particular, the "igraph" package (igraph.org/r), support fast handling of large network data with millions of nodes and fast implementation of network algorithms to analyze nodes, ties, and graphs. R is available for Windows PC or Apple's Mac operating environments. Social network users may ask: when should we forgo the ease of using UCINET in favor of computer command-based R? The answer is, when you have a large dataset, containing more than 10,000 nodes, for example, and/or you want to estimate an explanatory network model such as ERGM. R provides ready solutions to those challenges. However, if your network dataset has tens or even hundreds of nodes, and you need mostly descriptive analyses, then UCINET likely provides satisfactory solutions.

2. **SIENA** stands for Simulation Investigation for Empirical Network Analysis. This free program uses an actor-oriented model to analyze longitudinal data, that is, where a whole network has a repeated measure. SIENA was developed by Tom Snijders, Christian Steglich, and their associates (www.stats.ox.ac.uk/~snijders/siena). The early version, SIENA 3.0, also implemented ERGM and is still available but not maintained. The successor program, **RSiena**, is integrated in the package of the R statistical system, as is a more experimental version, **RSiena/Test**. Although both programs perform statistical analysis of changes in longitudinal networks using computer simulations, neither estimates ERGMs.

3. **PNet** is a specialty software that handles ERGM and a variety of methods related to ERGM. Developers behind PNet are Dean Lusher and associates at Swinburne University of Technology in Australia (www.melnet.org.au/ergm). An advantage of PNet is its solution to one-mode as well as two-mode and two-level ERGMs of network data. Although PNet was developed to deal with the one-mode network data, **MPNet** handles two-mode and two-level ERGMs. The team also developed **XPNet** for bivariate network analysis. The research team published a book about ERGM in general and PNet in particular (Lusher, Koskinen, & Robins, 2012).

4. **Gephi** is an open visualization program initially developed by students of the University of Technology of Compiègne in France (Grandjean, 2015) (gephi.org). It is widely used in academics of such disciplines as history, journalism, and political science as well as such social media practitioners as Twitter. Although capable of conducting basic social network analysis methods, Gephi is famous for its capability to map large data of online social networking activities.

5. **Pajek,** which means "spider" in Slovenian, was developed by Vladimir Batagelj and Andrej Mrvar at the University of Ljubljana in Slovenia. Like Gephi, Pajek is renowned for its ability to visualize very large social networks: it can process over one billion nodes. Pajek has programs to detect communities, analyze signed networks, and facilitate genealogy and citation research. And, like PNet, Pajek's developers published books to introduce basic social network analysis methods and their software (De Nooy, Mrvar, & Batagelj, 2018).

6. **NodeXL** stands for Network Overview Discovery and Exploration for Excel. The package was developed by an interdisciplinary team led by Ben Shneiderman and Marc Smith. NodeXL is designed specifically to mine data from popular social media network sites such as Twitter and Facebook. Researchers also use NodeXL to process and to analyze data from WWW Hyperlink, Flickr, Youtube, and Wiki networks (Hansen, Shneiderman, & Smith, 2011).

Table A Comparing Social Network Analysis Softwares

Names	*Platform*	*Pricing*	*Easiness of use*	*Strength*	*Weakness*
UCINET	PC (Windows); special handling for Mac users	First 90 days after download is free; after that varied prices apply	Easy to use, point and click to command functions	User-friendly	Unable to handle large data; no ERGM functionality
R	PC and Mac	Free	Computer command based, may have steep learning curve for beginners	Omnifunctional package for SNA	Steep learning curve
SIENA	PC and Mac for RSiena	Free	Siena 3.0 is relatively easy; RSiena is integrated with R	Easy to use, particularly Siena 3.0	Limited functions with ERGM
PNet	PC (Windows)	Free	Easy to use with PNet self-tutorial	Deals with various ERGM data structures	Standing-alone focusing on ERGM exclusively; only available in PC
Gephi	PC and Mac	Free	Self-tutorial can be challenging	Large data capability, graphing	Learning takes time
Pajek	PC (Windows)	Free	Self-tutorial can be challenging	Large data capability, graphing	Only available in PC
NodeXL	PC (Windows)	Basic version is free; the Pro version costs with different pricing	Easy to use with self-tutorial	Relatively large data, capabilities of mining social media data	Only available in PC

REFERENCES

adams, j. (2020). *Gathering social network data.* Thousand Oaks, CA: Sage.

Aggarwal, K., Kapoor, K., & Srivastava, J. (2017). Data mining techniques for social networks analysis. *Encyclopedia of social network analysis and mining* (pp. 1–11). New York, NY: Springer.

Aharonson, B., & Schilling, M. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. *Research Policy, 45,* 81–96.

Aickin, M. (2011). Patient-centered research from electronic medical records. *Permanent Journal, 15*, 89–91.

Akay, A., Dragomir, A., & Erlandsson, B.-E. (2015). Network-based modeling and intelligent data mining of social media for improving care. *IEEE Journal of Biomedical and Health Informatics, 19,* 210–218.

Aldenderfer, M., & Blashfield, R. (1984). *Cluster analysis.* Beverly Hills, CA: Sage.

Alderson, A., & Beckfield, J. (2004). Power and position in the world city system. *American Journal of Sociology, 109,* 811–851.

Alexander, M., & Danowski, J. (1990). Analysis of an ancient network: Personal communication and the study of social structure in a past society. *Social Networks, 12,* 313–335.

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior, 63,* 433–443.

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2017). Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method. *Physica A: Statistical Mechanics and its Applications, 468,* 278–288.

Allison, P. (2001). *Missing data.* Thousand Oaks, CA: Sage.

Alzahrani, T., & Horadam, K. (2016). Community detection in bipartite networks: algorithms and case studies. In *Complex systems and networks* (pp. 25–50). Berlin, Germany: Springer.

Anderson, J., & Jay, S. (1985). Computers and clinical judgement: The role of physician networks. *Social Science and Medicine, 20,* 969–979.

Andrevski, G., Brass, D. J., & Ferrier, W. J. (2016). Alliance portfolio configurations and competitive action frequency. *Journal of Management, 42,* 811–837.

Andris, C. (2016). Integrating social network data into GISystems. *International Journal of Geographical Information Science, 30,* 2009–2031.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 258–265).

Bailey, S., & Marsden, P. V. (1999). Interpretation and interview context: Examining the general social survey name generator using cognitive methods. *Social Networks, 21,* 287–309.

Barnes, C. M., Dang, C. T., Leavitt, K., Guarana, C. L., & Uhlmann, E. L. (2018). Archival data in micro-organizational research: A toolkit for moving to a broader set of topics. *Journal of Management, 44,* 1453–1478.

Barnes, J. (1954). Class and committees in a Norwegian island parish. *Human Relations, 7,* 39–58.

Barnes, J. (1979). Network analysis: Orienting notion, rigorous technique or substantive field of study. In P. W. Holland & S. Leinhardt (Eds.), *Perspectives on social network analysis* (pp. 403–423). New York, NY: Academic Press.

Basu, A. (2014). Social network analysis: A methodology for studying terrorism. In *Social networking* (pp. 215–242). Cham, Switzerland: Springer.

Batchelder, E. (2002). Comparing three simultaneous measurements of a sociocognitive network. *Social Networks, 24,* 261–277.

Benton, R. A. (2017). The decline of social entrenchment: Social network cohesion and board responsiveness to shareholder activism. *Organization Science, 28,* 262–282.

Berlusconi, G., Calderoni, F., Parolini, N., Verani, M., & Piccardi, C. (2016). Link prediction in criminal networks: A tool for criminal intelligence analysis. *Plos ONE, 11(4),* e0154244.

Berman, J. (2019). Big tech needs regulation, but DC must go to school before it goes to work. Vox. Accessed June 22, 2019. www.vox.com/recode/2019/6/14/18679675/big-tech-regulation-national-commission-technology-democracy

Bernard, H. R., Johnsen, E., Killworth, P., McCarty, C., Shelley, G., & Robinson, S. (1990). Comparing four different methods for measuring personal social networks. *Social Networks, 12,* 179–215.

Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology, 13,* 495–517.

Betancourt, N., Kovács, B., & Otner, S. (2018). The perception of status: How we infer the status of others from their social relationships. *Network Science*, *6,* 319–347.

Bliuc, A.-M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior,* 87, 75–86.

Blondel, V., Krings, G., & Thomas, I. (2010). Regions and borders of mobile telephony in Belgium and in the Brussels Metropolitan Zone. *Brussels Studies, 42,* 1–12.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10,* P10008 (12 pp.).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2011). The Louvain method for community detection in large networks. Accessed January 10, 2019. https://arxiv.org/pdf/1108.1502.pdf

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks, 29,* 555–564.

Boorman, S., & White, H. (1976). Social structure from multiple networks. II. Role structures. *American Journal of Sociology, 81,* 1384–1446.

Borg, I., Groenen, P. J. F., & Mair, P. (2017). *Applied multidimensional scaling and unfolding.* New York, NY: Springer.

Borgatti, S., & Everett, M. G. (1992). Notions of position in social network analysis. *Sociological Methodology, 22,* 1–35.

Borgatti, S., Everett, M. G.,, & Freeman, L. C. (2002). *UCINET 6 for Windows: Software for social network analysis.* Harvard, MA: Analytic Technologies.

Borgatti, S., Everett, M. G., & Shirey, P. (1990). LS sets, lambda sets and other cohesive subsets. *Social Networks*, *12,* 337–357.

Borgatti, S. P., (2005). Centrality and network flow. *Social Networks, 27,* 55–71.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks, 28,* 124–136.

Borgatti, S. P., & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks, 21,* 375–395.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks.* London, England: Sage.

Borgatti, S. P., & Molina, J. L. (2003). Ethical and strategic issues in organizational social network analysis. *Journal of Applied Behavioral Science, 39,* 337–349.

Bouchillon, B. C. (2018). Patching the melting pot: Sociability in Facebook groups for engagement, trust, and perceptions of difference. *Social Science Computer Review,* 1–20.

Box-Steffensmeier, J. M., & Christenson, D. P. (2015). Comparing membership interest group networks across space and time, size, issue and industry. *Network Science, 3,* 78–97.

Boyd, D. M., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13,* 210–230.

Brañas, A., Barrigón, M. L., Garrido-Torres, N., Perona-Garcelán, S., Rodriguez-Testal, J. F., Lahera, G., & Ruiz-Veguilla, M. (2016). U-shaped curve of psychosis according to cannabis use: New evidence from a snowball sample. *Journal of Psychopharmacology, 30,* 1331–1338.

Brands, R. A. (2013). Cognitive social structures in social network research: A review. *Journal of Organizational Behavior, 34*(S1), S82–S103.

Brashears, M. E. (2014). 'Trivial' topics and rich ties: The relationship between discussion topic, alter role, and resource availability using the 'important matters' name generator. *Sociological Science, 1,* 493.

Brayne, S. (2017). Big Data surveillance: The case of policing. *American Sociological Review, 82,* 977–1008.

Burgin, R. (1995). The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of the American Society for Information Science, 46,* 562–572.

Burt, R. S. (1978). Cohesion versus structural equivalence as a basis for network subgroups. *Sociological Methods & Research, 7*(2), 189–212.

Burt, R. S. (1979). Disaggregating the effect on profits in manufacturing industries of having imperfectly competitive consumers and suppliers. *Social Science Research, 8,* 120–143.

Burt, R. S. (1983). Distinguishing relational contents. In R. S. Burt & M. J. Minor (Eds.), *Applied network analysis* (pp. 35–74). Beverly Hills, CA: Sage.

Burt, R. S. (1985). General social survey network items. *Connections, 8,* 119–123.

Burt, R. S. (1987). A note on missing network data in the general social survey. *Social Networks, 9,* 63–73.

Burt, R. S. (1992). *Structural holes: The social structure of competition.* Cambridge, MA: Harvard University Press.

Burt, R. S. (2002). Bridge decay. *Social Networks, 24,* 333–363.

Cai, T. (2016). IGI Global. In *Business intelligence: Concepts, methodologies, tools, and applications* (pp. 543–557). Hershey, PA: IGI Global.

Calderoni, F., Brunetto, D., & Piccardi, C. (2017). Communities in criminal networks: A case study. *Social Networks, 48,* 116–125.

Campbell, K. E., & Lee, B. (1991). Name generators in surveys of personal networks. *Social Networks, 13,* 203–221.

Carley, K., & Krackhardt, D. (1996). Cognitive inconsistencies and non-symmetric friendship. *Social Networks, 18,* 1–27.

Carpenter, D., Esterling, K., & Lazer, D. (2004). Friends, brokers, and transitivity: Who informs whom in Washington politics. *Journal of Politics, 66,* 224–246.

Carrington, P. J., & Scott, J. (2011). Introduction. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 1–8). London, England: Sage.

Casciaro, T. (1998). Seeing things clearly: Social structure, personality, and accuracy in social network perception. *Social Networks, 20,* 331–351.

Chambers, D., Wilson, P., Thompson, C., & Harden, M. (2012). Social network analysis in healthcare settings: A systematic scoping review. *PloS one, 7*(8), e41911.

Child, J. (2005). *Organization: Contemporary principles and practice.* Chichester, West Sussex: Wiley.

Choe, H., Lee, D., Kim, H. D., & Seo, I. W. (2016). Structural properties and inter-organizational knowledge flows of patent citation network: The case of organic solar cells. *Renewable and Sustainable Energy Reviews, 55,* 361–370.

Chopade, P., Zhan, J., & Bikdash, M. (2015). Node attributes and edge structure for large-scale Big Data network analytics and community detection. *IEEE International Symposium on Technologies for Homeland Security.* Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Christakis, N. A., & Fowler. J. H. (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine, 358,* 2249–2258.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. SIGMOD'16. San Francisco, CA.

Coleman, J. S. (1986). Social theory, social research, and a theory of action. *American Journal of Sociology, 91,* 1309–1335.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using Big Data. *Journal of Communication, 64,* 317–332.

Crane, H., & Dempsey, W. (2017). Edge exchangeable models for network data. *Journal of the American Statistical Association, 113,* 1311–1326.

Crossley, N., Bellotti, E., Edwards, G., Everett, M. G., Koskinen, J., & Tranme, M. (2015). *Social network analysis for ego-nets.* London, England: Sage.

Crossley, N., Edwards, G., Harries, E., & Stevenson, R. (2012). Covert social movement networks and the secrecy-efficiency trade off: The case of the UK suffragettes (1906–1914). *Social Networks, 34,* 634–644.

Crossley, N., & Krinsky, J. (2016). *Social networks and social movements: Contentious connections.* New York, NY: Routledge.

Cui, A. S., Calantone, R. J., & Griffith, D. A. (2011). Strategic change and termination of interfirm partnerships. *Strategic Management Journal, 32,* 402–423.

Cullen, B., Mojtabai, R., Bordbar, E., Everett, A., Nugent, K., & Eaton, W. (2017). Social network, recovery attitudes and internal stigma among those with serious mental illness. *International Journal of Social Psychiatry, 63,* 448–458.

Curtis, B. (2014). Social networking and online recruiting for HIV research: Ethical challenges. *Journal of Empirical Research on Human Research Ethics, 9,* 58–70.

Daniel, J., Silva, R., Santos, A., Cardoso, J., Coelho, L., Freitas, M., & Ribeiro, O. (2017). Perception accuracy of affiliative relationships in elementary school children and young adolescents. *Frontiers in Psychology, 8,* 1936.

Davis, A., Gardner, B. B., & Gardner, M. R. (1941). *Deep south: A social anthropology of caste and class.* Chicago, IL: University of Chicago Press.

Davis, J. A. (1979). The Davis/Holland/Leinhardt studies: An overview. In P. W. Holland & S. Leinhardt (Eds.), *Perspectives on social network research* (pp. 51–62). New York, NY: Academic Press.

Delgado-Márquez, B. L., Hurtado-Torres, N. E., Pedauga, L. E., & Cordón-Pozo, E. (2018). A network view of innovation performance for multinational corporation subsidiaries. *Regional Studies, 52,* 47–67.

Dempsey, W., Oselio, B., & Hero, A. (2019). Hierarchical network models for structured exchangeable interaction processes. arXiv preprint arXiv:1901.09982.

De Nooy, W., Mrvar, A., & Batagelj, V. (2018). *Exploratory social network analysis with Pajek: Revised and expanded edition for updated software* (3rd ed.). Cambridge, England: Cambridge University Press.

Desmarais, B. A., & Cranmer, S. J. (2012). Statistical inference for valued-edge networks: The generalized exponential random graph model. *PloS ONE, 7,* e30136.

Diani, M. (2004). Networks and participation. In *The Blackwell companion to social movements* (pp. 339–359). Hoboken, NJ: Wiley.

Diani, M. (2016). Social networks and social movements. *European Political Science, 15,* 266–267.

Diviák, T., Dijkstra, J., & Snijders, T. A. B. (2019). Structure, multiplexity, and centrality in a corruption network: The Czech Rath affair. *Trends in Organized Crime, 22,* 274–297.

Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized block modeling.* Cambridge, England: Cambridge University Press.

Doreian, P., & Woodard, K. (1992). Fixed list versus snowball selection of social networks. *Social Science Research, 21,* 216–233.

Doreian, P., & Woodard, K. (1994). Defining and locating cores and boundaries of social networks. *Social Networks, 16,* 267–293.

Dorogovtsev, S. N., & Mendes, J. F. F. (2014). *Evolution of networks: From biological nets to the Internet and WWW.* Oxford, England: Oxford University Press.

Dowdle, A., Limbocker, S., Yang S., Sebold, K., & Stewart, P. A. (2013). *The invisible hands of political parties in presidential elections: Party activists and political aggregation from 2004 to 2012.* New York, NY: Palgrave Macmillan.

Du, R., Yu, Z., Mei, T., Wang, Z., Wang, Z., & Guo, B. (2014). Predicting activity attendance in event-based social networks: Content, context and social influence. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 425–434). New York, NY: Association for Computing Machinery.

Dunbar, R. I. (2016). Do online social media cut through the constraints that limit the size of offline social networks? *Royal Society Open Science, 3,* 150292.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world.* Cambridge, England: Cambridge University Press.

Ebadi, A., & Schiffauerova, A. (2016). How to boost scientific production? A statistical analysis of research funding and other influencing factors. *Scientometrics, 3,* 1093–1116.

Edwards, G., & Crossley, N. (2009). Measures and meanings: Exploring the ego-net of Helen Kirkpatrick Watts, militant suffragette. *Methodological Innovations Online, 4,* 37–61.

Ekblaw, A., Azaria, A., Halamka, J., & Lippman, A. (2016). A case study for blockchain in healthcare: 'MedRec' prototype for electronic health records and medical research data. In *Proceedings of IEEE Open & Big*

Data Conference (Vol. 13, p. 13). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology, 103,* 281–317.

Erdogan, B., Bauer, T., & Walter, J. (2015). Deeds that help and words that hurt: Helping and gossip as moderators of the relationship between leader-member exchange and advice network centrality. *Personnel Psychology, 68,* 185–214.

Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae, 6,* 290–297.

Erickson, B. H. (1996). Culture, class, and connections. *American Journal of Sociology, 102,* 217–251.

Erickson, B. H. (2004). The distribution of gendered social capital in Canada. In H. Flap & B. Volker (Eds.), *Creation and returns of social capital: A new research program* (pp. 27–51). New York, NY: Routledge.

Everett, M. (1985). Role similarity and complexity in social networks. *Social Networks, 7,* 353–359.

Everett, M. G., & Borgatti, S. P. (2013). The dual-projection approach for two-mode networks. *Social Networks, 35,* 204–210.

Everett, M., Boyd, J., & Borgatti, S. (1990). Ego-centered and local roles: A graph theoretic approach. *Journal of Mathematical Sociology, 15,* 163–172.

Falci, C., & McNeely, C. (2009). Too many friends: Social integration, network cohesion and adolescent depressive symptoms. *Social Forces, 87,* 2031–2061.

Fang, R., Duffy, M., & McAllister, D. J. (2015). When do coworkers undermine? The variable costs of having friends in high places. In *Academy of Management Proceedings*. Briarcliff Manor, NY: Academy of Management.

Fang, R., McAllister, D. J., & Duffy, M. (2017). Your connections made me undermine you: Provocative effect of coworker friendship contact status. In *Academy of Management Proceedings.* Briarcliff Manor, NY: Academy of Management.

Faust, K. (1988). Comparison of methods for positional analysis: Structural and general equivalences. *Social Networks, 10,* 313–341.

Faust, K. (1997). Centrality in affiliation networks. *Social Networks, 19,* 157–191.

Feehan, D. M., & Salganik, M. J. (2016). Generalizing the network scale-up method: A new estimator for the size of hidden populations. *Sociological Methodology, 46,* 153–186.

Feldman-Savelsberg, P., Ndonko, F., & Yang, S. (2005). How rumor begets rumor: Collective memory, ethnic conflict, and reproductive rumors in Cameroon. In G. A. Fine, V. Campion-Vincent, & C. Heath (Eds.), *Rumor mills: The social impact of rumor and legend* (pp. 141–159). New York, NY: Transaction Press.

Fernández-Avilés, G., & Montero, J.-M. (2016). Spatio-temporal modeling of financial maps from a joint multidimensional scaling-geostatistical perspective. *Expert Systems with Applications, 60,* 280–293.

Fischer, C. S. (1982). What do we mean by "friend"? An inductive study. *Social Networks, 3,* 287–306.

Fischer, C. S. (2009). The 2004 GSS finding of shrunken social networks: An artifact? *American Sociological Review, 74,* 657–669.

Fischer, C. S., & Shavit, Y. (1995). National differences in network density: Israel and the United States. *Social Networks, 17,* 129–145.

Folch-Fortuny, A., Villaverde, A. F., Ferrer, A., & Banga, J. R. (2015). Enabling network inference methods to handle missing data and outliers. *BMC Bioinformatics, 16,* 283. Accessed August 3, 2018. www.ncbi.nlm.nih.gov/pmc/articles/PMC4559359

Fortunato, S. (2010). Community detection in graphs. *Physics Report, 486,* 75–174.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2017). *Big Data and social science: A practical guide to methods and tools.* Boca Raton, FL: CRC Press.

Frank, O. (2005). Network sampling and model fitting. In P. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 31–45). New York, NY: Cambridge University Press.

Frank, O. (2011). Survey sampling in networks. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 389–403). London, England: Sage.

Freeman, L. C. (1979). Centrality in social networks: I. Conceptual clarification. *Social Networks, 1,* 215–239.

Freeman, L. C. (2000). Visualizing social networks. *Journal of Social Structure, 1,* 1–15.

Freeman, L. C. (2003). Finding social groups: A meta-analysis of the Southern Women data. In R. L. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic social network modeling and analysis: Workshop summary and papers* (pp. 39–97). Washington, DC: National Research Council, The National Acadamies Press.

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science.* Vancouver, Canada: Empirical Press.

Freeman, L. C. (2005). Graphic techniques for exploring social network data. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 248–270). Cambridge, MA: Cambridge University Press.

Freeman, L. C. (2011). The development of social network analysis—With an emphasis on recent events. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 26–39). London, England: Sage.

Freeman, L. C., Borgatti, S., & White, D. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks, 13,* 141–154.

Freeman, L. C., & Thompson, C. (1989). Estimating acquaintanceship volume. In M. Kochen (Ed.), *The small world* (pp. 147–158). Norwood, NJ: Ablex.

Freeman, L. C., & Webster, C. (1994). Interpersonal proximity in social and cognitive space. *Social Cognition, 12,* 223–247.

Frith, J. (2014). Communicating through location: The understood meaning of the foursquare check-in. *Journal of Computer-Mediated Communication, 19,* 890–905.

Gainforth, H. L., Latimer-Cheung, A. E., Athanasopoulos, P., Moore, S., & Ginis, K. A. M. (2014). The role of interpersonal communication in the process of knowledge mobilization within a community-based organization: A network analysis. *Implementation Science, 9,* 59.

Galaskiewicz, J. (1979). The structure of community organizational networks. *Social Forces, 57,* 1346–1364.

Gao, J., Nyhan, J., Duke-Williams, O., & Mahony, S. (2018). Visualising the digital humanities community: A comparison study between citation network and social network. In *Digital humanities 2018 Puentes-Bridges: Book of abstracts* (pp. 201–204). Mexico City: Red de Humanidades Digitales A.C.

Garlaschelli, D., & Loffredo, M. (2005). Structure and evolution of the world trade network. *Physica A, 355,* 138–144.

Gegov, E., Postorino, M. N., Atherton, M., & Gobet, F. (2013). Community structure detection in the evolution of the United States airport network. *Advances in Complex Systems, 16(1),* 1350003.

Gillath, O., Karantzas, G. C., & Selcuk, E. (2017). A net of friends: Investigating friendship by integrating attachment theory and social network analysis. *Personality and Social Psychology Bulletin, 43,* 1546–1565.

Glover, F. (1989). Tabu Search – Part I. *ORSA Journal on Computing, 1,* 190–206.

Glover, F. (1990). Tabu Search – Part II. *ORSA Journal on Computing, 2,* 4–32.

González-Bailón, S., Wang, N., River, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks, 38,* 16–27.

Goodman, L. (1961). Snowball sampling. *Annals of Mathematical Statistics, 32,* 148–170.

Grabowicz, P. A., Asco, J. J. R., Moro, E., Pujol, J. M., & Eguiluz, V. M. (2012). Social features of online networks: The strength of intermediary ties in online social media. *PLoS ONE, 7*(1), e29358.

Grandjean, M. (2015). Introduction à la visualisation de données: l'analyse de réseau en histoire [Introduction to data visualization: Network analysis in history]. *Geschichte und Informatik, 18,* 109–128.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78,* 1360–1380.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology, 91,* 481–510.

Greetham, D. V., Hurling, R., Osborne, G., & Linley, A. (2011). Social networks and positive and negative affect. *Procedia-social and Behavioral Sciences, 22,* 4–13.

Guo, L., Rohde, J. A., & Wu, H. D. (2018). Who is responsible for Twitter's echo chamber problem? Evidence from 2016 US election networks. *Information, Communication & Society, 21,* 1–18.

Hafner-Burton, E. M., Kahler, M., & Montgomery, A. H. (2009). Network analysis for international relations. *International Organization, 63,* 559–592.

Hällsten, M., Edling, C., & Rydgren, J. (2015). The effects of specific occupations in position generator measures of social capital. *Social Networks, 40,* 55–63.

Hansen, D., Shneiderman, B., & Smith, M. A. (2011). *Analyzing social media networks with NodeXL: Insights from a connected world.* New York, NY: Morgan Kaufmann.

Hargittai, E. (2018). Potential biases in Big Data: Omitted voices on social media. *Social Science Computer Review,* published electronically July 30, doi:10.1177/0894439318788322.

Harris, J. H. (2014). *An introduction to exponential random graph modeling.* Thousand Oaks, CA: Sage.

Haynes, C., & Sattler, J. (2017). The Twitter effect: How Trump used social media to stamp his brand and shape the media narrative on immigration. Accessed October 24, 2018. https://digitalcommons.newhaven.edu/politicalscience-facpubs/9

Healy, K., & Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology, 40,* 105–128.

Heaney, M. T. (2014). Multiplex networks and interest group influence reputation: An exponential random graph model. *Social Networks, 36,* 66–81.

Heckathorn, D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44,* 174–199.

Heckathorn, D. D., & Cameron, C. J. (2017). Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual Review of Sociology, 43,* 101–119.

Heemskerk, E. M., Fennema, M., & Carroll, W. K. (2016). The global corporate elite after the financial crisis: Evidence from the transnational network of interlocking directorates. *Global Networks, 16,* 68–88.

Heider, F. (1958). *The psychology of interpersonal relations.* New York, NY: Wiley.

Ho, J. C., Saw, E. C., Lu, L. Y., & Liu, J. S. (2014). Technological barriers and research trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Change, 82,* 66–79.

Ho, M. H., & Cheo, H. Y. (2014). Analyzing the brokerage roles of stakeholders in a technological network: A study of GMO plant technologies. In *Management of Engineering & Technology (PICMET)* (pp. 3144–3154). 2014 Portland International Conference. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Hogan, T. M., Olade, T. O., & Carpenter, C. R. (2014). A profile of acute care in an aging America: Snowball sample identification and characterization of United States geriatric emergency departments in 2013. *Academic Emergency Medicine, 21,* 337–346.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association, 76,* 33–65.

Hollywood, J. S., Vermeer, M. J., Woods, D., Goodison, S. E., & Jackson, B. A. (2018). *Using social media and social network analysis in law enforcement*. Santa Monica, CA: RAND Corporation.

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science, 4,* 93–103.

Hsung, R.-M., Fu,Y.-c., & Lin, N. (2017). The position generator: Measurement techniques for investigations of social capital. In N. Lin, K. Cook, & R. S. Burt (Eds.), *Social capital: Theory and research* (pp. 57–81). Hawthorne, NY: Aldine de Gruyter.

Huckfeldt, R., & Sprague, J. (1987). Networks in context: The social flow of political information. *American Political Science Review, 81,* 1197–1216.

Huisman, M. (2014). Imputation of missing network data: Some simple procedures. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of social network analysis and mining* (Vol. 1, pp. 707–715). Berlin, Germany: Springer.

Huisman, M., & Steglich, C. (2008). Treatment of non-response in longitudinal network studies. *Social Networks, 30,* 297–308.

Hussain, I., & Asghar, S. (2017). A survey of author name disambiguation techniques: 2010 to 2016. *Knowledge Engineering Review,* 32.

Ignatow, G., & Mihalcea, R. F. (2018). *An introduction to text mining: Research design, data collection, and analysis*. Thousand Oaks, CA: Sage.

Johnson, J., & Orbach, M. (2002). Perceiving the political landscape: Ego biases in cognitive political networks. *Social Networks, 24,* 291–310.

Junco, R. (2012). The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers & Education, 58,* 162–171.

Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings.* New York, NY: Oxford University Press.

Kamat, P., & Gautam, A. S. (2018). Recent trends in the era of cybercrime and the measures to control them. In *Handbook of e-business security* (pp. 243–258). Auerbach Publications. Milton Park, England: Taylor & Francis.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation, 2,* 10.

Killworth, P. D., & Bernard, H. R. (1978). The reversal small-world experiment. *Social Networks, 1,* 159–192.

Killworth, P. D., Bernard, H. R., McCarty, C., Doreian, P., Goldenberg, S., Underwood, C., . . . Von Sury Wemegah, M. (1984). Measuring patterns of acquaintanceship. *Current Anthropology, 25,* 381–392.

Killworth, P. D., Johnsen, E. C., Bernard, H. R., Shelley, G. A., & McCarty, C. (1990). Estimating the size of personal networks. *Social Networks, 12,* 289–312.

Killworth, P. D., Johnsen, E. C., McCarty, C., Shelley, G. A., & Bernard, H. R. (1998). A social network approach to estimating seroprevalence in the United States. *Social Networks, 20,* 23–50.

Kim, D. H., Lee, B. Y., & Sohn, S. Y. (2016). Quantifying technology-industry spillover effects based on patent citation network analysis of unmanned aerial vehicle (UAV). *Technological Forecasting and Social Change, 105,* 140–157.

King, P., & Howard, J. (2014). Children's perceptions of choice in relation to their play at home, in the school playground and at the out-of-school club. *Children & Society, 28,* 116–127.

Kleinbaum, A. M. (2017). Reorganization and tie decay choices. *Management Science, 64,* 2219–2237.

Kmetty, Z., Tardos, R., Albert, F., & Dávid, B. (2018). Mapping social milieus and cohesion patterns between 1997 and 2014: Exploiting the potential of the occupational position generator. *Social Networks, 55,* 116–129.

Knoke, D. (1990). Networks of political action: Toward theory construction. *Social Forces, 68,* 1041–1063.

Knoke, D. (2001). *Changing organizations: Business networks in the new political economy.* Boulder, CO: Westview Press.

Knoke, D. (2009). Playing well together: Creating corporate social capital in strategic alliance networks. *American Behavioral Scientist, 52,* 1690–1708.

Knoke, D., & Burt, R. S. (1983). Prominence. In R. S. Burt & M. J. Miner (Eds.), *Applied network analysis: A methodological introduction* (pp. 195–222). Beverly Hills, CA: Sage.

Knoke, D. & Kuklinski, J. H. (1982). *Network analysis.* Beverly Hills, CA: Sage.

Knoke, D., & Laumann, E. O. (1982). The social structure of national policy domains: An exploration of some structural hypotheses. In P. V. Marsden & N. Lin (Eds.), *Social structure and network analysis* (pp. 255–270). Beverly Hills, CA: Sage.

Knoke, D., & Rogers, D. (1979). A blockmodel analysis of interorganizational networks. *Sociology and Social Research, 64,* 28–52.

Knoke, D., & Yang, S. (2008). *Social network analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Knox, H., Savage, M., & Harvey, P. (2006). Social networks and the study of relations: Networks as method, metaphor and form. *Economy and Society, 35,* 113–140.

Kojaku, S., Cimini, G., Caldarelli, G., & Masuda, N. (2018). Structural changes in the interbank market across the financial crisis from multiple core-periphery analysis. *Journal of Network Theory in Finance, 4*(3), 33–51.

König, M., Liu, X., & Zenou, Y. (2014). *R&D networks: Theory, empirics and policy implications.* Department of Economic Working Paper No. 412. University of Zurich. Accessed August 3, 2018. https://www.econstor.eu/bitstream/10419/111199/1/econwp142.pdf

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks, 28,* 247–268.

Kowald, M., van den Berg, P., Frei, A., Carrasco, J.-A., Arentze, T., Axhausen, K., . . . Wellman, B. (2013). Distance patterns of personal networks in four countries: A comparative study. *Journal of Transport Geography, 31,* 236–248.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks, 9,* 109–134.

Krackhardt, D., & Kilduff, M. (1999). Whether close or far: Perceptions of balance in friendship networks in organizations. *Journal of Personality and Social Psychology, 76,* 770–782.

Krause, R. W., Huisman, M., & Snijders, T. A. B. (2018). Multiple imputation for longitudinal network data. *Italian Journal of Applied Statistics,* 30, 33–58.

Kreager, D. A., Molloy, L. E., Moody, J., & Feinberg, M. E. (2016). Friends first? The peer network origins of adolescent dating. *Journal of Research on Adolescence, 26,* 257–269.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling.* Beverly Hills, CA: Sage.

Kumbasar, E., Romney, A. K., & Batchelder, W. H. (1994). Systematic biases in social perception. *American Journal of Sociology, 100,* 477–505.

Labatut, V., Dugué, N., & Perez, A. (2014). Identifying the community roles of social capitalists in the Twitter network. Advances in Social Networks IEEE/ACM International Conference on Analysis and Mining. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E, 80,* 056117.

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc. Accessed February 6, 2018. http://blogs.gartner.com/doug-laney/files/2012/01/ad949–3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Larsen, A. G., & Ellersgaard, C. H. (2017). Identifying power elites—K-cores in heterogeneous affiliation networks. *Social Networks, 50,* 55–69.

Larson, J. M., & Lewis, J. I. (2017). Ethnic networks. *American Journal of Political Science, 61,* 350–364.

Laumann, E. O., Marsden, P. V., & Prensky, D. (1983). The boundary specification problem in network analysis. In R. S. Burt & M. Minor (Eds.), *Applied network analysis* (pp. 18–34). Beverly Hills, CA: Sage.

Laumann, E. O., Marsden, P. V., & Prensky, D. (1989). The boundary specification problem in network analysis. In L. C. Freeman, D. R. White, & A. K. Romney (Eds.), *Research methods in social network analysis* (pp. 61–87). Fairfax, VA: George Mason University Press.

Lazega, E., & Snijders, T. A. B. (Eds.). (2015). *Multilevel network analysis for the social sciences: Theory, methods and applications* (Vol. 12). Berlin, Germany: Springer.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Schudson, M. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.

Lee, H., & Sohn, I. (2015). *Fundamentals of Big Data network analysis for research and industry.* Indianapolis, IN: Wiley.

Lee, S. (2017). *Alliance networks, corporate investment, and firm valuation*. Atlanta, GA: Georgia State University dissertation. Accessed August 3, 2018. https://scholarworks.gsu.edu/finance_diss/32.

Leech, N. L., Collins, K. M. T., & Onwuegbuzie, A. J. (2017). Collecting qualitative data to enhance social network analysis and data mining. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of social network analysis and mining* (Vol. 1, pp. 124–133). Berlin, Germany: Springer.

Leifeld, P., & Cranmer, S. J. (2019). A theoretical and empirical comparison of the temporal exponential random graph model and the stochastic actor-oriented model. *Network Science, 7,* 20–51.

Leifeld, P., Cranmer, S. J., & Desmarais, B. A. (2018). Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *Journal of Statistical Software, 83*(6), 1–36.

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*. New York, NY: Association for Computing Machinery.

Levin, F. H., & Heuser, C. A. (2010). Evaluating the use of social networks in author name disambiguation in digital libraries. *Journal of Information and Data Management, 1(2),* 183–193.

Li, M., Westerholt, R., Fan, H., & Zipf, A. (2018). Assessing spatiotemporal predictability of LBSN: A case study of three foursquare datasets. *GeoInformatica, 22,* 541–561.

Lien, C.-H., Cao, Y., & Zhou, X. (2017). Service quality, satisfaction, stickiness, and usage intentions: An exploratory evaluation in the context of WeChat services. *Computers in Human Behavior*, *68,* 403–410.

Lin, N. (1982). Social resources and instrumental action. In P. V. Marsden & N. Lin (Eds.), *Social structure and network analysis* (pp. 131–145). Beverly Hills, CA: Sage.

Lin, N., & Dumin, M. (1986). Access to occupations through social ties. *Social Networks, 8,* 365–385.

Lin, N., Fu, Y.-c., & Hsung, R.-M. (2001). Measurement techniques for investigations of social capital. In N. Lin, K. Cook, & R. S. Burt (Eds.), *Social capital: Theory and research* (pp. 57–81). New Brunswick, NJ: Transaction.

Liu, X., Patacchini, E., Zenou, Y., & Lee, L.-F. (2012). *Criminal networks: Who is the key player?* CEPR Discussion Paper No. 8772.

Liu, Y., Tang, M., Zhou, T., & Do, Y. (2015). Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics. *Scientific Reports, 5,* 13172. doi:10.1038/srep13172

Logan, J. J., Jolly, A. M., & Blanford, J. I. (2016). The sociospatial network: Risk and the role of place in the transmission of infectious diseases. *PloS ONE, 11*(2), e0146915.

Lohr, S. (2013, February 1). The origins of 'Big Data': An etymological detective story. *New York Times*. Accessed August 10, 2018. https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story

Lu, W., & Hampton, K. N. (2017). Beyond the power of networks: Differentiating network structure from social media affordances for perceived social support. *New Media & Society, 19,* 861–879.

Ludkin, M., Eckley, I., & Neal, P. (2017). Dynamic stochastic block models: Parameter estimation and detection of changes in community structure. *Statistics and Computing, 28,* 1201–1213.

Lusher, D., Koskinen, J., & Robin, G. (2012). *Exponential random graph models for social networks: Theory, methods, and applications.* Cambridge, England: Cambridge University Press.

Macedo, A. Q., Marinho, L. B., & Santos, R. L. T. (2015). Context-aware event recommendation in event-based social networks. In *Proceedings of the 9th ACM Conference on Recommender Systems* (pp. 123–130). New York, NY: Association for Computing Machinery.

Madden, M., Gilman, M., Levy, K., & Marwick, A. (2017). Privacy, poverty, and Big Data: A matrix of vulnerabilities for poor Americans. *Washington University Law Review, 95,* 53–125.

Maggioni, M. A., Breschi, S., & Panzarasa, P. (2013). Multiplexity, growth mechanisms and structural variety in scientific collaboration networks. *Industry and Innovation, 20,* 185–194.

Maltiel, R., Raftery, A. E., McCormick, T. H., & Baraff, A. J. (2015). Estimating population size using the network scale up method. *Annals of Applied Statistics, 9,* 1247–1277.

Maoz, Z. (2012). Preferential attachment, homophily, and the structure of international networks, 1816–2003. *Conflict Management and Peace Science, 29,* 341–369.

Marcussen, C. H. (2017). Visualising the network of cruise destinations in the Baltic Sea: A multidimensional scaling approach. *Scandinavian Journal of Hospitality and Tourism, 17,* 208–222.

Marineau, J., Brass, D. J., Borgatti, S., & Vecchi, P. (2018). Individuals' formal power and their social network accuracy: A situated cognition perspective. *Social Networks, 54,* 145–161.

Marsden, P. V. (1987). Core discussion networks of Americans. *American Sociological Review, 52,* 122–131.

Marsden, P. V. (1993). The reliability of network density and composition measures. *Social Networks, 15,* 399–421.

Marsden, P. V. (2005). Recent developments in network measurement. In P. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 8–30). New York, NY: Cambridge University Press.

Marsden, P. V. (2011). Survey methods for network data. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 370–388). London, England: Sage.

Martí, J., Bolíbar, M., & Lozares, C. (2017). Network cohesion and social support. *Social Networks, 48,* 192–201.

Mashey, J. R. (1998). Big Data . . . and the next wave of InfraStress. Accessed August 10, 2018. http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf.

Masuda, Y. J., Liu, Y., Reddy, S. M. W., Frank, K. A., Burford, K., Fisher, J. R. B., & Montambault, J. (2018). Innovation diffusion within large environmental NGOs through informal network agents. *Nature Sustainability, 1,* 190.

Mazie, S. (2018). Do you have too many Facebook friends? Accessed August 3, 2018. https://bigthink.com/praxis/do-you-have-too-many-facebook-friends.

McCarty, C., Bernard, H. R., Killworth, P. D., Shelley, G. A., & Johnsen, E. C. (1997). Eliciting representative samples of personal networks. *Social Networks, 19,* 303–323.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., & Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization, 60,* 28–39.

McCormick, T. H., Salganik, M. J., & Zheng, T. (2010). How many people do you know? Efficiently estimating personal network size. *Journal of the American Statistical Association, 105,* 59–70.

McKay, A. S., Grygiel, P., & Karwowski, M. (2017). Connected to create: A social network analysis of friendship ties and creativity. *Psychology of Aesthetics, Creativity, and the Arts, 11,* 284–294.

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in America: Changes in core discussion networks over two decades. *American Sociological Review, 71,* 353–375.

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2009). Models and marginals: Using survey evidence to study social networks. *American Sociological Review, 74,* 670–681.

Medders, R. B., & Metzger, M. J. (2018). The role of news brands and leads in exposure to political information on the Internet. *Digital Journalism, 6*, 599–618.

Mendieta, J. G., Schmidt, S., Aastro, J., & Ruiz, A. A. (1997). A dynamic analysis of the Mexican power network. *Connections, 20*(2), 34–55.

Merrin, G. J., De La Haye, K., Espelage, D. L., Ewing, B., Tucker, J. S., Hoover, M., & Green, H. D. (2018). The co-evolution of bullying perpetration, homophobic teasing, and a school friendship network. *Journal of Youth and Adolescence, 47,* 601–618.

Merton, R. K. (1957). The role-set: Problems in sociological theory. *British Journal of Sociology, 8,* 106–120.

Metzger, M. J., Hartsell, E. H., & Flanagin, A. J. (2015). Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research, 30,* 414–431.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics, 49,* 93–123.

Mitchell, J. C. (1969). *Social networks in urban situations: Analyses of personal relationships in central African towns.* Manchester, England: Manchester University Press.

Mohdeb, D., Boubetra, A., & Charikhi, M. (2016). Tie persistence in academic social networks. *Informatica, 40,* 353–364.

Moreno, J. L. (1934). *Who shall survive?* Washington, DC: Nervous and Mental Disease.

Morimoto, S. A., & Yang, S. (2013). What friendship entails: An empirical analysis of graduate students' social networks. *Sociological Spectrum, 33,* 99–116.

Morrissey, J. P., Tausig, M., & Lindsey, M. L. (1985). *Network analytic methods for mental health service system research: A comparison of two community support systems.* Washington, DC: National Institute of Mental Health.

Mustanski, B., Birkett, M., Kuhns, L. M., Latkin, C. A., & Muth, S. Q. (2015). The role of geographic and network factors in racial disparities in HIV among young men who have sex with men: An egocentric network study. *AIDS and Behavior, 19,* 1037–1047.

Nadel, S. F. (1957). *The theory of social structure*. London, England: Cohen and West.

NasehiMoghaddam, S., & Ghazanfari, M. (2016). Structure and causality in the world system. *Social Network Analysis and Mining, 6,* 73.

Neal, J. W., Cappella, E., Wagner, C., & Atkins, M. S. (2011). Seeing eye to eye: Predicting teacher-student agreement on classroom social networks. *Social Development, 20,* 376–393.

Newcomb, T. (1961). *The acquaintance process.* New York, NY: Holt, Rinehart & Winston.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98,* 404–409.

Newman, M. E. J. (2006a). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103,* 8577–8582.

Newman, M. E. J. (2006b). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E, 74,* 036104.

Newman, M. E. J. (2010). *Networks: An introduction.* Oxford, England: Oxford University Press.

Newman, M. E. J. (2017). Measurement errors in network data. arXiv preprint arXiv:1703.07376.

Newman, M. E. J. (2018a). Network reconstruction and error estimation with noisy network data. arXiv preprint arXiv:1803.02427.

Newman, M. E. J. (2018b). Network structure from rich but noisy data. *Nature Physics, 14,* 542–545.

Ng, A. (2018). Facebook deleted 583 million fake accounts in the first three months of 2018. Accessed August 4, 2018. www.cnet.com/news/facebook-deleted-583-million-fake-accounts-in-the-first-three-months-of-2018

Niezink, N. M. D., Snijders, T. A. B., & van Duijn, M. (2019). No longer discrete: Modeling the dynamics of social networks and continuous behavior. *Sociological Methodology, 49,* 0081175019842263.

Norbutas, L., & Corten, R. (2018). Network structure and economic prosperity in municipalities: A large-scale test of social capital theory using social media data. *Social Networks, 52,* 120–134.

Nowell, B. L., Velez, A.-L. K., Hano, M. C., Sudweeks, J., Albrecht, K., & Steelman, T. (2018). Studying networks in complex problem domains: Advancing methods in boundary specification. *Perspectives on Public Management and Governance, 1,* 273–282.

Oberst, U., Wegmann, E., Stodt, B., Brand, M., & Chamarro, A. (2017). Negative consequences from heavy social networking in adolescents: The mediating role of fear of missing out. *Journal of Adolescence, 55,* 51–60.

O'Gorman, M. J. (2018). Donald Trump, naturally: Revisionist environmental policy, global warming Tweets, and the unexpected emphasis on climate change in the 2016 campaign. In *The role of Twitter in the 2016 US election* (pp. 75–95). Cham, Switzerland: Palgrave Pivot.

Olivella, S., Pratt, T., & Imai, K. (2018). *Dynamic stochastic blockmodel regression for social networks: Application to international conflicts.* Paper presented at PolMeth 2018 meetings. Provo, UT: Brigham Young University.

O'Reagan, D., & Fleming, L. (2018). The FinFET breakthrough and networks of innovation in the semiconductor industry, 1980–2005: Applying digital tools to the history of technology. *Technology and Culture, 59,* 251–288.

Orman, G. K., Labatut, V., & Cherifi, H. (2011). On accuracy of community structure discovery algorithms. *Journal of Convergence Information Technology, 6,* 283–292.

Orman, G. K., Labatut, V., & Cherifi, H. (2012). Comparative evaluation of community detection algorithms: A topological approach. *Journal of Statistical Mechanics: Theory and Experiment, 8,* P08001.

Ouellet, M., Bouchard, M., & Hart, M. (2017). Criminal collaboration and risk: The drivers of Al Qaeda's network structure before and after 9/11. *Social Networks, 51,* 171–177.

Padgett, J. F., & Ansell, C. K. (1993). Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology, 98,* 1259–1319.

Pallotti, F., Tubaro, P., Casilli, A. A., & Valente, T. W. (2018). 'You see yourself like in a mirror': The effects of internet-mediated personal networks on body image and eating disorders. *Health Communication, 33,* 1166–1176.

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences, 113,* 566–571.

Papachristos, A. V., Braga, A. A., & Hureau, D. M. (2012). Social networks and the risk of gunshot injury. *Journal of Urban Health, 89,* 992–1003.

Park, A. J., Tsang, H. H., & Brantingham, P. L. (2012). Dynalink: A framework for dynamic criminal network visualization. In *Intelligence and Security Informatics Conference (EISIC), 2012 European* (pp. 217–224). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Parsons, T. (1949). *Essays in sociological theory: Pure and applied.* New York, NY: Free Press.

Pattison, P. (1988). Network models: Some comments on papers in this special issue. *Social Networks, 10,* 383–411.

Perry, B. L., & Pescosolido, B. (2010). Functional specificity in discussion networks: The influence of general and problem-specific networks on health outcomes. *Social Networks, 32,* 345–357.

Porter, M. A., Mucha, P. J., Newman, M. E. J., & Warmbrand, C. M. (2005). A network analysis of committees in the U.S. House of Representatives. *Proceedings of the National Academy of Sciences of the United States of America, 102.20,* 7057–7062.

Press, G. (2013, May 9). A very short history of Big Data. *Forbes*. Accessed August 10, 2018. https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data

Purushotham, S., & Kuo, C.-C. J.. (2016). Personalized group recommender systems for location-and event-based social networks. *ACM Transactions on Spatial Algorithms and Systems (TSAS), 2*(4), 16.

Rafiei, D., Agichtein, E., Baeza-Yates, R., Kleinberg, J., & Leskovec, J. (2018). The shifting landscape of web search and mining: Past, present, and future. In *Companion proceedings of the Web Conference 2018* (p. 881). Geneva, Switzerland: International World Wide Web Conference Committee.

Reid, S. E. (2017). The curious case of loners: Social isolation and juvenile incarceration. *Legal and Criminological Psychology, 22,* 180–195.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In *Recommender systems handbook* (pp. 1–34). Boston, MA: Springer.

Riley, P., Wang, R., Wang, Y., & Feng, L. (2016). Global warming: Chinese narratives of the future. *Global Media and China, 1–2,* 12–31.

Rosen, N. O., Mooney, K., & Muise, A. (2017). Dyadic empathy predicts sexual and relationship well-being in couples transitioning to parenthood. *Journal of Sex & Marital Therapy, 43,* 543–559.

Roth, T. (2018). The influence of parents' social capital on their children's transition to vocational training in Germany. *Social Networks, 55,* 74–85.

Roversi, A. (2017). *Hate on the net: Extremist sites, neo-fascism on-line, electronic jihad.* New York, NY: Routledge.

Ruan, D. (1998). The content of the general social survey discussion networks: An exploration of general social survey discussion name generator in a Chinese context. *Social Networks, 20,* 247–264.

Ruan, D., Freeman, L. C., Dai, X., Pan, Y., & Zhang, W. (1997). On the changing structure of social networks in urban China. *Social Networks, 19,* 75–89.

Russell, M. A. (2011). *Mining the social web: Analyzing data from Facebook, Twitter, LinkedIn, and other social media sites.* Sebastopol, CA: O'Reilly Media.

Scheim, A. I., & Bauer, G. R. (2015). Sex and gender diversity among transgender persons in Ontario, Canada: Results from a respondent-driven sampling survey. *Journal of Sex Research, 52,* 1–14.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling.* New York, NY: Academic Press.

Schwartz, J. E. (1977). An examination of CONCOR and related methods for blocking sociometric data. *Sociological Methodology, 7,* 255–282.

Scott, J. (1991). *Social network analysis: A handbook.* London, England: Sage.

Scott, J. (2017). *Social network analysis* (4th ed.). London, England: Sage.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks, 5,* 269–287.

Senay, E., & Landrigan, P. J. (2018). Assessment of environmental sustainability and corporate social responsibility reporting by large health care organizations. *JAMA Network Open, 1*(4), e180975.

Sentis, K. P., & Burnstein, E. (1979). Remembering schema-consistent information: Effects of a balance schema on recognition memory. *Journal of Personality and Social Psychology, 37,* 2200–2211.

Shakya, H. B., & Christakis, N. A. (2017). Association of Facebook use with compromised well-being: A longitudinal study. *American Journal of Epidemiology, 185,* 203–211.

Shakya, H. B., Christakis, N. A., & Fowler, J. H. (2017). An exploratory comparison of name generator content: Data from rural India. *Social Networks, 48,* 157–168.

Shipilov, A. V., Rowley, T. J., & Aharonson, B. S. (2006). When do networks matter? A study of tie formation and decay. *Advances in Strategic Management, 23,* 481–519.

Simmel, G. (1908). *Sociology: Investigations on the forms of sociation.* Leipzig, Germany: Duncker and Humblot.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13,* 108–131.

Slez, A., & Martin, J. L. (2007). Political action and party formation in the United States Constitutional Convention. *American Sociological Review, 72,* 42–67.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43,* 1–43.

Smith, J. A., & Moody, J. (2013). Structural effects of network sampling coverage I: Nodes missing at random. *Social Networks, 35,* 652–668.

Smith, J. A., Moody, J., & Morgan, J. H. (2017). Network sampling coverage II: The effect of non-random missing data on network measurement. *Social Networks, 48,* 78–99.

Snijders, T., van de Bunt, G., & Steglich, C. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks, 32,* 44–60.

Snijders, T. A. B., Steglich, C., & Schweinberger, M. (2017). Modeling the coevolution of networks and behavior. In *Longitudinal models in the behavioral and related sciences* (pp. 41–71). New York, NY: Routledge.

Sopranzetti, S. (2018). Overlapping free trade agreements and international trade: A network approach. *World Economy, 41,* 1549–1566.

Stadtfeld, C., Snijders, T. A. B., Steglich, C., & van Duijn, M. A. J. (2018). Statistical power in longitudinal network studies. *Sociological Methods & Research,* 0049124118769113.

Statista. (2018). Number of media users worldwide from 2010 to 2021 (in billions). Accessed August 3, 2018. www.statista.com/statistics/278414/number-of-worldwide-social-network-users.

Steffensmeier, D. J., Schwartz, J., & Roche, M. (2013). Gender and twenty-first-century corporate crime: Female involvement and the gender gap in Enron-era corporate frauds. *American Sociological Review, 78,* 448–476.

Stevens, E., Jason, L. A., Ram, D., & Light, J. (2015). Investigating social support and network relationships in substance use disorder recovery. *Substance Abuse, 36,* 396–399.

Stork, D., & Richards, W. D. (1992). Nonrespondents in communication network studies: Problems and possibilities. *Group & Organization Management, 17,* 193–210.

Stroud, N. J. (2017). Selective exposure theories. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford handbook of political communication.* Oxford, England: Oxford University Press.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top-of-the-head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249–288). New York, NY: Academic Press.

Tsai, W. (2001). Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance. *Academy of Management Journal, 44,* 996–1004.

Tsang, A., & Larson, K. (2016). The echo chamber: Strategic voting and homophily in social networks. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (pp. 368–375). New York, NY: Association for Computing Machinery.

Valenzuela, S., Park, N., & Kee, K. F. (2009). Is there social capital in a social network site? Facebook use and college students' life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication, 14,* 875–901.

Van den Elzen, S., Holten, D., Blaas, J., & van Wijk, J. J. (2016). Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE Transactions on Visualization and Computer Graphics, 22,* 1–10.

Van der Gaag, M., & Snijders, T. A. B. (2004). Proposals for the measurement of individual social capital. In H. Flap & B. Volker (Eds.), *Creation and returns of social capital: A new research program* (pp. 199–219). New York, NY: Routledge.

Van Tubergen, F. (2014). Size and socio-economic resources of core discussion networks in the Netherlands: Differences by national-origin group and immigrant generation. *Ethnic and Racial Studies, 37,* 1020–1042.

Vaquera, E., & Kao, G. (2008). Do you like me as much as I like you? Friendship reciprocity and its effects on school outcomes among adolescents. *Social Science Research, 37,* 55–72.

Varela, A. R., Pratt, M., Harris, J., Lecy, J., Salvo, D., Brownson, R. C., & Hallal, P. C. (2018). Mapping the historical development of physical activity and health research: A structured literature review and citation network analysis. *Preventive Medicine, 111,* 466–472.

Verhaeghe, P.-P., Van de Putte, B., & Roose, H. (2013). Reliability of position generator measures across different occupational lists: A parallel test experiment. *Field Methods, 25,* 238–261.

Walsh, F., & Pozdnoukhov, A. (2011). Spatial structure and dynamics of urban communities. In *Proceedings of the 2011 Workshop on Pervasive Urban Applications (PURBA)*. San Francisco, CA: 1–8.

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks, 34,* 396–409.

Wang, Y., & Bowers, A. (2016). Mapping the field of educational administration research: A journal citation network analysis. *Journal of Educational Administration, 54,* 242–269.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge, England: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika, 61,* 401–425.

Wellman, B. (1999). From little boxes to loosely bounded networks: The privatization and domestication of community. In J. L. Abu-Lughod (Ed.), *Sociology for the twenty-first century* (pp. 94–114). Chicago, IL: University of Chicago Press.

White, D. R., & Reitz, K. P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks, 5,* 193–234.

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social structure from multiple networks, I: Blockmodels of roles and positions. *American Journal of Sociology, 81,* 730–780.

Wolff, H.-G., Schneider-Rahm, C. I., & Forret, M. L. (2011). Adaptation of a German multidimensional networking scale into English. *European Journal of Psychological Assessment, 27,* 244–250.

Wu, M., & Knoke, D. (2017). Dark networks: The terror-crime nexus. In *Palgrave handbook of inter-organizational relations in world politics* (pp. 471–484). London, England: Palgrave Macmillan.

Xu, K. S., & Hero, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing, 8,* 552–562.

Yang, J., & Leskovec, J. (2014). Overlapping communities explain core-periphery organization of networks. *Proceedings of the IEEE, 102,* 1892–1902.

Yang, J., McAuley, J., & Leskovec, J. (2014). Detecting cohesive and 2-mode communities in directed and undirected networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining.* (pp. 323–332). New York, NY: Association for Computing Machinery.

Yang, S., & Hexmoor, H. (2004). Measuring optimal connections in large networks: A new algorithm and its applications. *Journal of Mathematical Sociology, 28,* 197–213.

Yang, S., Keller, F., & Zheng, L. (2016). *Social network analysis: Methods and examples.* Thousand Oaks, CA: Sage.

Yang, S., Limbocker, S., Dowdle, A., Stewart, P. A., & Sebold, K. (2015). Party cohesion in presidential races applying social network theory to the preprimary multiple donor networks of 2004 and 2008. *Party Politics, 21,* 638–648.

Yenigün, D., Ertan, G., & Siciliano, M. (2017). Omission and commission errors in network cognition and network estimation using ROC curve. *Social Networks, 50,* 26–34.

Young, F. W. (1987). *Multidimensional scaling: History, theory, and applications.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Zhang, B., & Al Hasan, M. (2017). Name disambiguation in anonymized graphs using network embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* Singapore, November 6–10, 2017.

Zhang, M., Kong, X., Zheng, J., Wan, J.-b., Wang, Y., Hu, Y., & Shao, R. (2016). Research and development of antibiotics: Insights from patents and citation network. *Expert Opinion on Therapeutic Patents, 26,* 617–627.

Zhang, X., Zhao, J., & Cao, G. (2015). Who will attend?—Predicting event attendance in event-based social network. In Mobile Data Management (MDM), *2015 16th IEEE International Conference on Mobile Data Management,* Pittsburgh, PA. June 15 (Vol. 1, pp. 74–83). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Zipkin, J. R., Schoenberg, F. P., Coronges, K., & Bertozzi, A. L. (2016). Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics, 27,* 502–529.

Žnidaršič, A., Ferligoj, A., & Doreian, P. (2012). Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels. *Social Networks, 34,* 438–450.

Žnidaršič, A, Ferligoj, A., & Doreian, P. (2017). Actor non-response in valued social networks: The impact of different non-response treatments on the stability of blockmodels. *Social Networks, 48,* 46–56.

INDEX