Analyzing Social Networks

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

2nd Edition

Analyzing Social Networks

Stephen P Borgatti
Martin G Everett
Jeffrey C Johnson

SAGE

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

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

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

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

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First edition published April 2013. Reprinted 2013, 2015, 2016 (twice), and 2017.
This second edition first published 2018.

Editor: Jai Seaman
Assistant editor: Aly Owen
Production editor: Tom Bedford
Copyeditor: Christine Bitten
Proofreader:
Indexer:
Marketing manager: Susheel Gokarakonda
Cover design: Shaun Mercier
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

Library of Congress Control Number: 2017941096

British Library Cataloguing in Publication data

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

ISBN 978-1-5264-0409-1
ISBN 978-1-5264-0410-7 (pbk)

This book is for Lin Freeman, our mentor. Much of what is said in this book comes originally from Lin, either through classes or informal conversations.

Contents

Acknowledgements

We would like to acknowledge the considerable help of Bill Stevenson (Boston College), who wrote the original draft of the data collection chapter, as well as the guidance provided by the questions posed to us over the years from many workshop participants. We would also like to thank Michael Zurek for his help on the glossary.

Preface

Welcome to the world of social network analysis. This book is intended as a general introduction to doing network research. The focus is on methodology, from research design and data collection to data analysis. Of course, since methodology and theory are deeply intertwined, this book is also about network theory. What the book is not is a survey of empirical research on social networks.

The book is also meant to be relatively non-technical. We try not to simplify to the point of being inaccurate, but, when forced to make a choice, we have opted for intelligibility and transmitting the spirit of an idea. In each case, however, we provide pointers to the appropriate technical literature so that the reader can get a fuller picture if desired.

Doing network analysis implies using network analysis software. A number of packages exist, including UCINET (Borgatti et al., 2002) and Pajek (Batagelj and Mrvar, 1998). As two of the authors of UCINET are authors of this book, we use UCINET for most of our examples. However, we do *not* intend this book to be a tutorial on UCINET. This means we focus on generic data analysis issues and, in general, do not give detailed UCINET-specific instructions. For those interested, however, the book's website (http://analyzingsocialnetworks.com) gives detailed information on how all the UCINET examples are done. The one exception to all of this is Chapter 5, which is much more UCINET-focused than the rest of the book.

One of the issues we faced in writing this book was how to keep it down to a reasonable size and maintain an understandable flow. We wanted to write a guide rather than an encyclopedia. As a result, we had to leave some things out. In general, our approach to this was to include only methods and concepts that are in demand and tend to be useful in a variety of settings. For example, although the *k*-plex (Seidman and Foster, 1978) is one of our favorite network concepts, we left it out of the chapter on subgroups because, in general, other approaches tend to be more practical. Similarly, in the chapter on centrality, we successfully resisted the temptation to present even a small fraction of all the measures that are available in the literature.

Throughout the book, we use empirical examples to illustrate the material. Because social networks are studied in a variety of traditional academic disciplines, we draw

our examples from a wide variety of fields, including anthropology, sociology, management and health care.

The book consists of 15 chapters that, in our minds at least, are logically grouped into four sections. The first section consists of an introduction (Chapter 1) and some mathematical foundations (Chapter 2). Chapter 1 lays out our perspective on the network research enterprise as a whole. It discusses the kinds of things we try to explain, along with the main approaches to explaining them. Chapter 2 reviews – in very simple terms – some of the basic concepts in graph theory and matrix algebra. A reader familiar with network analysis could skip these two chapters, but we think it advisable to familiarize yourself with our notation and terminology.

The next section has six chapters which are all about research methods. Chapter 3, on research design, is about the choices we make in setting up a study to investigate a given research question. Some of it applies to social science research in general, but much of it presents issues that are specific to social network analysis, such as the special challenges to respondent privacy. A key concept introduced here is the distinction between whole-network research designs and personal-network (aka egocentric) research designs. Chapter 4 discusses different options for the collection of network data, focusing specifically on survey methods for full network designs. Chapter 5 is about the data manipulations we often do to prepare network data for different analyses. Because it also discusses the importing and exporting of data, this chapter is more closely tied to UCINET than any other chapter. Chapter 6 is about fundamental exploratory multivariate techniques that are not specifically designed for social network analysis but are often used as part of the analysis process. Chapter 7 is about ways of visualizing network data in order to reveal patterns. Finally, Chapter 8 is about statistical techniques for testing hypotheses with network data. These are techniques specifically tailored for the special challenges of network data, such as non-independence of observations. The first part of the chapter is about using permutation-based versions of standard techniques such as correlation and regression. The second part is about exponential random graph and SIENA models. These techniques are not available in UCINET, and the statistical underpinnings of the models are far outside the scope of this book. However, we have included a brief introduction so that the reader is at least familiar in the broadest terms with these options and can then decide whether to explore them further.

The third section of the book is about the core concepts and measures of network analysis. Chapter 9 discusses measures at the whole-network level of analysis, such as the density of ties and the degree of clustering. Chapter 10 is about measures of node centrality, which can be seen as characterizing each node's position in a network. Chapter 11 is about definitions and methods of

detecting groups (sometimes called 'clusters' or 'communities') within a network. Chapter 12 discusses ways of conceptualizing and measuring structural similarities in how nodes are connected in the network.

The final section of the book consists of three cross-cutting chapters organized around different kinds of data. Chapter 13 is about methods of analyzing affiliation-type data, as when we have persons' memberships in groups. Chapter 14 provides a set of heuristics useful in processing large networks, such as ways of breaking down the problem into a series of smaller ones, or changing the problem to analyze ties among clusters of nodes. Finally, Chapter 15 is concerned with designing, collecting and analyzing ego network data. We note that there is no chapter devoted to longitudinal data, but examples of longitudinal analyses can be found in many of the chapters in the network concepts section.

With certain exceptions, the chapters do not depend heavily on each other, so the book does not need to be read sequentially. One reviewer has suggested beginning with Chapters 1 and 2 for an introduction to networks, then Chapters 3–5, 15, 14 and 13 on study design and implementation, Chapters 9–12 on social network concepts and measures, and finally Chapters 6–8 on analyzing network data.

The second edition contains some new sections and additional material as well as some changes to the text in order to clarify and improve the flow. In addition, we have added some problems and exercises to each chapter.

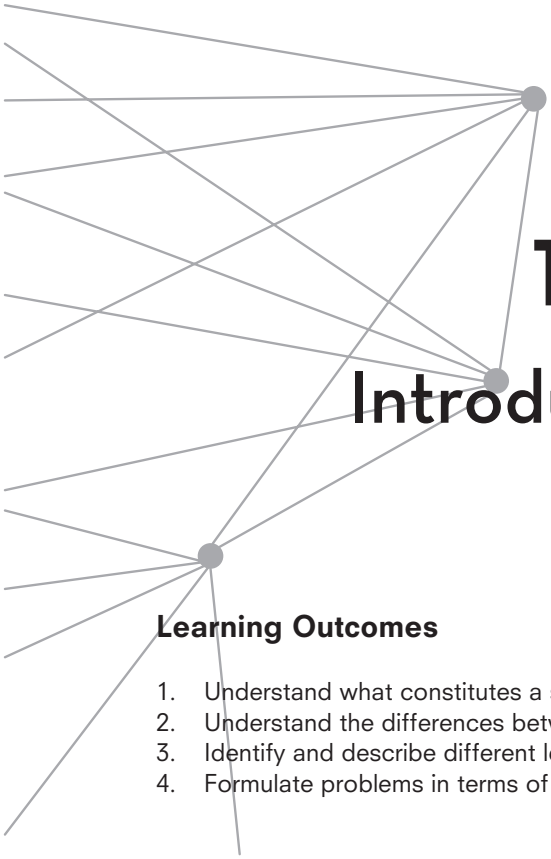
We are grateful to our mentor, Lin Freeman, for teaching us social network analysis in the first place. It is his take on the field that this book presents. We also thank Roberta Chase for many painful hours editing our less-than-perfect prose, Adam Jonas for managing all the (constantly changing) figures and tables, Chris Cooper for managing the references, and Filip Agneessens and Eric Quintane for their close reading of Chapter 8. We also thank Bill Stevenson for writing the original draft of the data collection chapter. Finally, we acknowledge NSF, DTRA, ARO and DARPA, whose grants have supported portions of this work.

We hope you find the book useful and will send us gently-worded suggestions for improvement.

Steve, Martin and Jeff
Dec 2017

Online Resources

[To come]



1 Introduction

Learning Outcomes

1. Understand what constitutes a social network
2. Understand the differences between different kinds of ties
3. Identify and describe different levels of analysis
4. Formulate problems in terms of network variables

1.1 Why networks?

An obvious question to ask is why anyone would want to analyze social network data. The incontestable answer, of course, is because they want to. But what are some sensible-sounding reasons that a researcher could use in polite company? One is that much of culture and nature seems to be structured as networks – from brains (e.g., neural networks) and organisms (e.g., circulatory systems) to organizations (e.g., who reports to whom), economies (e.g., who sells to whom) and ecologies (e.g., who eats whom). Furthermore, a generic hypothesis of network theory is that an actor's position in a network determines in part the constraints and opportunities that he or she will encounter, and therefore identifying that position is important for predicting actor outcomes such as performance, behavior or beliefs. Similarly, there is an analogous generic hypothesis at the group level stating that what happens to a group of actors is in part a function of the structure of connections among them. For example, a sports team may consist of a number of talented individuals, but they need to collaborate well to make full use of that talent.

1.2 What are networks?

Networks are a way of thinking about social systems that focus our attention on the relationships among the entities that make up the system, which we call actors or nodes. The nodes have characteristics – typically called 'attributes' – that distinguish among them, and these can be categorical traits, such as being male, or continuous attributes, such as being 56 years of age. The relationships between nodes also have characteristics, and in network analysis we think of these as kinds of ties or links. Thus, the relationships between Bill (male, 47 years old) and Jane (female, 43 years old) may be characterized by being married, living together, co-owners of a business, having friends in common, and a multitude of other relational characteristics that we refer to as ties. These relational characteristics can also be continuously or ordinally valued, as in having known each other for 12.5 years and having fights 3–5 times a year.

Of special interest in network analysis is the fact that ties interlink through common nodes (e.g., the A→B link shares a node in common with the B→C link), which creates chains or paths of nodes and links whose endpoints are now connected indirectly by the path. This in turn creates the connected web that we think of as a network.[1] Part of the power of the network concept is that it provides a mechanism – namely, indirect connection – by which disparate parts of a system may affect each other.

The nodes in a network can be almost anything, although when we talk about *social* networks we normally expect the nodes to be active agents rather than, say, inanimate objects.[2] Most often, nodes are individuals, such as individual persons or chimpanzees. But they can also be collectivities, such as teams, firms, cities, countries or whole species.

Whether actors are collectivities or individuals should not be confused with levels of analysis. In network analysis, it is useful to distinguish between three levels of analysis: the dyad, the node and the network (see Table 1.1). At the dyad level of analysis, we study pairwise relations between actors and ask research questions like 'do pairs of actors with business ties tend to develop affective ties?'. The dyad level is the fundamental unit of network data collection, and is the unit with the greatest frequency (i.e., most disaggregate). In Table 1.1, the notation $O(n^2)$ indicates that the number of dyads in a network is

[1] However, it should be understood that we do not require a network to be connected, nor to have any ties at all. This is important when analyzing networks over time, as initially a set of actors (say, a new task force charged with investigating unethical behavior in an organization) may have no ties at all to each other, but will develop ties over time. If the data are collected over time, we may see the network become connected.

[2] But this gets more complicated in the case of two-mode networks. See Chapter 13 for more on this.

Table 1.1 Examples of research questions by level of analysis and type of node.

	Type of node	
Level of analysis	**Individuals**	**Collectivities**
Dyad level $O(n^2)$	Are employees whose offices are near each other more likely to develop friendships than employees whose offices are further apart?	Are firms with similar organizational cultures more likely to form joint ventures with each other?
Node level $O(n^1)$	Are employees who are more central in their organization's friendship network less likely to leave for another company?	Are firms with more diverse technology partners more likely to introduce innovative products into the market?
Group/Network level $O(n^0)$	When a network of employees is characterized by many redundant paths between all pairs of persons, is the network less disrupted by individuals leaving the firm?	When a network of firms is densely connected, does this place the network at greater risk of catastrophic failure (because of cascade effects)?

of order n^2, where n is the number of nodes in the network.[3] At the node level of analysis, we ask questions like 'do actors with more friends tend to have stronger immune systems?'. Most node-level network properties are aggregations of dyad-level measurements, as when we count the number of ties that a node has. The number of nodes in the network is, of course, of order n.

At the group or network level, we ask questions like 'do well-connected networks tend to diffuse ideas faster?'. The number of objects at this level of analysis is of order n^0, which is to say, 1. This means, for example, that if we have a friendship network, a variable at this level of analysis will consist of a single quantity that characterizes the network as a whole (e.g., how densely connected it is). Note that at each level of analysis, the nodes could be individuals or collectivities, as shown in Table 1.1.

It is worth noting that the 'micro' versus 'macro' terminology used in many of the social sciences can refer to either the rows or the columns of Table 1.1. For instance, in the management literature, micro refers to studies in which the cases are persons and macro refers to studies in which the cases are firms. But in economics, it is more common to use micro to refer to the study of actor-level behavior (whether the actors are individuals or firms) and macro to refer to studies of the economy as a whole (i.e., the network level of analysis). Another source of confusion is the use of 'levels' in multilevel or mixed models in statistics. Here we might calculate the centrality of students within grade-level networks

[3] The use of this notation to represent levels of analysis is due to David Krackhardt (personal communication).

in order to predict future success, but use a multilevel regression model that takes into account characteristics of the students' school and school district. At the same time, people who study personal networks often regard ties or alters (level 1 cases) as nested within egos (level 2 cases).

1.3 Types of relations

Relations among actors can be of many different kinds, and each type gives rise to a corresponding network. So, if we measure friendship ties, we have a friendship network, and if we also measure kinship ties among the same people, we have both a friendship network and a kinship network. In the analysis we may choose to combine the networks in various ways, but fundamentally we have two networks. Perhaps the most commonly studied ties for persons are friendship ties, advice- or other support-giving, communication and, the most basic of all, simple acquaintanceship (who knows whom). Acquaintanceship is especially important in large networks, such as a firm of 160,000 employees or society as a whole. The latter is the basis for the famous Milgram (1967) small world or 'six degrees of separation' study. The process of how individuals become acquainted has been the subject of considerable research, including Newcomb's (1961) seminal book, *The Acquaintance Process*.

Table 1.2 provides a useful taxonomy of types of ties among persons. Inspired by Atkin's (1977) distinction between backcloth and traffic, the principal division in the table is between the relational states (on the left) and the relational events (on the right). Relational states refer to continuously present relationships between nodes, such as being someone's brother or friend. 'Continuously persistent' does not mean that the relationship will never end, but rather that, while it does exist, it exists continuously over that time. This contrasts with relational events, such as selling a house. Although the process may take months to execute, the concept of a sale is a discrete event. (Of course, we can always define a relational state based on a relational event simply by casting it in a timeless way. For example, if Bill sells a house to Jim, it is an event, but the relation 'has ever received a house from' is a state.) Events that recur can also be counted, as in the number of emails that X sent to Y last month. We often use recurring relational events as evidence of an underlying relational state, as in assuming that a frequent lunch partner is a friend. We may also regard recurring events as antecedents of relational states, so that if we frequently have lunch together (perhaps for work-related reasons), we may develop a friendship. It is difficult to develop friendships without any interactions at all.

Within relational events, the table distinguishes between interactions and flows. Interactions are behaviors with respect to others and often observable by third parties. Flows are the outcomes of interactions, and interactions form the

medium that enables things to flow. Flows may be intangibles, such as beliefs, attitudes, norms, and so on, that are passed from person to person. They can also consist of physical resources such as money or goods. In this book, we use flow in a relatively strict sense that doesn't include all types of causal chains. For example, if I tell you something that causes you to pick up a gun and shoot someone and then the police lock you up, we don't call that a flow. But if I tell you that grapefruit amplifies the effects of certain drugs, and you tell that to someone else who passes it on to someone else, we call it a flow. The difference is that in the second case it is the same state that is moving through the network. In the first case, it is something different in each person. But both cases involve a causal chain. Flows, then, are a special case of a more general category of causal cascades.

Within relational states, the table distinguishes between similarities, relational roles and relational cognition. Taking these in reverse order, relational cognition refers to thoughts and feelings that people have about each other. This includes acquaintance – who knows whom. Relational cognitions are essentially unobservable by other network members except as inferred from interactions. A highly consequential example of relational cognition is the trust relation, which can determine whether transactions will take place, and at what cost.

The relational roles category includes some of the most permanent of human relations, such as 'parent of' and 'sibling of'. Typically, the persons we have these relationships with are named or categorized by the relationship. Hence the person we have a friendship tie with is called a friend and is seen as enacting the friend role. When these relationships are asymmetric (such as 'mother of'), our culture typically provides us with named reciprocal roles. Hence we have parents and children, students and teachers, bosses and subordinates, and so on.

The similarities category refers to relational phenomena that are not quite social ties but can be treated as such methodologically, and which are often seen as both antecedents and consequences of social ties. For example, physical proximity (i.e., similarity in physical location) provides opportunities for face-to-face interactions.

Table 1.2 Taxonomy of types of relations.

Relational states								
Similarities			Relational roles		Relational cognition		Relational events	
Location	**Participation**	**Attribute**	**Kinship**	**Other role**	**Affective**	**Perceptual**	**Interactions**	**Flows**
Same spatial and temporal space	Same clubs, same events	Same gender, same attitude	Mother of, sibling of	Friend of, boss of, student of, competitor	Likes, hates	Knows, knows of, sees as happy	Sold to, talked to, helped, fought with	Information, beliefs, money

At the same time, certain social relations (e.g., romantic) often lead to radical increases in proximity (as in moving in together). Co-membership in groups (such as universities, gyms, teams, workplaces) provides many opportunities for interaction. Co-participation in events (such as attending the same conference or the same political rally) also provides opportunities for interaction. We can also define similarities in terms of attributes of nodes, such as gender and race. An enduring finding in social psychology is homophily – the tendency for people to like people who are similar to themselves on socially significant attributes.

One reason for pointing out the difference between relational states and relational events is that most of network analysis is built on relational states. For example, most centrality measures are best understood as generating predictions of the amount or timing of flow that is expected to arrive at a node as a function of its position in a network of relational states. The network is an observable system of roads. The centrality measures estimate the amount or timing of traffic that might flow to each node, given a set of assumptions about how things flow (e.g., whether they travel only along shortest paths). In most cases, if we were able to measure flow directly, we would not need to calculate centrality: we would simply use the observed flow instead.

It is worth pointing out that when nodes are collectivities, such as firms, there are two different kinds of ties possible. First, there are ties among the firms qua firms – that is, ties that are explicitly between the firms as single entities, such as a joint venture between two firms, an alliance, a purchase agreement, and so on. Second, there are ties between the individual members of the firms. Even though these are not 'official' ties between the organizations, they may serve all the same functions. For example, if the chief executive officers of two companies are friends, they may well share considerable information about each other's organization, constituting a flow of information between the firms. Table 1.3 provides examples of both kinds of ties among firms, cross-classified using the typology in Table 1.2.

Table 1.3 Relations among firms.

Type	Firms as entities	Via individuals
Similarities	Joint membership in trade association; co-located in Silicon Valley	CEO of organization A sits on same board as CEO of organization B
Relations	Joint ventures; alliances; distribution agreements; owns shares in; regards as competitor	Chief scientist of A is friends with chief scientist of B
Interactions	Sells product to; makes competitive move in response to	Representatives of A converse with representatives of B
Flows	Technology transfers; cash infusions	Employee of A leaks information to employee of B

1.4 Goals of analysis

Network analyses can be applied or basic.[4] By 'applied' we mean that the study consists of calculating a number of metrics to describe the structure of the network or capture aspects of individuals' positions in the network. The results are then interpreted and acted upon directly. For example, in an applied setting such as public health, we might use a centrality analysis of a network of drug addicts to detect good candidates for costly training in healthful practices, with the hope that these individuals would then diffuse the practices through the network. Or in management consulting, we might detect groups of employees from one organization in a merger situation who are not integrating well with the other company and create some kind of intervention with them. Applied studies are basically univariate in the sense that the variables measured are not correlated with each other. Rather, the correlations are assumed – because they have been observed or deduced in other, basic, research. For example, in the drug addict case, we choose to identify central players because previous research has suggested that getting central players to adopt a behavior will have add-on effects through diffusion to others. The causal relationships have been established, so we need only measure the predictor variables.

In contrast, basic research studies are multivariate and correlative – they try to describe the variance in certain variables as a function of others. The objective is to understand the dependent variables (i.e., outcomes) as the result of a causal process acting on a set of starting conditions. The independent variables serve to capture the initial conditions as well as traces of the theorized process. These are the kinds of studies we usually see in academic research. The function of network analysis in these studies is often to generate the variables that will be correlated, either as independent/explanatory variables or as dependent/outcome variables. As an example of the former, we might construct a measure of the centrality of each actor in a network, and use that to predict each actor's ability to get things done (i.e., their power). Studies of this type seek to create a network theory of ____, where we fill in the blank with the dependent variable, such as aggression or status attainment, yielding a 'network theory of aggression' or a 'network theory of status attainment'. As an example of using network variables as dependent variables, we might use the similarity of actors on attitudinal and behavioral variables (e.g., political views and smoking behavior) to predict who becomes friends with whom. Studies of this type seek to generate a ____ theory of networks, where we fill in the blank with a mechanism relating to the independent variables, such as a 'utility-maximization theory of network tie formation' or a 'balance theory perspective on network change'.

[4] Some might use 'descriptive' or 'explanatory', but explanation is theory and a theory is a description of how a system works.

Table 1.4 Types of network studies classified by direction of causality and level of analysis.

	Network variables as independent/ explanatory	**Network variables as dependent/ outcomes**
Dyad level	Friendship between pairs of farmers to predict which pairs of farmers make the same decision about going organic	Similarity of interests (e.g., sky diving) to predict who becomes friends with each other
Node level	Centrality in organizational trust network to predict who is chosen for promotion	Extraversion to predict who becomes central in friendship network
Network level	Shortness of paths in a group's communication network to predict group's ability to solve problems	Type of organizational culture (emphasizing either cooperation or competition) to predict structure of the trust network

Whether we use network variables as the independent variables in our analyses or as the dependent variables, they can be at any of the three levels of analysis discussed earlier. Table 1.4 gives examples of studies representing six possible combinations.[5]

1.5 Network variables as explanatory variables

When network variables are used as independent variables, the researcher is implicitly or explicitly using network theory to explain outcomes. These outcomes can be highly varied given that networks are studied in so many different fields – anything from individual weight gain to firm profitability. But because network processes are being used to explain these outcomes, there is a certain amount of unity in the logic that is used to predict the outcomes.

Most network theorizing is based on a view of ties as conduits through which things flow – material goods, ideas, instructions, diseases, and so on. Atkin (1977) referred to this as the backcloth and traffic model, where the backcloth is a medium, like a road system, that enables some kind of traffic to flow between locations. Within this basic conception, however, there are many different mechanisms that have been proposed to relate flows to outcomes. To discuss these, it is helpful to classify the outcomes being studied into a few broad categories. One basic category of outcomes consists of some sort of achievement, performance or benefit, either for individual nodes or for whole networks. Studies of this sort are known as social capital studies. An example is

[5] For simplicity, the table excludes cases where network variables are both the independent and dependent variables, as when friendship ties are used to predict business ties, or one kind of node centrality is used to predict another.

social resource theory (Lin, 2001), which argues that an actor's achievement is in part a function of the resources that their social ties enable them to access. Thus, an entrepreneur who is well connected to people who control a variety of important resources (e.g., money, power, knowledge) should be better positioned to succeed than one who has only her own resources to draw on. Thus, the key here is the inflow of resources that the entrepreneur's ties afford her.

Another perspective, which we refer to as *arbitrage* theory,[6] argues that a node B can benefit if it has ties to A and C, who are otherwise unconnected and who have achieved differing levels of progress toward a common goal. For example, if C has already solved something that A is still struggling with, B can make herself useful by bringing C's solution to A (for a price!). Here, the benefit is derived from the combination of an inflow and an outflow. Yet another network mechanism linking networks to achievement is *auctioning*. Here, if B has something that both A and C need, B can play them off against each other to bid up the price or extract favors from each. In this case, the benefit comes from the potential outflow from B to her contacts. In all of these cases, achievement is some sort of function of social ties. That is, the structure of the network and the position of individual nodes within it are crucial factors in predicting outcomes. This is very clear in the last two examples, in which a node B occupies a position between two others. But it is also true of the first case (social resource theory), because the resources of a node's connections may themselves be a function of their connections.

Another basic category of outcomes is what we might call 'style'. Unlike achievement, where one outcome is 'better' than another, style is about choices. Studies in this category look at things like political views, decisions to adopt an innovation, acquisition of practices and behaviors, and so on. These outcomes are often phrased in dyadic terms, so that what we are trying to explain is why, say, two firms have adopted similar internal structures, or why two people have made the same decision on the kind of smartphone to buy. The classic network explanation for these observed similarities is diffusion or influence. Through interactions, actors affect each other and come to hold similar views or become aware of similar bits of information. This is a perspective that clearly stems from a view of ties as conduits. But it is not necessarily the case that node A resembles node B because they influenced each other. It could be that a third party is tied to both of them and is influencing them both. It could also be something more subtle. For example, consider predicting employees' reaction to their phone ringing. Suppose some people cringe when it rings and others enjoy it. It could be that the people who cringe are those who are highly central in the advice network, meaning that lots of people are constantly calling to get their help and this gets annoying. Notice it is not that the central people are infecting each other with a bad attitude toward the phone, or even that third parties are

[6] Arbitrage is our term for one specific mechanism in Burt's (2004) discussion of brokerage.

infecting both of them with that attitude. It is a reaction that both have to the same situation, namely receiving so much flow. Essentially, the argument is that nodes are shaped by their social environments, hence nodes that have similar environments (such as both being central) will have similar outcomes.[7]

1.6 Network variables as outcome variables

It is often asserted that there is more research examining the consequences of network variables than the antecedents. This could be true, but it could also be a misperception due to the fact that the various factors that impinge on network variables come from a wide variety of different fields and will not have any particular theoretical unity. This is especially clear when you consider that the network properties being explained can be at different levels of analysis (i.e., the dyad, the node and the whole network), and that they may not be talked about using network terms. For example, there is a large and venerable literature on the acquaintance process (Newcomb, 1961) that never uses the term 'network'.

One of the oldest and most frequently replicated findings in social psychology is homophily – the tendency for people to have positive ties to those who are similar to themselves on socially significant attributes such as gender, race, religion, ethnicity and class. One way of thinking about these findings is in terms of a logistic regression in which the cases are dyads, the dependent variable is whether or not the nodes in the dyad have a positive tie, and the independent variables are things like *samegender* (a variable that is 1 if the nodes in the dyad are the same gender and 0 otherwise) and *agediff* (the absolute value of the difference between their ages).[8] In predicting most kinds of positive ties (but not marriage or other romantic relationships) we find a positive coefficient for *samegender* and a negative coefficient for *agediff*.

It is worth noting that having positive ties with people similar to oneself need not be solely the result of a preference. It could also reflect the availability of suitable partners. For example, if most people in an organization were women, we would expect most of these women's work friends to be women as well, simply because of availability. At the same time, we would expect most men to have quite a few women as friends, again because of availability. We would not want to conclude from such data that women are homophilous whereas men are heterophilous. One of the historical roots of social network analysis is in structuralist sociology, which, in the name of parsimony, urges us to seek answers in opportunities and constraints before turning to preferences.

---

[7] Note this is an example of a causal cascade that is not a flow, as discussed earlier.

[8] See Chapter 8 for a discussion of how to deal with issues of non-independence of observations that arise in an analysis of this type.

This suggests two basic types of factors in tie formation – opportunity and preference – and these are often intertwined. As an example of an opportunity-based mechanism, another well-known finding in the literature is that one tie leads to another. For example, business ties can lead to friendship ties, and vice versa. The presence of one tie sets up the opportunity for another kind of tie to form. More generally, as discussed in the third section of this chapter, we often expect relational states like friendship to lead to interactions (e.g., talking) through which things like information can flow, and which in turn can change the relationship (e.g., sharing intimacies deepens the relationship).

An example of a preference-based mechanism is balance theory (Festinger, 1957; Heider, 1958). In this theory, a person tries to be congruent with those she likes. So, if Jane likes Sally, and Sally likes Mary, it would cause Jane cognitive dissonance to dislike Mary. Based on balance theory, we would expect either that Jane's estimation of Mary would rise, or her estimation of Sally would decline. Note that an opportunity-based perspective would also predict the development of a positive tie between Jane and Mary because both of them are friends with Sally and Sally might well invite both to the same events, where they might interact and learn to like each other.

1.7 Summary

Network analysis is about structure and position. To this end, the field has developed an impressive array of concepts to characterize position and structure. In large part, the field has been able to express these concepts formally (i.e., in mathematical terms). This is a huge advantage because it means we can program computers to detect and measure these concepts in data, which in turn allows us to test hypotheses empirically. One downside, however, has been that some social scientists, unfamiliar with formal theorizing, have misconceived of the field as a methodology. It does indeed have a distinctive methodology that is born of its fundamentally relational view of social phenomena. But the theoretical concepts that are so emblematic of the field, such as centrality and structural equivalence, are just that: theoretical concepts that are part of a distinctive approach to explaining the social world (Borgatti and Halgin, 2011).

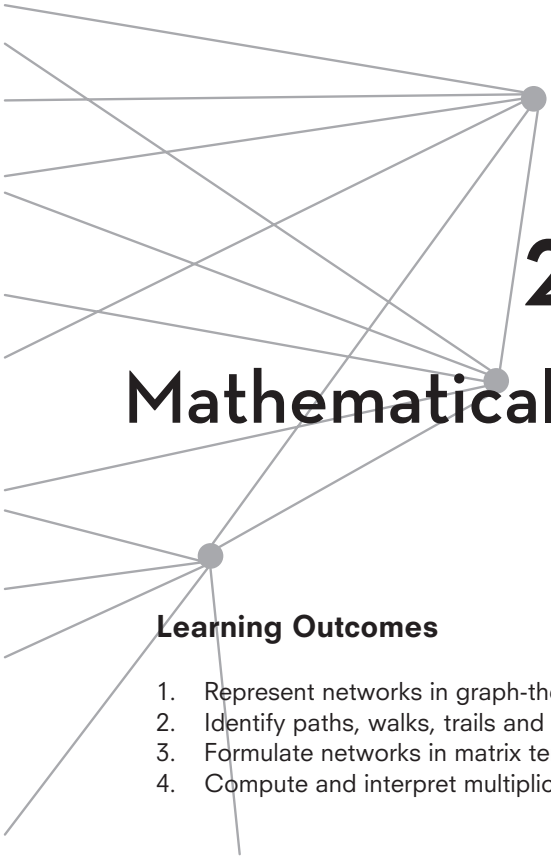
1.8 Problems and Exercises

1. There are three levels of analysis in the study of social networks: the dyadic level, node level and network level. For each of the research problems described below, what level of analysis is appropriate?

a. In a coeducational summer camp for teens, researchers want to know the extent to which attitudes about religion play a role in the formation of friendships within the first week of coming to camp.
b. An anthropologist is interested in studying the relationship between Canadian Inuit hunters' structural position in a hunting advice network, as measured by indegree centrality, and their hunting success.
c. A sports psychologist is interested in studying the relationship between basketball team cohesion off the court and number of regular season wins among a sample of 30 US universities.
d. A political scientist hypothesizes a relationship between the presence of international trade relations and the formation of bilateral defense agreements.
e. An agricultural extension researcher proposes that time of adoption of a new fertilizer among Iowa corn farmers is related to the structural centrality of farmers in a communication network.
f. An organizational sociologist hypothesizes that the more regional sales teams have a centralized information-sharing network the greater the team's overall sales.
g. An educational researcher is interested in how the political views of incoming freshmen at a large university affect the formation of friendship ties over the first semester.
h. A network researcher is interested in the relationship between astronaut knowledge of mission network structure and psychological well-being over the course of a 30-day simulated mission.
i. A management researcher advocates that highly centralized networks are more efficient at a variety of task settings than distributed networks, and designs an experiment to test this hypothesis.

2. For each of the research problems identified in Problem 1, which is the explanatory variable, and is it a network or non-network variable?
3. Based on the taxonomy of relations in Table 1.2, what type of relation best reflects each of the following? Explain your answer.

a. International trade
b. Financial transactions among banks
c. Preschool children's stated play preferences
d. College student attendance at university functions
e. Who one trusts in an organization
f. Advice-seeking among scientific research team members
g. Who one talks to about important matters
h. Money lending in a rural Indian community
i. Conflict among ethnic groups in South Sudan
j. Enjoys working with in small project teams
k. Would want to work with on future projects with others in a high-tech firm
l. Sexual relationships among IV drug users
m. Lab proximity of scientists in a research institute
n. Observed interactions at a company picnic
o. County commissioners and their votes on policy issues



2 Mathematical Foundations

Learning Outcomes

1. Represent networks in graph-theoretic language
2. Identify paths, walks, trails and components
3. Formulate networks in matrix terms
4. Compute and interpret multiplication of adjacency matrices

2.1 Introduction

As should be evident from Chapter 1, social network analysis is a social science. The actors we study are typically individuals (specifically humans, but also other social species such as apes and dolphins) or organizations (such as corporations). But networks are encountered in many other fields as well, including physics, ecology, chemistry, neurology, genetics and computer science. What these instances of network analysis have in common is an underpinning in a branch of discrete mathematics called graph theory. In this chapter we introduce the terminology and basic conceptual building blocks of graph theory. In addition, we present a short introduction to matrices, which can also be used to represent networks, and matrix algebra, which has proved very useful in network analysis.

2.2 Graphs

One way of conceptualizing networks mathematically is as *graphs*. The term 'graph' here does not refer to a diagram but rather a mathematical object (Harary, 1969). A graph $G(V, E)$ consists of a set of vertices V (also called nodes

or points), and a set of edges E (or links or lines). The edges connect pairs of vertices. To express that an edge connecting vertices u and v exists in a graph G, we write $(u, v) \in E(G)$. If we think of G as a binary relation, then we could also write uGv. For example, if G represents the 'likes' relation, the uGv would indicate that u likes v. When two vertices are joined by an edge, we say the vertices are adjacent. So, adjacent just means 'has a tie'. If an edge connects A with B, and another edge connects A with C, we say that the two edges are incident upon A. The number of edges incident on a node is called the 'degree' of that node.

Graphs may be directed or undirected. In a directed graph, the edges are like arrows – they have direction. Edges in directed graphs are often referred to as arcs, and can be thought of as ordered pairs of vertices. For example, the graph depicted visually in Figure 2.1 consists of a set of vertices $V = \{A, B, C, D, E\}$, and a set of ordered pairs $E = \{(A, B), (B, C), (C, D), (D, A), (D, E)\}$. The (C, D) pair indicates that C sends a tie to D. If the tie is reciprocated, the pair (D, C) would also be a member of the set E (but, in the example, it is not). Directed graphs are used to represent relational phenomena that logically have a sense of direction – for example, 'is the parent of' and 'gives advice to'. Note that directed relations can be reciprocated. It could be, for example, that in a certain group of people, every time someone gives advice to someone else, they receive advice from that person as well.

In undirected graphs, the edges are unordered pairs. Undirected graphs are used for relations where direction does not make sense or logically must always be reciprocated, as in 'was seen with' or 'is kin to'.

Although not a mathematical necessity, in social network analysis we usually organize things such that every edge in a graph means the same thing – that is, represents the same social relation. So a given graph $G(V, E)$ contains only friendship ties, while another graph $H(V, A)$ contains only advice ties among the

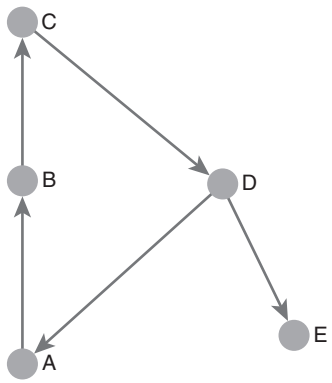


Figure 2.1 A simple directed graph.

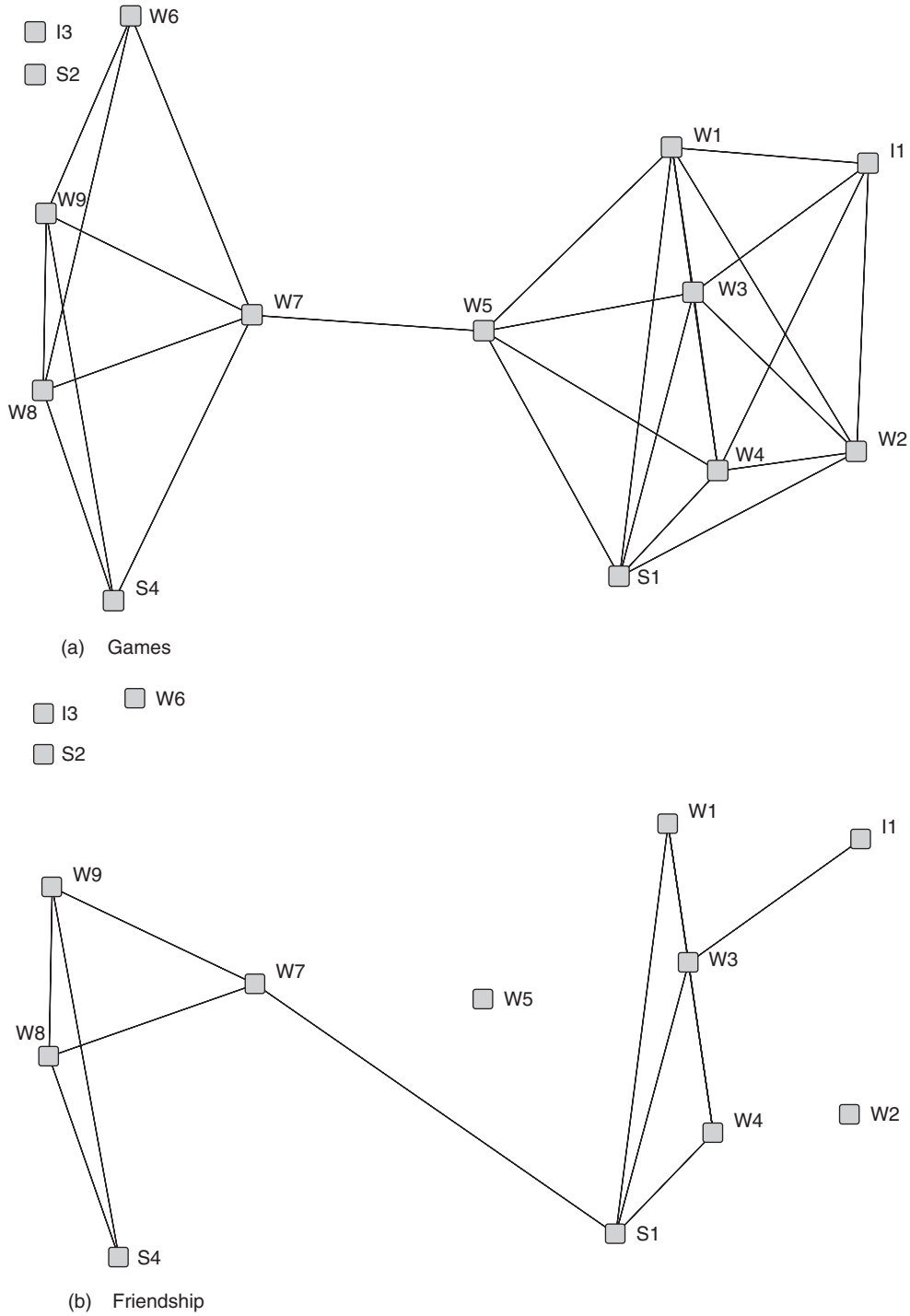


Figure 2.2 A multirelational network consisting of two relations: (a) who plays games with whom; (b) friendship ties.

same set of vertices. This means we think of the friendship and advice networks as different from each other and analyze them separately, though there are certainly exceptions to this.[1] In general, we expect each kind of social relation to have a different structure and to have different implications for the nodes involved. For example, being highly central in a friendship network might be very pleasant, while being central in a hatred network could be quite the opposite. Similarly, having many ties in a sexual network could imply a high risk of contracting a sexually transmitted disease, while having many ties in a gossip network implies an accumulation of (possibly incorrect!) information about one's social environment.

When we have more than one relation on the same set of vertices, we often refer to our data as a multirelational dataset, or (confusingly) as a network. Thus, the term 'network' in its largest graph-theoretic sense can refer to a collection of graphs in which each graph represents a different kind of social tie. As an example, Figure 2.2 shows some data from the Roethlisberger and Dickson bank wiring room dataset (Roethlisberger and Dickson, 1939). Figure 2.2a shows who plays games with whom, while Figure 2.2b shows friendship ties. As we can see, the two graphs have many points of similarity but are by no means identical. For example, in the games network, W5 and W7 are adjacent, whereas they are not in the friendship network. On the other hand, W3 and I1 are tied in both networks. When the relationship between two nodes includes multiple ties, the relationship is said to be 'multiplex'. We might even define a new network in which a tie exists between two nodes if their relationship is multiplex.

In the games network, there are two vertices that have no connections, I3 and S2. These are called 'isolates'. The friendship network has five isolates. We call the number of connections an actor has her 'degree'. Nodes with just one tie (i.e., degree 1) are called 'pendants'.[2] The friendship network has one pendant (I1).

2.3 Paths and components

A key concept in graph theory is the notion of a path. In the friendship network, vertices W1 and W7 do not have a tie, but information passed along between friends could reach W7 from W1 through the intermediary S1. A sequence of adjacent nodes forms a path. If the graph is directed, the sequence must respect

[1] For example, we might regard all ties between nodes as implying acquaintance, so we include all of them and call the resulting network the acquaintance network.

[2] Technically, a node is only a pendant if the one node it has a tie to has more than one tie.

the direction of the edges to be called a 'path'. Actually, the term 'path' refers to a particular kind of sequence, namely one which never revisits a vertex. For example, in the friendship graph, the sequence S4–W9–W8–W7 is a path, but W9–W8–S4–W9–W7 is not because it visits W9 twice. A sequence that revisits nodes but never revisits an edge is called a 'trail'. The sequence W9–W8–S4–W9–W7 is a trail, but W8–W7–W9–W8–W7–S1 is not because the line from W8 to W7 is used twice. Such a sequence is called a 'walk'. A walk is any sequence of adjacent nodes, without restriction. Obviously, every path is a trail, and every trail is a walk.

Paths, trails and walks matter because they correspond to different processes that we might want to model. Consider a coin changing hands as it moves through the economy. The coin does not know where it has been before, and neither do the people passing it along. As a result, the sequence it follows is completely unrestricted and is best described as a walk – perhaps even a random walk. In contrast, consider a juicy bit of gossip flowing through the network. Looking at the friendship graph in Figure 2.2b, does it seem likely that the gossip would follow the sequence W8–W7–W9–W8–W7–S1? Probably not. In most cases, W8 would remember having told W7 the story, and would not do it again anytime soon. Barring a few well-known exceptions (Alzheimer's cases; any family gathering) people do not tell the same stories again to the same people. So gossip probably does not traverse the network in an unrestricted way. One question, though, is whether it would revisit a node, as in the sequence W9–W8–S4–W9–W7. In many cases, the answer would be yes, because S4 does not know that W9 has already received this particular bit of information. Hence, an appropriate way of modeling the flow of gossip would be as trails. Finally, consider the case of a deadly virus that is spread by contact. In fact, let us suppose it is so virulent that it kills anyone it infects. To model the movement of this virus we would probably use paths because it never revisits a node. (Less gruesomely, we could imagine that it does not revisit nodes because once they get it, they become immune.)

The length of a walk (and therefore a trail and a path) is defined as the number of edges it has. The shortest path between two vertices is called a 'geodesic'. Geodesics are not necessarily unique as there could be multiple paths of equally short length between a given pair of vertices. In the friendship graph, W3–W4–S1 and W3–W1–S1 are both geodesics. The length of a geodesic path between two vertices is called the 'geodesic distance', or simply the 'distance'. If we assume that it takes a unit of time for something to traverse a link, the distance between two nodes indicates the fastest something could travel from one node to the other. A long geodesic distance implies that, even under the very best conditions, it would be a long time before something gets from one node to the other.

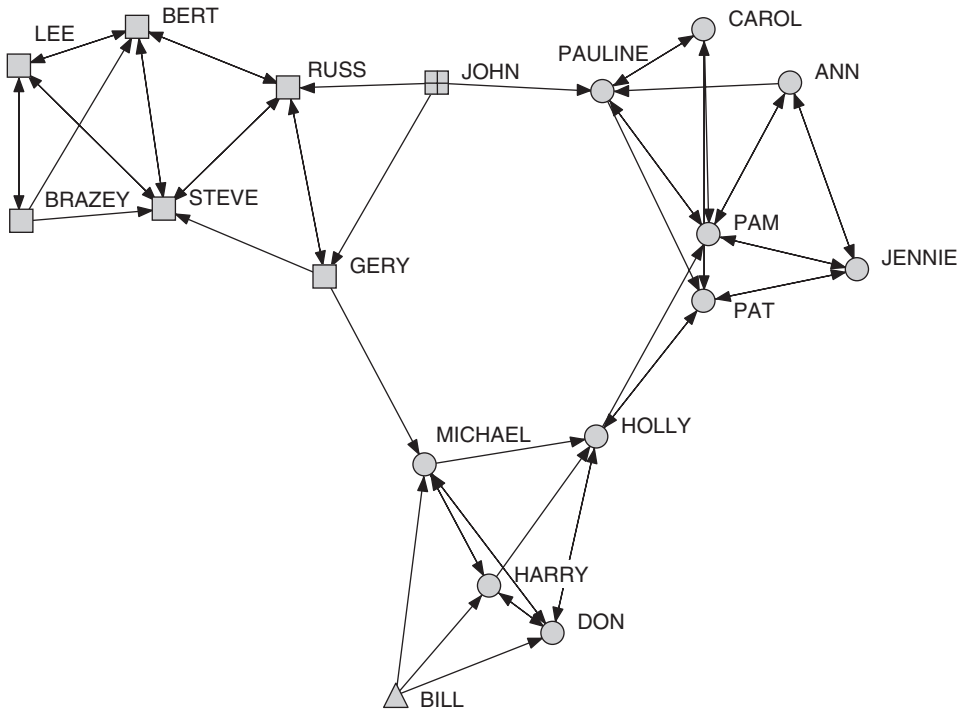


Figure 2.3 The Campnet dataset.

Some nodes cannot reach each other by any path. For example, consider the graph in Figure 2.3 (the standard UCINET Campnet dataset, which we describe in detail in Chapter 8). Try to find a path (respecting the direction of ties) from Holly to Brazey. There is no way to do it. The basic problem is that Michael and Pauline have no outgoing ties toward the left-hand side of the graph, and so there is no way for anyone on the right-hand side to reach anyone on the left-hand side. In this sense, the left- and right-hand sides belong to different components of the graph. A component is defined as a maximal set of nodes in which every node can reach every other by some path. The 'maximal' part means that if you can add a node to the set without violating the condition that everyone can reach everyone, you must do so. This means that the set {Lee, Bert, Brazey} is not a component, because we could add Steve and everyone could still reach each other. The component they are part of consists of Lee, Bert, Brazey, Steve, Russ and Gery. By this definition, there are four components in the graph: {Lee, Bert, Brazey, Steve, Russ, Gery}, {Michael, Harry, Don, Holly, Pam, Pat, Jennifer, Ann, Pauline, Carol}, {Bill}, and {John}. Bill and John form singleton components. These components are depicted by different node shapes in Figure 2.3.

In directed graphs like Figure 2.3, it is sometimes useful to consider 'weak components', which are the components you would find if you disregarded the directions of the edges. To distinguish weak components from the kind where we respect the direction of the edges, we can refer to the latter as 'strong components'. If the data are not directed, we just use the term 'components'.

The games network in Figure 2.2a has an interesting bowtie-like structure. It is worth noting the importance of the edge connecting W5 with W7. If this edge were not there, the group on the left would be separated from the group on the right. We call such edges 'bridges'. The friendship relation Figure 2.2b has two bridges: the one connecting S1 and W7 and the one connecting I1 with W3. Vertices with that same property are called 'cutpoints'. In the games relation W5 and W7 are cutpoints, and in the friendship relation S1, W7 and W3 are cutpoints. Note that I1 in the friendship relation is not a cutpoint since its removal does not separate any part of the network. In these examples the cutpoints are at the ends of bridges, but it is possible to have cutpoints that are not part of a bridge.

In many circumstances we have values associated with our edges. These may represent the strength of the tie, the frequency of interaction or even a probability. This applies to both directed and undirected network data. In our diagrams we can put the value on the edge, but for complex networks this is often not practical and we discuss other approaches in the chapter on visualization. Figure 2.4 gives a valued network where the values are 1, 2 or 3. If A sends a tie to B with a value of 2, the value is placed closer to the sending vertex. It can be seen that some ties are reciprocated but not always with the same value.

How socially close actors are to each other in a network is known as 'dyadic cohesion'. The simplest, most fundamental measure of dyadic cohesion is adjacency itself. If you and I have a tie (say, a trust tie) then we are more cohesive than if we did not have a tie. Of course, we have to be careful to think about what kind of relation is being measured. A graph in which every node has a 'hates' tie to every other node may be mathematically cohesive, but the sociological reality is that the network is maximally non-cohesive. If the data consist of valued ties (e.g., strengths or frequencies), so much the better, because then we have degrees of cohesion instead of simple presence or absence.

It is useful to note that some nodes that are not adjacent may still be indirectly related. All nodes that belong to the same component are far more cohesive than a pair of nodes that are in separate components. If a virus is spreading in one component, it will eventually reach every node in the component – but it cannot jump to another component. Naturally, if we are using the existence of a path from one node to another as a measure of cohesion, it is only a small stretch to consider counting the number of links in the shortest path between two nodes as an inverse measure of dyadic cohesion. However, one

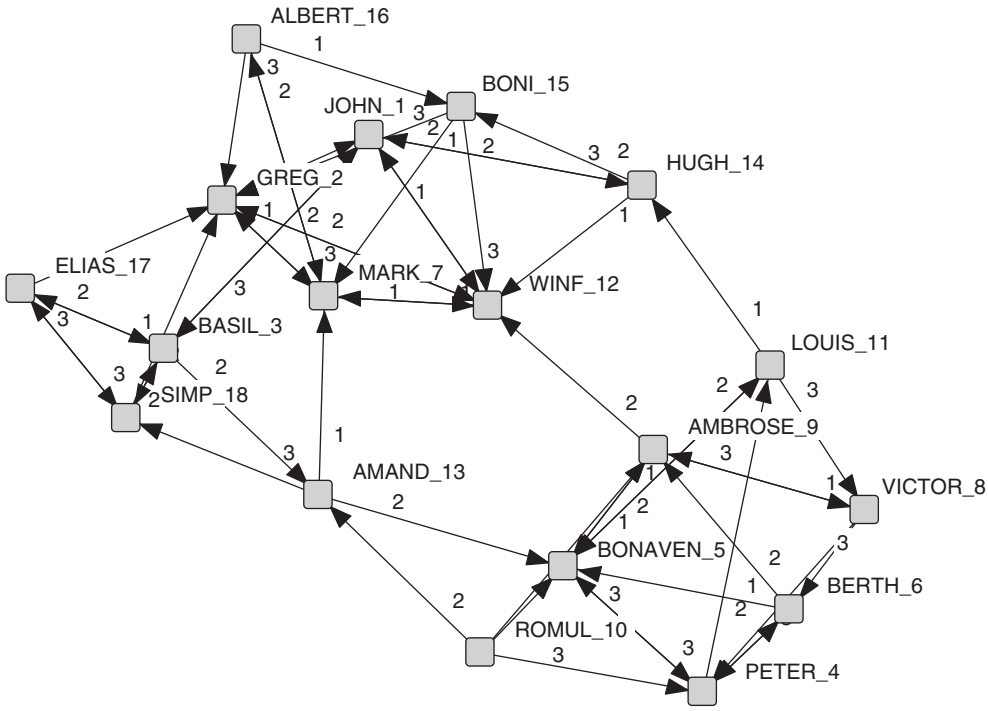


Figure 2.4 A valued network.

problem with geodesic distance is that the distance between nodes in separate components is technically undefined (or, popularly, infinite). A solution is to use the reciprocal of geodesic distance ($1/d_{ij}$) with the convention that if the distance is undefined, then the reciprocal is zero. This also has the advantage of making it so that larger values indicate more cohesion. We explore these ideas in more detail in Chapter 9.

2.4 Adjacency matrices

Another way to conceptualize networks mathematically is by using matrices. An adjacency matrix is a matrix in which the rows and columns represent nodes and an entry in row i and column j represents a tie from i to j. In other words, the adjacency matrix A of a non-valued graph is defined as a matrix in which $a_{ij} = 1$ if there is a tie from i to j, and $a_{ij} = 0$ otherwise. The direction is important and it must be remembered that, by convention, the direction goes from the rows to the columns. If the graph has valued edges, then we can simply use those values as the entries in the adjacency matrix. When the values are all

positive, we often use the convention that a zero indicates no tie (alternatively, we can use a specially designated missing-value code to indicate no tie; this is a necessity when the matrix can include negative values). If the graph is reflexive – that is, vertices can have ties to themselves – then there can be values down the main diagonal. For most relations self-loops are not allowed, and in this case the diagonal is often filled with zeros (a convention we shall use), but it could be argued that the diagonal should be blank. If the graph is undirected, then the matrix will be symmetric, meaning that the top right half of the matrix (above the main diagonal) will be the mirror image of the bottom half of the matrix, and x_{ij} will always equal x_{ji}. If the graph is directed, x_{ij} need not equal x_{ji} (although it may). The adjacency matrix for the games relation in Figure 2.2 is given in Matrix 2.1.

We can also use matrices to represent derived connections between pairs of nodes such as geodesic distance. Given a graph then the geodesic distance matrix D has elements d_{ij} equal to the geodesic distance between i and j. When there is no possibility of confusion, we simply use the term 'distance matrix'.[3] The matrix D is in fact extremely similar to the adjacency matrix of a graph; where there are 1s in the adjacency matrix, there are 1s in the distance matrix. But where there are 0s in the adjacency matrix, there is a range of values in the

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
		I1	I3	W1	W2	W3	W4	W5	W6	W7	W8	W9	S1	S2	S4
1	I1	0	0	1	1	1	1	0	0	0	0	0	0	0	0
2	I3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	W1	1	0	0	1	1	1	1	0	0	0	0	1	0	0
4	W2	1	0	1	0	1	1	0	0	0	0	0	1	0	0
5	W3	1	0	1	1	0	1	1	0	0	0	0	1	0	0
6	W4	1	0	1	1	1	0	1	0	0	0	0	1	0	0
7	W5	0	0	1	0	1	1	0	0	1	0	0	1	0	0
8	W6	0	0	0	0	0	0	0	0	1	1	1	0	0	0
9	W7	0	0	0	0	0	0	1	1	0	1	1	0	0	1
10	W8	0	0	0	0	0	0	0	1	1	0	1	0	0	1
11	W9	0	0	0	0	0	0	0	1	1	1	0	0	0	1
12	S1	0	0	1	1	1	1	1	0	0	0	0	0	0	0
13	S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	S4	0	0	0	0	0	0	0	0	1	1	1	0	0	0

Matrix 2.1 Adjacency matrix of relation 1 in Figure 2.2.

[3] More generally, elsewhere in the book we refer to 'proximity matrices', which is a general category of matrices – including distance matrices – that record the closeness or similarity (or farness or dissimilarity) of pairs of entities.

		1	2	3	4	5	6	7	8	9	1 0	1 1	1 2	1 3	1 4	1 5	1 6	1 7	1 8
		H	B	C	P	P	J	P	A	M	B	L	D	J	H	G	S	B	R
1	HOLLY	0		2	1	1	2	2	2	2			1		2				
2	BRAZEY	5	0	7	6	6	7	7	7	4		1	5		5	3	1	1	2
3	CAROL	2		0	1	1	2	1	2	4			3		4				
4	PAM	3		2	0	2	1	1	1	5			4		5				
5	PAT	1		1	2	0	1	2	2	3			2		3				
6	JENNIE	2		2	1	1	0	2	1	4			3		4				
7	PAULINE	2		1	1	1	2	0	2	4			3		4				
8	ANN	3		2	1	2	1	1	0	5			4		5				
9	MICHAEL	1		3	2	2	3	3	3	0			1		1				
10	BILL	2		4	3	3	4	4	4	1	0		1		1				
11	LEE	5	1	7	6	6	7	7	7	4		0	5		5	3	1	1	2
12	DON	1		3	2	2	3	3	3	1			0		1				
13	JOHN	3	4	2	2	2	3	1	3	2		3	3	0	3	1	2	2	1
14	HARRY	1		3	2	2	3	3	3	1			1		0				
15	GERY	2	3	4	3	3	4	4	4	1		2	2		2	0	1	2	1
16	STEVE	4	2	6	5	5	6	6	6	3		1	4		4	2	0	1	1
17	BERT	4	2	6	5	5	6	6	6	3		1	4		4	2	1	0	1
18	RUSS	3	3	5	4	4	5	5	5	2		2	3		3	1	1	1	0

Matrix 2.2 Geodesic distance matrix for Campnet data.

distance matrix, providing a more nuanced account of lack of adjacency. The distance matrix must be symmetric for undirected data. Matrix 2.2 gives the geodesic matrix for the network in Figure 2.3. This was produced by running the geodesic distance routine in UCINET. We see that the distance from Brazey to Pam is 6, and there is no path from Brazey to Bill as this entry is blank. Note that in this case the zeros on the diagonal do have meaning.

2.5 Ways and modes

The adjacency matrix of a graph is always square: it has the same number of rows as columns. Moreover, it is a one-mode matrix, which means that the rows and columns both refer to the same single set of entities. In contrast, in a two-mode matrix the rows and columns refer to different sets of (non-interchangeable) nodes, and would only coincidentally be square. For example, the classic dataset collected by Davis et al. (1941) in their book, *Deep South*, is shown in Matrix 2.3. The 18 rows of the matrix correspond to

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
EVELYN	1	1	1	1	1	1	0	1	1	0	0	0	0	0
LAURA	1	1	1	0	1	1	1	1	0	0	0	0	0	0
THERESA	0	1	1	1	1	1	1	1	1	0	0	0	0	0
BRENDA	1	0	1	1	1	1	1	1	0	0	0	0	0	0
CHARLOTTE	0	0	1	1	1	0	1	0	0	0	0	0	0	0
FRANCES	0	0	1	0	1	1	0	1	0	0	0	0	0	0
ELEANOR	0	0	0	0	1	1	1	1	0	0	0	0	0	0
PEARL	0	0	0	0	0	1	0	1	1	0	0	0	0	0
RUTH	0	0	0	0	1	0	1	1	1	0	0	0	0	0
VERNE	0	0	0	0	0	0	1	1	1	0	0	1	0	0
MYRNA	0	0	0	0	0	0	0	1	1	1	0	1	0	0
KATHERINE	0	0	0	0	0	0	0	1	1	1	0	1	1	1
SYLVIA	0	0	0	0	0	0	1	1	1	1	0	1	1	1
NORA	0	0	0	0	0	1	1	0	1	1	1	1	1	1
HELEN	0	0	0	0	0	0	1	1	0	1	1	1	0	0
DOROTHY	0	0	0	0	0	0	0	1	1	0	0	0	0	0
OLIVIA	0	0	0	0	0	0	0	0	1	0	1	0	0	0
FLORA	0	0	0	0	0	0	0	0	1	0	1	0	0	0

Matrix 2.3 Two-mode Southern women dataset.

women, and the 14 columns correspond to events the women attended. In the matrix, $x_{ij} = 1$ if woman i attended event j; this is sometimes called an 'affiliation matrix'.

More generally, matrices can be described as having ways and modes. The ways are the dimensions of the matrix – normally two, as when we have rows and columns – while the modes are the kinds of entities being represented. A three-way matrix has rows, columns and levels, as in a data cube. For example, suppose we have data indicating which persons attended which annual conferences in each year. This could be represented by a three-way, three-mode matrix. As an example of a three-way, one-mode matrix, consider the cognitive social structure data that David Krackhardt (1987) pioneered. He asked each member of a group to tell him which people in the group had friendship ties with which others. So, they were not just being asked about their own ties to others, but their perceptions of everyone else's ties to everyone else. The result is a person-by-person matrix for each person. Combining these into a single data matrix we get a three-way, one-mode matrix X in which $x_{ijk} = 1$ if person k perceives a tie from i to j. Note that x_{iji} is person i's perception of his or her own tie to j.

2.6 Matrix products

A cornerstone of matrix algebra is matrix multiplication, an operation that is defined as follows. If A and B are conformable matrices (which means that the number of columns in A equals the number of rows in B), then the product of A and B is written $C = AB$ and is calculated as follows:

$$C_{ij} = \sum_k a_{ik} b_{kj} \tag{2.1}$$

We can use matrix multiplication to construct compound social relations. For example, if F is an adjacency matrix representing the 'friend of' relation and matrix E represents the 'enemy of' relation, then the product FE is a compound relation we might call 'enemy of a friend of'. If the (i, j) cell of FE is greater than 0, this indicates that i has at least one friend for whom j is an enemy. In other words, j is the enemy of i's friend. More generally, $FE(i, j)$ gives the number of i's friends who have j as an enemy.

It is worth remembering that matrix multiplication is not commutative, so that AB need not be the same as BA (and, because of lack of conformability, may not even be calculable). For example, if F is the friendship relation and B is the 'boss of' relation, then if I have an FB relationship with Jane, she is the boss of at least one of my friends. But if I have a BF relationship with Jane, she is a friend of my boss. These are very different relationships.

We can also compute products of matrices with themselves. For example, if F is the friendship matrix, then FF is the 'friend of friend' relation. When the (i, j) cell of FF is greater than 0, it indicates that i has at least one friend who considers j a friend. The magnitude of the (i, j) cell gives the number of times that i has a friend that has j as a friend, which is to say, it is the number of friends of i who have j as a friend. Another way to think of this is in terms of walks. The FF matrix, or F^2, gives the number of walks from i to j that are of length 2 – that is, walks with one intermediary. Multiplying the FF matrix again by F gives us F^3, whose entries give the number of walks of length 3 from any node to any other.

More generally, the matrix F^k gives the number of walks of length k that start at the row node and end at the column node. It is worth remembering that these are walks rather than simple paths, which means they can double back on themselves. For example, suppose $F^3(i, j) = 2$. This means that there are two walks of length 3 that start at i and end at j. One such walk might be i–k–i–j. Another walk might be i–k–m–j. Both are walks of length 3. An important application of matrix powers is given in Chapter 10, where we discuss Bonacich's (1987) beta centrality concept.

A useful application of matrix products is to express social theories in formal form. For example, the notion that 'the friend of my friend is my friend, the friend of my enemy is my enemy, the enemy of my friend is my enemy, and the enemy of my enemy is my friend' can be expressed compactly as four equations:

$$F = FF$$

$$E = FE$$

$$E = EF$$

$$F = EE$$

Of course, once these principles are expressed as equations, they can also be tested. Using methods covered in Chapter 8, we can count how often (i.e., for how many i, j pairs) it is true that, say, $E = EF$. For example, we could count how often $EF(i, j) = E(i, j)$ across all i and j, and test whether this quantity is larger than we would expect by chance.

2.7 Summary

Social networks can be represented mathematically as graphs – mathematical objects consisting of two sets: a set of vertices and a set of edges connecting the vertices. The edges may or may not have a direction associated with them and can also have values. Each graph represents a single relation, but we often have multirelational networks with different sets of edges from different relations on the same set of vertices. A path is a sequence of non-repeated vertices such that adjacent pairs form an edge. Sets of vertices that are mutually reachable form components, and if they take account of direction, they are known as strong components. The length of the shortest path between any two vertices is called the geodesic distance. An alternative representation is to use matrices, with the adjacency matrix being by far the most common way to do this. We can also use matrices to capture the distance between all pairs of vertices in a network. If the data have two modes, we can use a reduced form of the adjacency matrix called an affiliation matrix.

2.8 Problems and Exercises

1. For the types of relations listed in Problem 3 in Chapter 1, are the ties implied by these relations directed or undirected?
2. Re-express the simple graph below as an adjacency matrix.

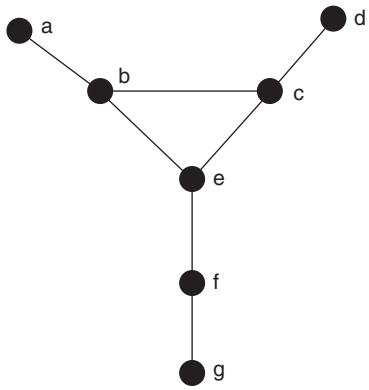


Figure 2.5

3. Re-express the graph in Problem 2 as a geodesic distance matrix. What do those distances mean?
4. For the graph in Problem 2 above, provide examples of each of the following:
   a. Paths
   b. Trails
   c. Walks
5. Re-express the directed valued graph below as a valued proximity matrix.

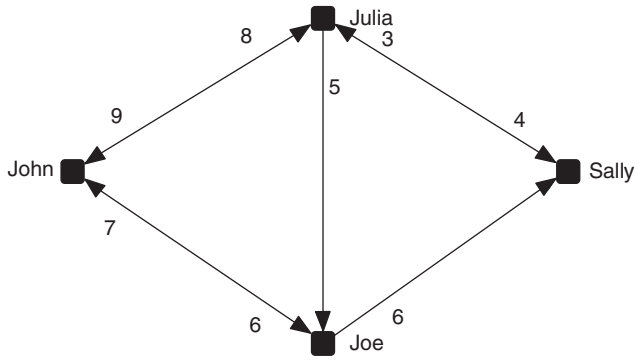


Figure 2.6

6. For the bank wiring room games graph in Figure 2.2a, if the edge between nodes W7 and W5 were removed, how many components would the graph now have? If we were to calculate geodesic distances for the new graph, what would be the distance between W9 and W3? Explain.
7. For each of the network examples in Chapter 1, Problem 3, are the associated matrices one-mode or two-mode?
8. Given a friendship relation and 'is the boss of' relation, use matrix multiplication to hand-calculate the 'is friends with the boss of' relation.

Friend of

	A	B	C	D
A	0	1	1	0
B	1	0	0	0
C	1	0	0	1
D	0	0	1	0

X

Boss of

	A	B	C	D
A	0	1	0	0
B	0	0	1	0
C	0	0	0	0
D	1	0	0	0

=

Friend of boss of

	A	B	C	D
A				
B				
C				
D				

Matrix 2.4

Recall that if cell (A, C) has a value greater than zero in the 'is friends with the boss of' matrix, it means that A is friends with C's boss. From a power perspective, how would you view the row sums of the 'Friend of boss of' matrix?

3
Research Design

Learning Outcomes

1. Design effective and reliable network research projects
2. Identify sources and boundaries of network data
3. Understand and minimize the effects of data error

3.1 Introduction

This chapter is about designing network research. We try to lay out some of the issues that need to be considered when constructing a network study. In particular, we try to highlight the implications in terms of the interpretation, validity and feasibility of different combinations of design choices. The reader will recognize that many of the issues discussed here are common to all social science research and are not particular to social network analysis. For example, social networks can be studied via experiments, quasi-experiments, field studies and so on. They can be studied not only quantitatively but also ethnographically. The data collection can be cross-sectional or longitudinal. In addition, however, we also discuss some special design issues specific to social network analysis, such as the decision between whole-network and personal-network designs, as well as how to bound the network, what kinds of ties to measure, and so on.

It should be noted that this chapter touches on several topics that are discussed in greater depth in other chapters. In particular, the reader should consult Chapters 1, 4, 8 and 15.

3.2 Experiments and field studies

While most social network research has been carried out using field studies, typically survey-based, there is a well-known body of network research employing experimental designs of one form or another. True experiments are the gold standard for the study of causation. For a study to be a true experiment it must have a pre-post (i.e., variables measure before some intervention or treatment, and then again after) or post-only design, together with random assignment of units to treatment groups and manipulation of the independent variable (such as a social intervention) while controlling for all other factors, both known and unknown. The key element of a true experiment is the random assignment; in quasi-experimental research there is usually a pre-post design, along with manipulation of the independent variable, but the units of analysis are not randomly assigned to treatments (usually because it is technically not feasible or is unethical, such as assigning smoking to randomly chosen participants). Field studies, or observational studies, may be longitudinal but more often are cross-sectional and do not involve a manipulation of the independent variable. As one goes from true experiments to field designs, there is less control over various threats to the studies' reliability and validity.

Table 3.1 provides some examples of experiments and quasi-experiments in social network research. As we move from the Rand et al. (2011) study down to that of Soyez et al. (2006), researcher control over the various aspects of the study declines. In the Rand et al. (2011) experiment, subjects are randomly assigned to one of four conditions. In each of the conditions the links in the who-can-communicate-with-whom network are manipulated. The study objective is to examine the evolution of cooperation in each of the experimental conditions. In the Barr et al. (2009) field experiment, subjects were randomly chosen from Orma villages in Kenya and small industrial operations in Ghana and were assigned to one of two player conditions. The behavior of subjects in the two reflected either degrees of trust or trustworthiness on the part of the players. Although there was random assignment of subjects, there was no direct manipulation of one of the primary independent variables of interest, individual-level social capital. Instead, social capital was measured separately and used as one of several independent variables in attempts to account for individual-level variation in subjects' game-playing behaviors. Finally, the Soyez et al. (2006) study used a classic quasi-experimental design in which subjects from four different clinics or cohorts were assigned in sequence, and not randomly, to one of two experimental conditions. Subjects in the control condition received standard treatment for drug abuse while subjects in the experimental condition received standard treatment plus a social network intervention.

Table 3.1 Experimental and quasi-experimental designs in social network research.

Study	Design	Conditions and manipulation	Independent and dependent variables
Rand et al. (2011)	Experiment. Random assignment of 785 subjects to one of four conditions. A repeated cooperative dilemma game is played with other subjects in an artificial network.	1 Random link condition 2 Fixed link condition 3 Strategic link condition 3a Viscous condition 3b Fluid condition	The independent variable is the experimental network condition, and the dependent variable is the evolution of cooperation in an iterative game.
Barr et al. (2009)	Experiment. Assign players of the Trust Game to two player conditions, one reflecting trust and the other trustworthiness. Players were randomly selected from the community to play.	Two cultural settings (Orma of Kenya and workers in Accra, Ghana). Two player conditions in the Investment Game protocol.	The independent variable was social capital (betweenness centrality) and the dependent variables were player 1 and player 2 behaviors reflecting degrees of trust (player 1 behavior) and trustworthiness (player 2 behavior).
Soyez et al. (2006)	Quasi-experiment. Members of four cohorts were assigned to social network intervention involving three elements. Assignment to conditions was based on sequential program admission, not random assignment.	The first group was taken from the four cohorts admitted between 1 May 2000 and 30 April 2002 (N = 94) and received standard treatment (control group). The second group was taken from the four cohorts admitted between 1 January 2001 and 30 April 2002 and received standard treatment plus network intervention (experimental group).	Test how treatment factors predicted substance abuse treatment retention, where one of the treatment factors was social networks intervention (ego network).

Most research involving social networks employs field/observational designs of one form or another. Data can be collected at a single point in time (cross-sectional and lagged cross-sectional) or at multiple points in time (longitudinal). Collecting data at two or more points in time allows for the study of change. Examples of cross-sectional and longitudinal research in social networks are shown in Table 3.2. These examples were chosen to help illustrate not only basic elements of research design but also studies that examine both the causes and the consequences of social network structure. The study by Christakis and Fowler (2007) analyzed data repurposed from the Framingham Heart Study, which had a prospective (longitudinal) design. Christakis and Fowler used it to study the 'contagion' of obesity through social networks in a population.

Table 3.2 Examples of field studies in social network research.

Study	Design	Independent variable	Dependent variable
Christakis and Fowler's (2007) study of networks and obesity (Framingham data)	Prospective design; follow a cohort through time	Ties to obese actors at time t_i	Weight at time t_{i+1}
Burt (1995) structural holes	Cross-sectional design	Individual level social capital (e.g., constraint)	Performance evaluations; bonuses received
Casciaro (1998) study of personality and network accuracy	Cross-sectional design	Strong need for achievement	Social network accuracy
Johnson et al.'s (2003) study of group dynamics in polar research stations	Prospective, repeated measures design	Core–periphery structure	Morale and individual level psychological well-being
Padgett and Ansell's (1993) study of marriage among elite Renaissance Florentine families	Retrospective design; data gathered from historical records	Spanning structural holes in marriage relations	Financial gain and power

The advantage of this design is that network relations at one point in time can be used to predict outcomes such as obesity at some future time, providing some help in sorting out the direction of causation (even though it did not use a true experimental design, something that would be totally impractical in this case). They found evidence to suggest a contagion effect even for something that would appear on the surface not to be 'catchable' in the medical sense.

The cross-sectional studies by Burt (1995) and Casciaro (1998) both collected data via surveys at one point in time. However, Burt was interested in understanding how an actor's position in a network – the spanning of structural holes – influences outcomes such as evaluations of employee performance and sizes of bonuses received. Thus, some element of network structure is influencing some outcome of interest (e.g., performance). In contrast, Casciaro (1998), although also using a cross-sectional design, was interested in how personality influences an actor's accuracy in cognitive social structures. So here an attribute of an actor is influencing the ability of that actor to report accurately on the network relations of others. In other words, network accuracy is a consequence of personality.

In the Johnson et al. (2003) study a longitudinal repeated measures design was used. Although the table depicts a study about the role of core–periphery structure in influencing group morale, the premise of the research was more complex and illustrates a slight spin on the simple structural causes-and-consequences dichotomy.

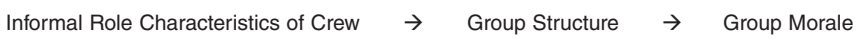


Figure 3.1 Relationship among variables for Johnson et al. (2003) study.

As shown in Figure 3.1, the study was more broadly focused on the relationship among three variables in which network structure was a mediating variable. The study focused on how the emergence of informal roles in the network (e.g., clown, expressive leader) affected the evolution of network structure. If certain roles emerged in certain combinations, it was expected that the network would form a core–periphery structure, and the more the network evolved a core–periphery structure, the higher the morale and individual-level psychological well-being (e.g., lower levels of depression). So, in a sense, this is an example of research that viewed structure as both a cause and a consequence.

Finally, the Padgett and Ansell (1993) study is not unlike the Burt example in that network structure is found to influence power among elite families in Renaissance Florence; here the network consists of connecting families by marriage that are not otherwise connected. This is an example of a retrospective case study where we are given an outcome – the rise of the Medici family in terms of power and wealth – and use historical data to speculate about why it happened.

3.3 Whole-network and personal-network research designs

There are two fundamental kinds of network research designs: 'whole-network' designs and 'personal-network' designs.[1] In general, when people talk about network analysis, they are referring to whole-network studies. In whole-network research, we study the set of ties among all pairs of nodes in a given set. For example, we might study who is friends with whom among all members of a given department in an organization. In whole-network studies, we can think of the relation being measured as a dyadic variable that has a value for every pair of nodes. For example, in the friendship case, every dyad might be assigned a 1 or a 0 indicating whether they are friends or not.

In personal-network studies, there is a set of focal nodes called 'egos' or 'index nodes', and their ties to others, called 'alters', are assessed, but the alters are not necessarily among the set of egos. An example of a personal-network

[1] In the literature, what we have called 'whole' network studies are known by a wide variety of names, such as 'socio-centric', 'complete' and 'full'. Studies using a personal-network research design are also known as 'ego-network studies' and 'ego-centered or egocentric studies'.

study is the General Social Survey of 1985, in which approximately 1500 egos were sampled using a probability sample from the population of Americans. Each was then asked for a list of up to five people with whom they discussed important matters. The aim was simply to understand something about the social environment of each of the egos, not to construct a network of ties among the 1500 (which would probably be completely empty), nor to connect the alters of one ego to the alters of any other ego (typically, the names of the alters are not even given in full).

In general, whole-network designs enable researchers to employ the full set of network concepts and techniques, which often assume that the entire network is available. This is particularly true of positional concepts such as betweenness centrality or regular equivalence. However, because the cost (to the researcher and the respondent) of whole-network designs increases quickly with network size, the richness of the data often suffers as the researcher has to scale back the number of questions he asks (see Chapter 4 for more information on this). In that sense, personal-network designs can yield richer, more detailed data about the network area local to the respondent, but at the cost of losing information on the global pattern of connections. Personal-network designs also have the advantage of simplifying issues of bounding the network, since there is no cost to allowing a respondent to mention any alter they like. Personal-network designs also have significant advantages with respect to confidentiality, as personal-network surveys can be entirely anonymous (with respect to the respondents), and when the respondents mention alters, they do not need to give the alters' real names. This can improve the quality of the data (because the respondent feels safer in giving these) and simplify the process of getting approval from human subject review boards.[2]

As we devote a separate chapter to personal-network designs, the rest of this chapter focuses on whole-network designs, although many of the points we make apply to personal-network designs as well.

3.4 Sources of network data

Network data can be collected from either primary or secondary sources. In primary data collection, we directly ask people questions or observe their behavior. What is asked or observed is determined by the objectives of the study, and the researcher has a lot of control over the types of relations studied and the types of actor attributes collected. In secondary data collection, we gather data that already exist somewhere, whether in paper records (e.g., fish exchange

[2] Institutional review boards or IRBs in the US.

records, historical marriage records), or electronic databases (e.g., Enron emails, social networking sites). Secondary data are often easier and quicker to collect but impose strong and arbitrary limits on the type of relations studied. Some of the computer-based data generated by social media such as Facebook and even email represent a transitional form between primary and secondary data. Although the data are collected directly, as in primary research, there are limitations on the types of relations available for study, as in secondary research.

Most published network research in the social sciences is based on primary data sources. Much of this is based on surveys, in which respondents are asked to report on their ties with others. However, there are also some well-known examples of direct observation. One of the stages of the well-known Hawthorne studies (Roethlisberger and Dickson, 1939) involved planting an observer at the back of the room where a set of employees constructed telephone wiring apparatuses. The observer was there for several months and recorded all kinds of interactions among the workers, including who played games with whom during breaks, who had conflicts with whom, who traded jobs with whom, and so on.

In recent years, we have seen a significant increase in the use of secondary sources. One reason for this is the increased availability of electronic records of all kinds, including bibliometric data (e.g., who cites whom), membership data (e.g., who is on what corporate board, who was in what movie) and of course social media (e.g., who follows whom on Twitter). Another reason is the increasing importance in the social science literature of longitudinal data, which are often only feasible to collect from secondary sources. However, not all secondary research is electronic. As previously mentioned, one of the best-known network analyses of archival data is the study by Padgett and Ansell (1993), who analyzed the pattern of marriages among Renaissance Florentine families.

3.5 Types of nodes and types of ties

As noted in Chapter 1 (see Table 1.2), there are many kinds of ties one could measure. Most network studies involve persons as the nodes and interpersonal relations as the ties. However, the nodes can be all kinds of entities – monkeys, firms, countries and so on. And the type of node obviously has a major impact on what kinds of ties are collected and how they are collected. These decisions – who to study, what ties to study, and where to obtain the data – are interlinked and must to some extent be considered together.

Table 3.3 reproduces in simplified form the typology of types of ties presented in Chapter 1. At the top left of the table are co-occurrences. One advantage of co-occurrence data is that they are relatively easy to collect. One reason is that membership-type data are often not thought of as particularly private or sensitive.

In addition, they are often available via archival sources. For example, we can look up the names of people serving on the boards of directors of firms. We can use the Internet Movie Database (IMDb) to find people who have served as cast or crew together on films. A frequently reanalyzed dataset in the network literature was collected by Davis et al. (1941) for their *Deep South* book. To obtain these data, they used the society pages of a local newspaper to learn which women attended which social events.

Next, we have true social relations – ties that have a continuous nature in the sense that they can be seen as relational states (such as being friends) rather than events (such as 'having sent an email to'). Many social relations have a quality of being institutionalized such that they have a degree of reality apart from the perceptions of the individuals involved. An example is marriage, where two people are married to one another even if they deny it. As such, information on such ties can be collected from sources other than the two people involved, such as others in the community, family members, archival records, and so on. Other types of social relations, such as affective and perceptual ties, have no independent existence or corroboration: short of inferring the tie based on some behavioral theory, such data have to be obtained by surveying the perceiver.

The third type of dyadic phenomenon, interactions, can be either directly observed or reported on by respondents. Who people talk to, watch movies with, hang out with, or communicate with via ham radio are all interactions, and are all, in principle, observable. Of course, there are always issues of interpretation. For example, if two people are verbally sparring, are they having a conflict or a friendly competition?

Table 3.3 Types of dyadic phenomena commonly studied.

Category	Varieties and examples
Co-occurrences	Co-membership in groups
	Co-participation in events
	Physical distances
	Similarities in attributes (e.g., political views)
Social relations	Kinship relations
	Affective relations (e.g., dislikes)
	Perceptual relations (e.g., knows)
Interactions	Transactions (e.g., 'sells to')
	Activities (e.g., 'sleeps with')
Flows	Ideas and information
	Goods
	Infections

In a network study of a fish camp, Johnson and Miller (1983) observed two Italian fishers engaged in what appeared to be a heated discussion. Johnson asked a younger Italian fisher, who was also observing, what the conflict was all about. The younger Italian explained that there was no conflict, but that the two men – who were brothers – were simply having a friendly discussion about a nephew. Johnson was interpreting that interaction from his cultural perspective rather than from the perspective of the two Italians engaged in the interaction.

It is worth noting here the difference between using a highly interpreted label such as 'friendly competition', versus a less interpreted label, such as 'verbal sparring', versus something even less interpreted, such as 'communicated face to face'. The higher the level of interpretation, the more theoretically useful the data are likely to be, but the greater the chance of being wrong. It is also worth noting that interactions are often collected as a proxy for unseen underlying social relations. For example, we might record who talks to whom outside of work and assume this means they are friends. Again, making these kinds of interpretations is often more powerful but may be quite unwarranted.

Electronic interactions are often available in archival form, as when we mine the email logs of a company's email server. Although convenient to collect, email interactions are particularly difficult to interpret with respect to inferring an underlying social relation. People email their friends, but they also email work colleagues, family members, acquaintances, and strangers, even on a corporate email account. Even when we have access to the content of the emails, it may be very difficult to determine what the underlying relationships are between the interactants.

Finally, the fourth type of dyadic phenomena, flows, can be seen as the outcomes of interactions. When two people interact, information is exchanged. Knowledge is transferred. Material goods can also be transferred, as in the sharing networks of subsistence hunters, where the catch is distributed among group members or traded for other commodities. In general, these kinds of data are rarely collected because they are difficult to obtain. More often, interactions are recorded, and flows are assumed. For example, many studies ask, 'Who do you seek advice from?', and the assumption is that the resulting data can be used as a proxy for the flow of information (from the alter to the ego). But, in fact, we don't know which bit of information actor A received from actor B. In a few cases, however, direct measures of flows are obtainable, as in tables of the dollar values of flows of raw materials and manufactured goods between countries. Similarly, personnel flows between companies, universities, football teams and the like are readily observable. In general, flows among collective actors like countries and organizations are easier to measure than flows between individuals, since they are public actors that are under observation by many.

3.6 Actor attributes

As noted in Chapter 1, the analysis of social networks involves more than networks. For example, node-level research normally combines network-derived variables, such as node centrality, with non-network attributes of the actors, such as demographic characteristics or personality characteristics. In some cases, the network-based variable will be among the independent variables (as when we predict performance based on centrality, controlling for competence), and sometimes it will be the dependent variable, as when we use personality characteristics to predict centrality. Either way, an important part of the research design will be to collect non-network data that will be combined in the analysis with network data.

A particularly important class of node attributes is the set of behaviors, attitudes, ideas, perceptions, beliefs, etc., that individuals have. Because these are changeable, they provide opportunities to study how networks influence individuals. Choices like what clothes to wear and what words to use are strongly influenced by the choices made by our friends, family and others in our personal networks. For some choices, the number of friends making the same choice is crucial, as in which chat system to use: it isn't useful to choose the better communication platform unless the people you want to communicate with also choose that platform.

3.7 Sampling and bounding

One of the most vexing problems for those just starting out in network research is the problem of bounding the network, although this is a bit of a misnomer. It is not the network that needs bounding, but the study. In some cases the decision seems easy and may even be made tacitly without conscious effort. Well-known examples include Sampson's (1969) study of a monastery, Zachary's (1977) study of a karate club, Bernard and Killworth's (1973) study of a research ship, Krackhardt's (1987) study of a company in Silicon Valley, and Johnson et al.'s (2003) study of a polar research station. All these involve groups that have obvious boundaries. In other cases, the problem seems nearly insurmountable. For example, if we are interested in studying social influences on consumer purchasing, we know we cannot study the entire network – all 7 billion living humans. For convenience, we might choose residents of the city in which we are located, or, more realistically, a small neighborhood. The problem is that no matter whom we choose, we can be sure that a large number of influencers of these people will be outside the sample.

Notice that the problem is not really the size of the network but rather the nature of the research question. If the research interest is social influence on

decision-making, the studies we cited above as examples of easy boundary specification do not look so simple: while the monks may be fairly isolated, the employees of a company are not. The principal influencers of an employee's decisions (say, to leave the company) may well be outside the company, such as family members and members of competing companies.

To deal with this problem, we offer two suggestions. First, if your research question does not allow you to restrict the set of alters that a respondent could name, use a personal-network research design. You still have to decide who will be your respondents, but this could be as simple as a random sample from the population to which you wish to generalize. In a sense, the boundary specification problem involves two sets of actors that need bounding: the egos (in whose ties we are interested), and the alters (those with whom egos have ties). In the case of a whole-network study, these two sets of actors are the same. In personal-network studies, however, they are not, and this is quite liberating. It is also substantively interesting: in a very real sense, in a personal-network design each respondent has their own custom social world with its own boundary.

The second suggestion is to consider whether you are studying a sociological group or not. The realist school of network research design restricts itself to studying only groups, but the nominalist school sees no essential problem with studying networks that are not groups (Laumann et al., 1983). Groups are sociologically real – they are recognized by their members and, in principle at least, they have boundaries; part of the concept of a group is that there are members and non-members, even if in fact the boundaries are fuzzy and/or contested. If one is studying the internal network of a group, then getting the boundaries more or less right is important. One does not want to miss bona fide members of the group, nor does one want to include non-members. Both errors threaten the validity of the study, and the inclusion of non-members can also add considerably to the scope and complexity of the project.

If you are not studying natural groups, then the study boundaries are determined by the research question (see Table 3.4). For example, you might be interested in how the structure of the trust network in different classrooms affects the class's ability to successfully perform group projects. In this case, the network of trust ties within each classroom is assessed, and ties outside the class are not measured. This is not a problem; it does not imply that no ties to the outside world exist, nor that these ties are unimportant. It is just that the research is specifically about how the ties within a classroom affect classroom outcomes. Whether you think that is a fruitful research question is another matter. The point is that choosing an 'artificial' boundary (i.e., one that may not correspond to a sociological group) is not necessarily a threat to the validity of

Table 3.4 Sampling and bounding networks.

Type of sample	Nominalist/etic (researcher-defined networks)	Realist/emic ('natural' groups)
Random sample	Random sample of persons matching researcher's criteria.	Do ethnographic pre-study to determine group members, then sample from it.
Snowball sample	Interview any qualifying actor with a tie to any actor already selected, up to K waves or until quotas or cost limits reached.	Get starter set of group members. Select all group members with tie to previously selected member. Repeat until few new names appearing.
	E.g., ask each person who they inject drugs with, then interview those people. Repeat.	E.g., get self-identified members of gang. Ask them for other members. Repeat.
Census	All persons matching researcher criteria. E.g., all members of the Anthropology dept.	Get list of 'members' from somebody in group. E.g., locate gang member, obtain list of members, interview all/adjust on basis of subjective information.

the research design. And, as explained above, even if one is studying a natural group and knows what the boundaries are, the research objectives may necessitate studying group members' ties to people outside the group. Note that this does not mean that the choice of boundary is irrelevant. A well-known finding in the field is Brass's (1984) finding that an individual's centrality within their department was positively related to power and promotions, but their centrality within the organization as a whole was not.

Most groups have fuzzy boundaries. Even formal groups such as corporations which have membership lists have part-timers, virtual workers, temps, new hires, current applicants, retirees, consultants, etc. One basic approach to approximating the boundaries is snowball or other respondent-driven sampling methods (Johnson, 1990). In a study of communication networks in the king mackerel fishery in the southeastern United States, Maiolo and Johnson (1992) used key informant free-lists (see Borgatti, 1994, for a description of the technique) and commercial license lists to identify an initial set of seeds for a snowball sample. Although a commercial license list existed for commercial fishers and commercial dealers, there was no such list for sportfishers who targeted king mackerel, which is both an important commercial and sport species. In addition, the fact that someone was a commercial fisher did not mean they necessarily targeted king mackerel. Key informants known to target king mackerel in both the commercial and recreational sectors were asked to free-list fishers they knew who regularly targeted that species. This list provided a seed list from which to begin the snowball sample of actors 'who talked to each other about king mackerel fishing'. However, the problem was where to stop. During the course of the snowball sample, which was conducted by both phone and

face-to-face interviews across North Carolina, South Carolina, Georgia and Florida, there were periods of time when there was considerable sample saturation or name redundancy in the elicitation of alters. This saturation represented fuzzy boundaries around the fisher community that were often related to geographical factors. Thus, boundaries were placed on the basis of tie intensity and redundancy in the course of the snowball sample. However, it should be noted that if the purpose of the study is to discover the nature of ties that connect various areas of high redundancy or density in social networks, then ties bridging these areas of high density need to be pursued, and the redundancy criteria may need to be applied across several waves of a snowball sample.

Many studies use a combination of nominalist (or etic) criteria and realist (or emic) criteria. An example of this is Johnson's (1986) study of the diffusion of innovations through a network of commercial fishers. Initially Johnson used the commercial license list obtained from the North Carolina Division of Marine Fisheries to identify commercial license holders in a small fishing community in North Carolina. He could have used the list as the boundary for the network, but the list included anyone who had purchased a commercial fishing license no matter how much they actually fished, and he wanted to weed out the people who really were unconnected to the local fishing world. His solution was to use the fishers' own perceptions to refine the sample. Using the licensing list, he wrote the names of each fisher on a card and asked fishers in the community to sort the names into piles according to how similar they perceived the fishers to be to one another (i.e., an unconstrained pile-sort task). Based on the pile-sort results, it was clear that there were perceived differences among the various license holders based on amount of income derived from commercial fishing. A multidimensional scaling plot (see Chapter 6) revealed two clear clusters that basically broke down by those perceived as full-time fishers as opposed to those viewed as part-time fishers. The final set of actors used for the network survey included only full-time fishers identified in the analysis. Thus, actors' perceptions were used in combination with a researcher-derived list to determine the final set of actors used in the study.

3.8 Sources of data reliability and validity issues

Errors in network data can arise from a multitude of sources. The way questions are framed, the manner in which network boundaries are specified, the willingness of respondents to answer questions, the manner in which data are aggregated, informant accuracy, the erroneous attribution of behaviors, etc., can all create error in terms of missing data or the presence of data that lack validity and may be misleading. Borgatti et al. (2006) examined the effect of data error

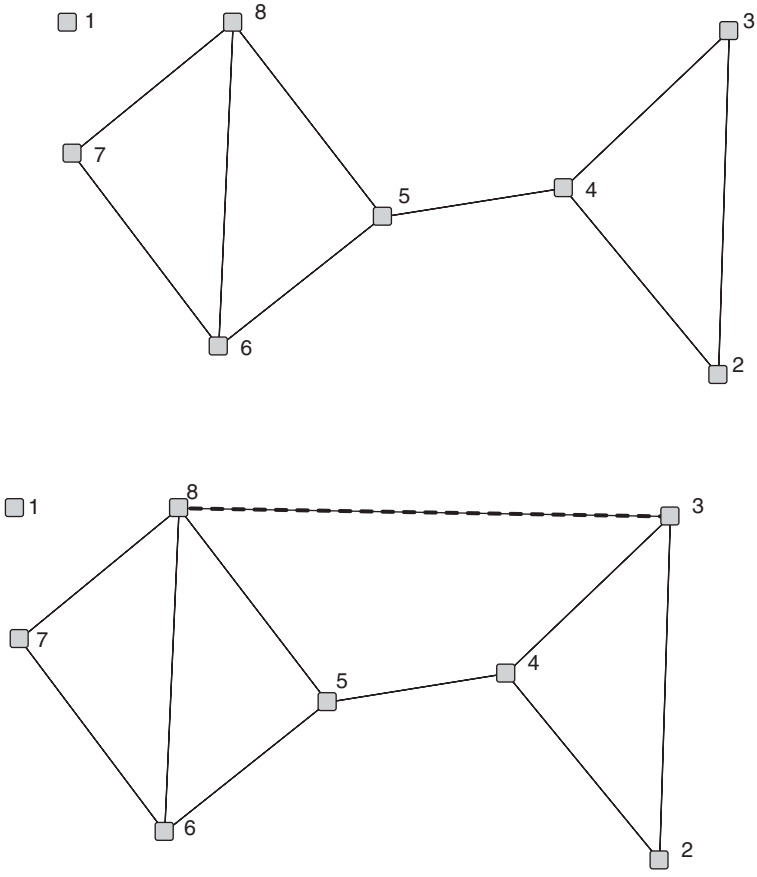
on the measurement of centrality. They examined four kinds of error: omission of nodes, omission of ties, inclusion of false nodes and inclusion of false ties. They found that errors in various centrality measures resulting from the random exclusion and inclusion of edges in random graphs vary as a function of characteristics of the network itself (e.g., density, sparseness), and that the accuracy of measures declines predictably with the amount of error introduced. The latter is good news for network researchers, but it must be remembered that Borgatti et al. studied only artificial networks, not empirically collected networks. Moreover, their results are average values across thousands of trials. Even if the accuracy of a betweenness measurement declines an *average* of 10% when 10% error is introduced in the data, there can be individual cases where the introduction of 10% error causes a great deal more error in the measurement of centrality. For example, consider the network shown in Figure 3.2. A missing tie between nodes 4 and 5 in the top figure would completely hide the brokering importance of these two nodes. Conversely, the erroneous addition of a tie between nodes 3 and 8 would make node 3 look far more important than it really was. The lower graph in Figure 3.2 has the tie between 3 and 8 added and the changes to the betweenness scores are shown in the panel at the top right.

Johnson et al. (1989) used a Monte Carlo simulation approach to study error in networks derived from snowball samples employing a fixed-choice methodology. They found that degree centrality was relatively robust under different sampling conditions, a finding echoed by Costenbader and Valente (2003) and Wang et al. (2012).

Compared to the collection of other types of data in the social sciences (e.g., attribute-based survey data), the collection of social network data can be quite challenging. A major threat to validity in social network research stems from problems of missing data that are due to a number of different sources at a number of different stages in the research process. In addition, errors can arise from data (e.g., a network tie) that are erroneously included – what we have been referring to as 'commission errors'. These sources of error all can lead to model misspecification.

One major contributor to missing data is non-response in network surveys. This can happen if the network boundaries are not properly specified on theoretical or other grounds. Network surveys are extremely susceptible to non-response bias in that missing actors and their links can affect structural and analytical outcomes at both the network and individual levels. Respondents can refuse to participate in the survey at all or can refuse to answer some or all survey questions due to such things as interviewee burden or question sensitivity; they may drop out of a longitudinal study prematurely as a result.

The design of the study and subsequent sample or instrument design (e.g., types and forms of relational questions) for a given social network



Betweenness Centrality

Node	Present	Absent
1	0	0
2	0	0
3	0.159	0
4	0.095	0.381
5	0.127	0.429
6	0.040	0.095
7	0	0
8	0.246	0.095

Figure 3.2 Effects of adding a tie on the betweenness centrality of nodes in a network.

problem and context can also be important in limiting threats to validity (and this can vary cross-culturally). Respondent unreliability and inaccuracy have been shown to produce error of various kinds (but the error is often well behaved, as discussed below). The following provides a summary of some types of error that are of concern, beginning with two that were discussed earlier in the chapter.

- **Omission errors**. Missing edges and nodes can have large impacts on errors in network variables, particularly for some centrality measures. Such missing data can make networks appear to be more disconnected than they really are or make other nodes and edges in the network appear to be more important than they really are (as evidenced by the missing of a single tie between nodes 4 and 5 in Figure 3.2). If data are being collected in surveys using open-ended format questions, omission errors are most frequently a result of the insufficient elicitation of respondent's alters (see Chapter 4 for more discussion).

- **Commission errors**. Like omission errors, the erroneous inclusion of nodes and edges can affect the ultimate determination of node-level measures and the identification of key nodes (as is clear in Figure 3.2).
- **Edge/node attribution errors**. These result from assigning a behavior or attributing something to either an edge or node incorrectly. Misassignment of a behavior to a node can yield attributed linkages in a network that in reality do not exist. Attribution error is a common problem in the interpretation of two-mode data that has been converted to one-mode. For example, two individuals in a university program may co-attend a large number of classes. We therefore assume a connection (either pre-existing or as a result of meeting in class). But it could easily be that one of the individuals is a non-traditional student who is older and married and does not hang out with other students in the program. Treating co-attendance as a tie is, in this case, a mistake. Collection of other relational data could help in determining whether an active tie actually exists in this case (e.g., triangulation).
- **Data collection and retrospective errors**. Care should likewise be taken when using network data collected from individuals where the network elicitation question deals with reports of behavior, particularly having to do with social interactions of a temporally discrete nature. For example, questions that are of the kind 'who are the people you interacted with yesterday in the plaza?' are notoriously prone to error. Bernard and Killworth (1977), Bernard et al. (1980, 1982), and Killworth and Bernard (1976, 1979) conducted a series of studies on informant accuracy in social networks involving fraternity members, ham radio operators, and deaf people communicating with teletype machines, to mention a few. They found that people were inaccurate in their reporting of interactions with others. For example, ham radio operators, who kept logs of radio conversations, made both omission and commission errors in their retrospective reporting of radio interactions. Bernard, Killworth and Sailer asked the operators to list all the people they talked to on the radio the day before; the researchers then checked the accuracy of the reported communications with the actual communications as recorded in the log books and found them to be woefully inaccurate. The overall conclusion of their studies was, in their words, that 'what people say, despite their presumed good intentions, bears no useful resemblance to their behavior' (Bernard et al., 1982: 63).

The Bernard et al. research led to a flurry of other studies on the topic. An important study by Freeman et al. (1987) and Freeman and Romney (1987) found that informants are more accurate in reporting long-term patterns of behavior than behaviors at some point in time. They observed the participants in a colloquium series at the University of California Irvine throughout the quarter. On the day after the last colloquium of the quarter, the people who attended were asked to list all the participants present at that last colloquium. There were inaccuracies, as expected, but these inaccuracies were patterned and predictable. Omission errors included people who normally did not attend the colloquium but happened to be at the last one, while commission errors included people who usually came to the colloquium but

happened to not be there for the final one. Thus, individual informants were reporting more on what usually happened rather than on what happened during a specific colloquium. A better way to ask the question posed at the beginning of this section about plaza interactions would be 'who are the people you usually interact with in the plaza?' or 'who are the people you interacted with most in the plaza over the last two weeks?' These reports of long-term patterns of behavior are much less prone to error.

Research on ego biases in cognitive networks (Krackhardt, 1987, 1990; Kumbasar et al., 1994; Johnson and Orbach, 2002) has shown that some individuals in the network are more accurate about reporting linkages than others. They find that active, more powerful nodes tend to be more accurate. Johnson and Orbach (2002), for example, found that the more central an actor is in the political network, the more accurate their cognitive networks. These all have implications for methods for assessing and weighting the reliability and validity of network data and for potentially fixing missing data problems.

- **Data management/data entry**. Errors due to data entry, coding and transcription/translation are well known in other analytical and modeling domains, but they can be even more problematic in the network context as they can have larger effects. Fischer (2006), for example, suggests that some of the contested results of the McPherson et al. (2006) research on the shrinking of Americans' social networks in a longitudinal study of the General Social Survey may be due in part to what Fischer refers to as random or technical errors (e.g., software problems, interview procedures, coding errors).
- **Data fusion/aggregation**. Decisions often have to be made on aggregating data on different temporal, relational and spatial scales. Such aggregations, if done improperly, can create errors at a variety of levels. For example, when aggregating longitudinal real-time or streaming data for analysis, important individual nodes and edges may be excluded because they have lost their importance in the network. As in the boundary specification problem, there should be some guiding principles, preferably of a theoretical nature, for making aggregation decisions (e.g., before and after a hypothesized important event).
- **Errors in secondary sources and data mining**. Various forms of secondary source data may have inherent biases which should be considered in any analysis. This type of data can be easier to collect than primary types of data (e.g., data scraped from the Web), but it can be fraught with errors at a variety of levels. Examples of important questions one should ask when using secondary source data include: if, instead of obtaining this tie from some records, we asked a survey question, what survey question would the tie correspond to? Are nodes really the same? For example, telephone records show ties between phones. But the phones may be used by multiple people, and a given person may have multiple phones. Does the observation of two individuals at the same event imply a tie? Are records temporally comparable, at the same scale, etc.? For further discussion of this issue, see Section 4.6 in Chapter 4.

- **Formatting errors**. In data mining or Web scraping there are errors that can be due to differences in document or website formatting. These errors can lead to the over- or under-representation of terms, actors, attributes, etc. in the data retrieval process. Care should be taken that any relations assigned among nodes are not an artifact of formatting errors. In addition, Web scraping and automated data mining methods should be scrutinized for consistency in the operationalization of important concepts. The bottom line is that the quality of a study is a function of the quality of the data: garbage in, garbage out.

3.9 Ethical considerations

Network research poses different ethical challenges from those of other kinds of social research, particularly in whole-network research designs. In whole network designs, it is impossible to collect the data anonymously. Personal-network research designs can be anonymous, both with respect to the respondent and the people they mention (e.g., they can use nicknames). But for all practical purposes, full network designs require that the respondent identify themselves, which means the researcher can only offer confidentiality. This makes it imperative that the researcher make it clear to the respondent who will see the raw data what can reasonably be predicted to happen to the respondent as a result of an accidental breach of confidentiality.

A related issue is that, unlike other research, non-participation by a respondent in a network study does not necessarily mean that they are not included in the study. Even if an actor chooses not to fill out the survey, other respondents may still list that person as a friend, enemy, etc. A person who does not wish to be embarrassed by their poor standing in the group will still be found to be the person most often named as difficult to work with. This can be remedied by eliminating all non-respondents from the dataset altogether. However, this may wreck the quality and representativeness of the data, which introduces its own ethical issues. This is particularly a problem in applied settings, where decisions will be based on the results of the study. The researcher can, of course, consider it enough to warn management of the problem, but realistically the researcher knows that the management is not going to fully appreciate the depth of the problem, especially since it may be difficult to explain just how the picture is misleading without revealing the very information that the researcher is trying to suppress.

The non-participation issue points to a more subtle underlying difference between network research (of all kinds) and conventional social science. Whereas in conventional studies the respondent usually reports only on herself, in network studies the respondent reports on other people, some of whom may not wish to be reported on. And if these people were not also intended to be

respondents in the study, they will not have been contacted to sign consent forms. As a matter of general principle, this does not seem unethical, as the respondent owns her own perceptions. This needs to be considered on a case-by-case basis, however. For example, if the respondent reports seeing someone engage in illegal activities, there is a clear implication that the named party does in fact do illegal things; it is not 'just' a perception as in 'I think John respects me'. In general, the researcher needs to balance his research need against the dangers posed by revelation to both the alters and the respondents who tell on them. Also, an interesting aspect of many social ties, particularly those based on role relationships such as 'is a friend of', is that neither person owns the relationship exclusively; it is a joint creation, and so it is at least plausible to argue that neither party can ethically report on it without the consent of the other.

The issue of which ties it is acceptable to ask about is particularly important in organizational research, especially when the price of getting access to the organization is providing feedback to management. It is generally accepted that the behavior of employees of an organization is open to scrutiny by management. Supervisors base their evaluations of subordinates on all kinds of factors, both formal and informal. How employees relate to each other is something that is of legitimate interest to managers and, indeed, in the case of sexual harassment, an obligation. It is also generally accepted that some things are private; what employees do in the privacy of their own bedrooms with their spouses is none of the organization's business. But what of employee friendships? This is one of the most commonly asked questions in organizational network studies. As a general rule, the network researcher is far more interested in informal ties, including negative ties, than those dictated by the formal structure of the organization. It seems at least plausible to argue that these sorts of questions fall into a gray area between acceptable management scrutiny and invasion of privacy.

Another way in which network studies (of the full network type) differ from conventional social science studies is that missing data are exceptionally troublesome. If a few highly central players are missing, the resulting network could be quite different from what it would have been had those people responded. This creates unfortunate incentives for network researchers to discourage respondents from opting out of a study. As a result, they may not do a fully adequate job of outlining the risks to respondents. In organizational settings, they will also be sorely tempted to get the boss to send a clear message to employees that they should participate in the survey. This might not be coercive from a workplace legality standpoint, but many academic human subject review boards (institutional review boards in the US) would disagree.

Another issue that is special to full network research has to do with data visualization. In most social science research it is variables, not respondents, that are the focus of interest. Respondents are merely anonymous replications – the

more the better. Fundamentally, they are treated as bundles of attribute values. Consequently, it is rarely useful to express the results of quantitative research by providing displays of individual data.[3] But in network analysis it is extremely common to present a network diagram that shows who is connected to whom. Such diagrams are not highly digested outputs of analysis, but rather low-level displays that represent the raw data; the outgoing arrows from any node have a one-to-one correspondence with that person's filled-out questionnaire, compactly revealing each person's responses. The obvious solution, of course, is to suppress the node labels or identify nodes only categorically, such as by department or gender. But the level of risk to respondents here is much higher than most consent forms would suggest, because organizational members can often deduce the identity of one person (e.g., the only high-ranking woman in the Boston office), and once that person has been identified, their known associates can sometimes be deduced as well, eventually unraveling the whole network. At the very least, participants can often identify themselves (e.g., when they remember listing exactly seven friends and no other node in the graph has exactly seven ties).

A final point of difference is not fundamental to the field but has to do with the fact that most potential respondents do not know much about social network analysis. Most people today have a great deal of experience filling out survey questionnaires in a variety of contexts from political polls to marketing research to job applications. Although new media like Facebook present some new challenges, when it comes to simple questionnaires we would argue that people have an intuitive feel for the potential consequences of disclosing personal information. Coupled with explicit consent forms that outline some of the risks, most researchers would agree that respondents' common sense provides adequate protection. Network surveys, on the other hand, are a whole new ballgame. Most respondents in a network study will not have participated in one before, and, in organizational contexts, managers receiving network information will not have done so either. As a result, there is a greater burden on researchers to be clear about the risks. Even if a consent form were to clearly state that the data would *not* be kept confidential and would be reported back to the group, many respondents would not be able to fully imagine how it would feel to be identified in the analysis as, say, a peripheral player whom nobody really likes.

In short, the design of a network study generally requires more attention to ethical issues than ordinary studies, particularly in organizational settings. We advise using an expanded consent form that explains more about the outputs of network analysis than is customary in other types of research. For more suggestions, see the papers by Borgatti and Molina (2003, 2005).

---

[3] This is not true of qualitative research, however, where it is common to provide direct quotations (albeit anonymously) from individual respondents.

3.10 Summary

Network studies need to be carefully designed to take account of the particular features inherent in social networks. Personal-network research designs, in which information is gathered from a random sample of actors who give information on their connections, pose fewer data collection problems than whole-network designs. However, the downside is that ego-network studies fail to capture the full structural properties of the whole network. Determining which actors to include in a study can be challenging, and network boundaries are not always clear. Even when formal groups are considered, there is often a degree of ambiguity about membership. When the boundary is not clear, snowball or respondent-driven sampling can be used to determine a population. Errors in network data can occur from a variety of sources, and any study needs to take steps to try and reduce these errors as much as possible. The nature of network data and subsequent analysis and visualization give rise to a number of ethical considerations which are particular to network studies, and these need to be clearly thought through before data collection begins.

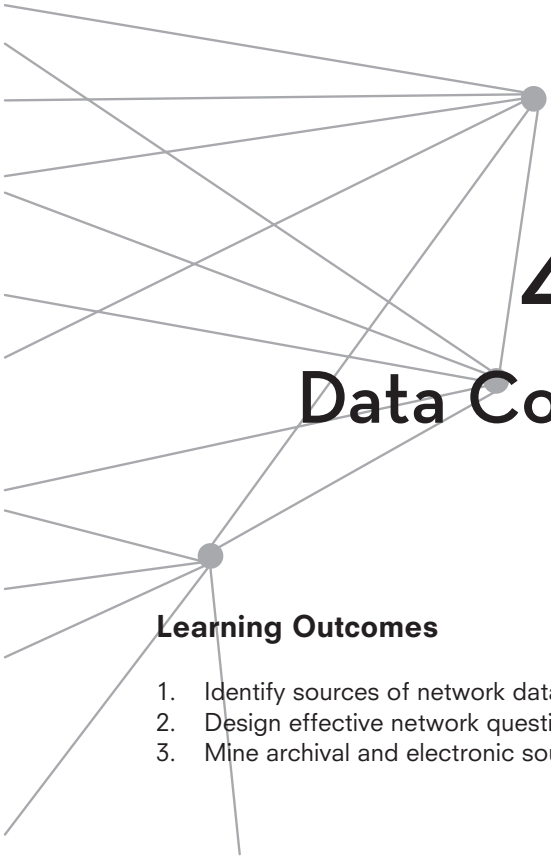
3.11 Problems and Exercises

1. For each of the research problems in Chapter 1, Problem 1, what are the independent and dependent variables? Based on the designs outlined in Tables 3.1 and 3.2, what type of research design does each of the research problems suggest?
2. For each of the research questions below, discuss whether a whole or personal network approach is more appropriate.
    a. How do social relations in a university sports club influence members' attitudes towards university sports policies?
    b. To what extent is smoking behavior among adolescents affected by their social networks?
    c. How much do voting patterns in a state legislature conform to political party affiliation?
    d. How do immigrants' social networks affect cultural assimilation?
    e. Does the network structural position of a manager in a financial firm impact that manager's performance?
    f. To what extent is toothpaste brand selection affected by consumers' social networks?
    g. To what extent is the social network structure at a commercial fish camp in Canada influenced by ethnicity?
    h. What factors influence the development of cooperative social relations among activists in an environmental social movement?
3. Produce a hypothetical social network study example for each of the three major types of research designs: experimental, quasi-experimental and observational. For each example, identify the independent and dependent variables.

4. For the hypothetical observational design presented in Problem 3 above, is the design cross-sectional, retrospective, or prospective? What are the advantages and disadvantages of each in the study of social networks?
5. Bounding the network in whole social network studies can be challenging and is important in designing valid research. For each of the social network examples below, provide a discussion for how the networks might be bounded in the design of a whole network study.

    a. A study of fraternities at a medium-sized Midwestern university
    b. The study of an informal gay group in an urban neighborhood
    c. The network relations among active hunters in a small village in the Amazon
    d. Relationships among NGOs involving a dam project in West Africa
    e. Food-sharing networks in a village in Central Asia
    f. The political network of community activists in a moderate-sized city

6. Nonresponse in social network surveys can be a major threat to the validity of a social network study. What are some of the ways researchers can minimize survey nonresponse?
7. What are some of the key ethical concerns in social network research as opposed to other types of more traditional research, such as classical social surveys?



4 Data Collection

Learning Outcomes

1. Identify sources of network data
2. Design effective network questionnaires
3. Mine archival and electronic sources for network data

4.1 Introduction

On the surface, asking network questions might seem pretty straightforward. For example, we just ask 'Please tell me the names of all your friends'. But there is a lot more to it than that. First, how will respondents interpret the term 'friend'? Can we expect 'friend' to have the same meaning for all respondents no matter what their ethnic, regional, educational or social class? Second, do we ask the question in an open-ended format, or do we provide the respondent with a roster of names to choose from (i.e., a closed-ended format)? If we use an open-ended question, respondents may forget to list people, and if we use a list or aided format we need to know all the names of network members in advance. Third, do we just want to know whether or not a tie exists between two people, or do we want to know the strength of that tie? And if we want to know something about its strength, do we use an absolute or relative scale? Finally, do we use pen and paper or some type of electronic format for collecting the data? The answers to these questions will ultimately depend on characteristics of the population, the type of social relations being studied and, above all, the research objectives.

In this chapter we discuss a variety of issues relating to the collection of primary network data in full network research designs. This includes working

around respondent sensitivities and selecting the right question formats, including closed-ended versus open-ended rosters, use of rating scales, multi-item batteries, electronic surveys, and so on. Many of the issues we discuss apply to personal-network research designs as well, but we note that there is a chapter (Chapter 15) devoted entirely to personal-network designs. We close the chapter with a discussion of collecting archival and electronic data.

4.2 Network questions

In principle, we can study networks of all kinds of entities and all kinds of relations. Our research objectives, for example, may call for us to study trust ties among terrorists over time. Unfortunately, that may not be possible. There are always practical considerations that get in the way. Even if we can get respondents to talk to us, we will rarely have carte blanche with respect to what we ask, and how much we ask. Depending on the context, some types of relational questions are more sensitive than others, and this respondent sensitivity can impact interviewees' willingness to answer questions or, worse, answer honestly and competently. Further, such sensitivity can vary by cultural context (e.g., economic relations may be more sensitive in some cultures than others), can vary over time (e.g., some relational questions may be of a more sensitive nature at the beginning of a longitudinal study than toward the end), and can vary as a function of the data collection methods employed (e.g., face-to-face versus online interviews).

The proper selection of the network questions and formats is critical to the success of any network study. The structure of network questions greatly influences the validity and reliability of respondent answers due to such things as question clarity, burden, sensitivity, and cognitive demand. Many of the issues concerning standard survey and questionnaire development and design apply equally to the study of social networks. However, social network questions are somewhat unique in that we are not simply asking about some attribute of the respondent or ego (e.g., age); we are asking them about their web of social relations that may evoke emotional responses or tax their abilities to recall aspects of their network relations and/or network behaviors.

A short case study illustrates the point. Johnson et al. (2003) studied the network dynamics at polar research stations. At the beginning of the four-year study, the researchers were initially interested in the formation of friendships and the ability of individuals to assess potential friendships one day following the initial contact. One of the researchers attended the first training exercise of the first winter-over crew preparing to deploy to the South Pole station. During a break in training, the crew members were given a questionnaire asking

them to rank the other members of the crew from 1 to n – 1 with respect to how likely they were to form a friendship with each one over the coming winter. The exact request was as follows:

> Please rank the following members of the winter-over crew in order of their friendship or potential friendship to you from 1 to 20. The member you feel closest to should be ranked '1' while the individual you feel most distant from should be ranked 20. We realize this task is difficult because of the short amount of time you have known other members of the group; your judgments may be based more on your sense of the potential for friendship than on any current relationship. Whatever the difficulty it is important you fill the form out completely. Thank you!

Immediately, several of the crew began to grumble and protest and one crew member threw down his pencil and walked out of the room. One respondent wrote on the survey form 'I would really like to do my best to cooperate and help but *please* [respondent's emphasis] no more rankings', while yet another put '1' next to each crew member's name. This resistance to the administered network question was related to two primary problems. First, it was discovered that the initial period of group formation was filled with great optimism (i.e., a utopian stage), where there was a general perception that everyone would get along and be friends over the course of the austral winter. The task of having people rank-order one another in terms of potential friendship created quite a negative emotional response on the part of crew members, since they believed at this point in the group formation process that 'everyone' would be friends and ranking people meant that some people would be ranked near the bottom of people's list, therefore implying a possible lack of friendship. Thus, both the type of relation and how it was measured (i.e., rank-order) were problematic in practice. The mix of a rank-order collection method and actors' judgments as to expectations of friendship fostered a 'perfect storm' in terms of sensitivity and interviewee burden. Eventually the researcher met with the crew at an 'all-hands meeting'. They all discussed the survey, and agreed on a compromise: the survey would ask the crew about 'who one interacts with socially' rather than 'friendship' and would measure it on an 11-point Likert scale (from 0 to 10) anchored with words from never (0) to most often (10), as shown in Figure 4.1. Thus, the relational question and the method of measurement were ultimately determined in concert with those being studied. A beneficial side-effect of this process was that it created a sense of investment in the design of the study on the part of the crew and helped foster an extremely high and sustained response rate over the winter and in subsequent years.

It is worth noting that while respondents were initially very sensitive about discussing their feelings about each other, they had fewer problems doing so

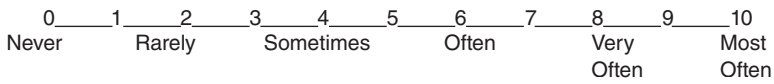


Figure 4.1 Relative interaction scale developed in cooperation with the South Pole winter-over crew.

later, and were even willing to answer questions about their negative feelings toward others. This reflects a temporal component in question sensitivity and its ultimate impact on potential non-response bias. Thus, the maturity and other characteristics (e.g., cultural context) of the group itself may have an impact on the level of emotional reaction to one or a given set of network questions. This variability means that network surveys have to be pre-tested and in some cases co-developed with the research subjects. This is particularly true in management consulting settings where the reason for collecting the network data is that there is some kind of political or interpersonal problem. Under those conditions, people become very wary of researchers asking sensitive questions like 'who do you trust?' and 'who don't you get along with?'

It is essential to do some ethnographic background research to explore the types of network relations and labels or terms that are appropriate for a study and to discover the best way to word the questions. Once the questions are developed they should be pre-tested to make sure respondents are clear about what they mean. The greater the heterogeneity of the backgrounds of the members of the group, the more critical this becomes.

It is also very useful to do some ethnographic work at the back end of a study. Patterns found in the analysis can often be quickly explained quite readily by the members of the group themselves. It is also useful to test the results – which could be spurious – against their insider knowledge to see if they have validity from a native's point of view. We refer to the practice of doing ethnography at either end of a quantitative study as 'the ethnographic sandwich'.

4.3 Question formats

A fundamental issue in the design of network questions is whether to use an open- or closed-ended format. Figure 4.2 provides examples of the types of questions used in each. With a closed-ended question format, the set of nodes comprising the network are chosen in advance and a roster created; respondents are then asked about each person on the roster. The main advantage of using rosters (besides guaranteeing that the set of respondents matches the set of actors being asked about) is that respondents are less subject to recall error. All they have to do is respond to each name they are asked about. Some may think this recall problem in open-ended question formats is overblown, but a simple

empirical example provides some insight. In the South Pole study discussed extensively throughout this book, winter-over crews were debriefed by the researchers at the end of each winter. The crew sizes across three separate years were quite small, ranging from 22 to 28 people. The crews in each station had been together for well over a year and spent 8.5 months together in total isolation over the austral winter. If asked about any randomly chosen crewmate, crew members could fill books of information about them. But when asked in the debriefing interviews to list all their crewmates, it was found that they would commonly forget about up to 25% of their fellows. So, recall error is a significant issue.

Another advantage of the roster is that it limits potential biases affecting the probability of an actor being selected by a respondent. Imagine someone in an organization who works in the basement in a physically isolated part of the building. If actors in the organization were asked in an open-ended format to list 'people that they don't know but would like to get to know', this person may systematically be left out because of the limitations of human recall. The lack of selection of these physically isolated people may not have anything to do with who they are as human beings or as potential friends; it may just be a matter of people forgetting about them because of their location. The disadvantages of the roster method are that (a) it requires deciding ahead of time which nodes will be asked about, and (b) it can be cumbersome and intimidating when the list of potential alters gets large – say, more than 500 names. The latter problem can be ameliorated by the use of hierarchically organized rosters (especially in online surveys), such as having the respondent first select an organizational unit, then respond to each of the names in that unit. The same can be done with lists organized alphabetically.

In comparison, unaided or open-ended question formats require no prior decisions about who to obtain information about. So, in cases where the list of potentially relevant alters is large (e.g., the population of American consumers), and/or insufficient ethnographic work has been done to have a clear idea of who to ask about, the open-ended approach may be the only way to ask questions. In this case respondents are asked to list people that they, for example, 'talk to' or 'share needles with'. Besides recall issues, open-ended questions have a number of potential disadvantages in a full network research context. The biggest issue is identifying the actor whom a respondent names. If they mention 'Bob Smith', is that the same as the 'Bobby Smith' whom someone else mentioned? And is it the same as another respondent in the study whom the researchers know as 'Robert Smith'? This is particularly a problem in populations where full or even real names are rarely known, such as drug injectors on the streets of Hartford, CT (Weeks et al., 2002). In personal-network research designs, this is not a problem because we do not need to draw connections across different respondents.

Closed-ended (aided)	**Example**
• Boundaries are known and all actors listed • Becomes cumbersome as networks grow in size • Fewer concerns about respondent recall and accuracy • Each actor has approximately an equal chance of being selected	Who would you converse with if you met on the street? (check as many as apply) Felicia Hardy ☐ Steve Rogers ☐ Sam Wilson ☐ Patsy Walker ☐ Bruce Banner ☐ Ted Salis ☐ Kitty Pride ☐
Open-ended (unaided)	**Example**
• Much more subject to recall error • Can use a fixed-choice method limiting the number of actors elicited • Each actor in the network does not have an equal chance of being chosen given recall and free-listing issues • Better for face-to-face interviews where probing can be used	If you wanted to learn more about what goes on in the Avengers organization, who would you talk to? (Please list as many relevant names as you can in the spaces provided) _____ _____ _____ _____

Figure 4.2 Examples of closed- and open-ended network question formats.

Another potential problem with the open-ended format concerns the size of respondent lists. For example, in an unlimited choice format, if respondent A lists 30 alters while respondent B lists 15 alters, can we conclude that respondent A has a larger network than respondent B? That might be the case, but it might not. Perhaps A is very energetic and really thinks long and hard about the question, while B is tired and bored with the survey and just wants to get it over with. One way to deal with this is to limit the number of names people can provide. Unfortunately, this has problems as well because people use varying heuristics for recalling names (Brewer, 1995a, 1995b). In the South Pole debriefing, it was apparent that some respondents were mentally walking through the station to remember names. They moved from the garage through the carpenter shop to the generator room into the bar then the galley, and so on, as they recalled fellow crew members. Others recalled names based on social groupings. All of this is to say that limiting names to a certain number can systematically bias the resulting networks.

The recall problem in open-ended elicitation can be somewhat ameliorated with the use of different types of interview probes (Brewer, 2000). For instance, in the free-list example above, if the interviewer had verbally repeated the names listed by each respondent and simply asked 'So you have listed John, Susie, … and Sarah: are there any other crew not listed that wintered-over with you?', possibly followed by 'Can you think of anyone else?', there is a good

Repeated roster

Q1. Please indicate with which of the following you would converse if you met them on the street.

Felicia Hardy	☐
Steve Rogers	☐
Sam Wilson	☐
Patsy Walker	☐
Bruce Banner	☐
Ted Salis	☐
Kitty Pryde	☐

Q2. Please indicate with which of the following people you work.

Felicia Hardy	☐
Steve Rogers	☐
Sam Wilson	☐
Patsy Walker	☐
Bruce Banner	☐
Ted Salis	☐
Kitty Pryde	☐

Multigrid

Q1. In the grid below, please indicate those people you would converse with if you met them on the street.

Q2. In the grid below, please check off the names of the people you work with.

Q3. In the grid below, please check off the names of a selected set of people whom you don't know but would like to know, based on things you've heard, or their interests, etc.

	Q1 Converse with	Q2 Work with	Q3 Want to Know
Felicia Hardy	☐	☐	☐
Steve Rogers	☐	☐	☐
Sam Wilson	☐	☐	☐
Patsy Walker	☐	☐	☐
Bruce Banner	☐	☐	☐
Ted Salis	☐	☐	☐
Kitty Pryde	☐	☐	☐

Figure 4.3 Repeated roster versus multigrid formats.

chance that the respondent would have noticed the missing crew member(s) and provided the additional name(s). Of course, the use of probes can be readily done in face-to-face interviews but would be much more difficult in other types of self-administered survey formats. In some settings, such as organizations, it is possible to use visual aids to help stimulate recall, such as providing office maps or unit-level organizational charts.

If a roster format is chosen, there are a number of further decisions to make about how the questions and lists should be structured. The two primary formats for closed-ended questions are repeated rosters and multigrids. Figure 4.3 provides examples of each type of format. In repeated rosters the same list of network members is repeated following each network question. Respondents can then circle or check the appropriate names in response to each of the questions. The multigrid format places the lists in a series of columns with each column associated with a relational question. Again, respondents can check or circle the appropriate answers. The two are similar in terms of potential reliability and validity, but the multigrid is a more compact format; if one is using

pencil-and-paper collection, the latter format can help reduce the number of pages in the survey. Sometimes this has a beneficial psychological effect on respondents in that it makes the survey appear shorter than it would using the repeated roster. There is probably nothing more daunting to a potential respondent than when a researcher pulls out a one-inch thick survey and places it on the table in front of them.

The roster examples in Figure 4.3 involve respondents making simple yes/no decisions about a given tie: either there was a tie or there was not. This checklist method has an advantage in that it is less cognitively demanding on respondents. Moreover, it is quick and easy to administer. However, one drawback is that it provides no discrimination with respect to a tie's value such as tie frequency or strength. For that, we need to use some kind of ratings approach that allows for the assessment of the frequency of contacts or the strength of relationships. We can elicit tie values by using either an absolute or a relative scale. Figure 4.4 shows types of questions in the two approaches. With absolute scales, we are attempting to assign to each person on the list that is given to respondents or listed by respondents the degree of interaction within a specified period of time. So we might ask 'Do you seek advice from _____ once a year, once a month, once a week?' When using absolute scales, it is important to do sufficient preliminary research to determine the appropriate time intervals, or risk getting no variance (e.g., all respondents choose 'every day' because in that setting everyone interacts multiple times a day). Another issue with absolute scales is that people are not particularly good at them. Given two alters, respondents probably have a good idea which one they interact with more often, but they may be inaccurate about whether it is once a week, once every couple of weeks, or once a month.

Some researchers use questions that are more explicitly ordinal and generic, such as an *n*-point Likert scale anchored with words such as 'very infrequently', 'somewhat infrequently', 'neither infrequently nor frequently', 'somewhat frequently' and 'very frequently'. Such questions are easier to write, since the researcher does not have to know what range of frequencies to ask about. However, such scales are also quite vulnerable to response sets. Some respondents are very liberal and rate everyone as somewhat or very frequent, while others are more conservative, rarely venturing above the middle point. Others are conservative in another way: they rarely venture far from the middle point in either direction. We can try to mitigate this problem by using a relative scale such as 'much less than average', 'less than average', 'about average', 'more than average' and 'much more than average'. A thoughtful respondent should see this as an invitation to use all ends of the scale. Unfortunately, there is also less information in this scale than in an absolute scale. Greater than average interaction for one person might be interacting once a day but for another

Tie frequency	*Format of question*	*Comments*
Absolute	'How often do you talk to _____, on average?' – Once a year or less – Every few months – Every few weeks – Once a week – Every day	Need to do pre-testing to determine appropriate time scale Danger of getting no variance Assumes a lot from respondents
Relative	'How often do you speak to each person on the list below?' – Very infrequently – Somewhat infrequently – About average – Somewhat frequently – Very frequently	Assumes less of respondents; easier task Is automatically normalized within respondent • Removes response set issues • Makes it hard to compare values in different rows

Figure 4.4 Question formats for assessing frequency of contact.

person it might be once every two weeks, and there is no way to distinguish them. A value of '4' for one person may refer to wildly different levels of interaction than a '4' for someone else.

The same kinds of trade-offs are seen in full ranking data. In the full ranking task, the respondent is asked to rank everyone, except themselves, from 1 to $n - 1$. Ranking has the advantage of asking for only ordinal judgments (is A more than B?). These are more natural than rating scales, which ask the respondent to assign a number between, say, 1 and 7 to represent their feelings about each other person. However, as the list of names gets longer, respondents find full rankings increasingly difficult to do and find ratings much easier and faster to do.

One other technique that is worth mentioning is the idea of breaking a single complicated question into more numerous but simpler questions. Rating every individual on a list on a 1–5 scale is a slow process compared to checking names off. So, one possibility is to convert the rating question into multiple check-off tasks. For example, instead of asking 'How often do you see each of the following people?' using a scale of 1 = once a year, 2 = once a month, and 3 = once a week, we can instead ask three separate questions: 'who are the people on this list you see at least once a year?', 'who are the people on this list you see at least once a month?', and 'who are the people on this list you see at least once a week?'. It may seem counterintuitive, but it is often faster and easier this way. This is especially true if electronic surveys are used because the second question only has to list the names that were selected in the first question, and the third question only has to list the names checked off in the second question. The task then becomes lightning fast.

4.4 Interviewee burden

Sometimes the size and particular boundaries for a network are dictated by the methods employed. Some data collection methods are labor-intensive and burdensome, where such burden varies as a function of network size. Two examples of this are personal-network studies and cognitive social structure studies, where respondents are asked to report on the network connections of all other actors in the network (Krackhardt, 1987; Kumbasar et al., 1994; Johnson and Orbach, 2002). In these types of studies, the number of data points needed from a respondent increases with the square of the number of alters, rapidly increasing the burden on the respondent.

In a study by Johnson and Orbach (2002) on political networks and the passing of a piece of environmental legislation, there were potentially over 400 actors in the political network involving legislators, staff, resource managers, lobbyists and private citizens. The researchers were interested in the relationship between knowledge of the political landscape and political power. However, the respondents were very high-status people (e.g., the President Pro Tem of the North Carolina Senate, cabinet-level secretaries, legislative committee chairs and co-chairs, etc.), who will not grant a researcher a 3-hour interview. To deal with this, the study began by asking 10 politically knowledgeable key informants (Johnson, 1990) to free-list actors who were seen as 'important' in the development and passing of a particular piece of environmental legislation. The top 45 names most frequently listed by the key informants were used to bound the network. This is like the data-driven, emic or realist strategy for bounding a network discussed in Chapter 3, except that in the legislative case described here methodological realities were partially driving boundary specifications.

In addition, for the cognitive network data collection, the respondents were asked to name only three people on the list whom they thought each of the political actors talked to most about a given piece of environmental legislation over a period of time. This reduced the task to approximately 135 reported dyads, which was much more reasonable, although still daunting, given the research population. As we shall see in the chapter on personal networks, this is also a very important consideration in designing personal-network surveys.

Interviewee burden, more generally, can lead to various kinds of non-response on the part of actors. There is plenty of literature discussing these issues in survey research (Dillman, 1977; Church, 2001) and this is also a real concern in personal-network approaches. However, unlike typical survey research where researchers are willing to accept at least some level of missing data and non-response bias, in whole-network surveys such levels are unacceptable and pose real threats to the validity of any study (as discussed in Chapter 3). As we have seen from earlier examples, the goal is to minimize respondent anger

and frustration. One potential source of respondent displeasure, which can result in non-response or early withdrawal from the interview, is the length of the interview itself. A major reason people state for their unwillingness to participate in surveys is being 'too busy' or having a lack of motivation (Sosdian and Sharp, 1980). It is important to keep in mind that network interviews and the complexity of certain social network methods can place huge temporal and cognitive demands on respondents.

There are no hard-and-fast rules about what makes an interview too long or too demanding. However, the shorter the network survey instrument, the better, particularly if one is engaged in a longitudinal study where sustained participation is crucial. One rule of thumb for achieving an optimally sized network survey instrument is to include only those questions that are theoretically critical for the study at hand – no more and no less. If you are uncertain about the theoretical relevance of a network question, you should conduct exploratory or ethnographic research to find out. Conducting ethnographic work prior to a full network survey can help in assuring the reliability and validity of network questions and in understanding the capacity of respondents to answer instruments of a given size (e.g., chief executives of companies and fishers in Cuba may face different time constraints and different levels of enthusiasm).

One final note is that the placement of network questions in any survey may impact outcomes. As discussed earlier, network questions are often cognitively demanding. If such questions are placed at the end of an already extensive survey, there are chances that respondents may be less thorough in their responses or may even refuse to answer. It is to issues of cognitive demand and interviewee burden that we now turn.

4.5 Data collection and reliability

As discussed in Chapter 3, whole network approaches can be sensitive to missing data (Borgatti et al., 2006). This is particularly true for smaller networks, where the absence of actors or ties can have relatively large effects. The manner in which we collect network data can have a profound impact on actor participation and on the reliability and validity of the social network data sought.

Table 4.1 shows some of the ways in which researchers have typically collected network data. The columns in the table represent a few of the trade-offs one should consider in the course of choosing a data collection method for a network study. As we have seen from the polar research station network earlier in this chapter, some network questions may be more emotionally sensitive than others. Self-administered network surveys, including mail-out and online surveys, may minimize the degree of self-consciousness on the

part of respondents. In addition, they do not suffer from reactions to the interviewer, and they are very convenient for the researcher. On the other hand, self-administered surveys that are not hand-delivered typically have much lower response rates.

An important means for reducing non-response on the part of actors is the building of rapport with respondents before administering the survey (Johnson, 1990). This is particularly a problem with self-administered mail-out and online surveys, where there is limited opportunity to establish contact and create a relationship. Dillman (1977) provides suggestions for overcoming some of the disadvantages of mail-out and phone surveys in terms of increasing response rates. However, face-to-face data collection provides the greatest opportunity for establishing rapport with respondents. Additionally, it facilitates the use of elicitation interviewing techniques for the collection of network data, such as various probing techniques to improve respondent recall (Brewer, 2000; Johnson and Weller, 2002). Network elicitation is difficult to do in a less interactive context and limited in phone and group interview formats. Mail-out surveys are particularly at a disadvantage when using network questions that are open-ended.

Some studies have argued that low response rates in surveys are due less to potential respondents' resistance to participation and more to the researcher's inability to simply find and interview respondents (Sosdian and Sharp, 1980). These issues have become increasingly problematic for methods such as phone surveys, where people may have been overwhelmed by telemarketers, donation solicitations and political canvassing. In addition, the use of mobile phones is creating new challenges to the valid use of phone interviews and surveys. With technologies such as caller ID, people can now monitor calls and choose not to answer the phone. Johnson (1990), for example, found that using respondents to

Table 4.1 Features of different survey types.

Type of data collection	Issues of sensitivity	Interviewer response effects	Data handling errors	Cost of administering	Ability to establish rapport	Ability to maximize elicitation
Face-to-face	High	High	Moderate	High	High	High
Self-administered	Low	Low	Moderate	Moderate	Low	Low
Mail-out	Low	Low	Moderate	Low	Low	Low
Electronic	Low	Low	Low	Low	Low	Low
Phone	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
Group setting	High	Moderate	Moderate	High	Moderate	Moderate

call ahead to their listed alters (who would be more likely to answer phone calls from friends) in a snowball sample limited this problem, particularly in a population of older adults in a small Midwestern town who were suspicious of strangers (Johnson and Griffith, 2010a).

Comparative research has shown that the different survey approaches vary in response or return rates. Such studies have also found that differences in response rates can vary depending on the social, organizational or cultural context. For example, in a comparison of different survey approaches, whereas mail-out surveys win out in one context, they may just as readily lose out to other methods, such as online surveys, in another. The point here is that the data collection method you choose should be sensitive to the given cultural and social context in which you plan to work (Church, 2001). In addition, we advocate making as much contact with potential respondents as possible independent of the type of data collection approach. In fact, the more you can engage in ethnographic on-the-ground efforts, the better your chances for maximizing response rates in network surveys. In the polar research example (Johnson et al., 2003), two of the investigators spent months training and deploying with winter-over crews. This enabled them to build rapport so as to maximize the chances of study participation over the austral winter, a period when members of the research team were not present and monthly questionnaires were distributed by the station physician.

4.6 Archival data collection

In order for data collected from archival sources to be of use in the study of social networks, it must contain information on social relations that are amenable to either a one-mode or two-mode network format. Some archival sources are inherently relational and very structured, such as church marriage records, records of business partnerships, legislative voting records, ledger sheets, and accounts of trades. In such cases, ties may be readily determined among and between social entities such as individuals, firms, families, tribes, and businesses (one-mode). Or they can be inferred indirectly through co-occurrences, such as overlaps in voting behavior among members of Parliament at the start of World War II, the co-occurrence of patrons among seventeenth-century scientists in Italy, or co-attendance at early twentieth-century political rallies or events in New York (two-mode). Additionally, the nature and structure of the archival data frames just which network relations a study can use. If you are interested in economic exchange among villagers in Tuscany in the sixteenth century, but all that exists are marriage records, then the relational data available are not suitable for your research problem.

Relational data in archival sources can also be extracted from less structured historical sources. Many accessible historical records may not be as well structured as in the examples above and may be freer flowing, as in the form of a narrative. If these narratives – such as letters between luminaries of some historical period – mention names, events, locations, etc., then it is possible to build a social network database by coding the narratives. For example, in a series of letters from Galileo's daughter, Maria Celeste, to her father, there are many mentions of people and places that can be used to piece together social relations among actors of the time. In the following excerpt from Maria's letter of 10 August 1623, there is clear reference to an exchange of letters between Galileo and the new Pope:[1]

> The happiness I derived from the gift of the letters you sent me, Sire, written to you by that most distinguished Cardinal, now elevated to the exalted position of Supreme Pontiff, was ineffable, for his letters so clearly express the affection he has for you, and also show how highly he values your abilities. I have read and reread them, savoring them in private, and I return them to you, as you insist, without having shown them to anyone else except Suor Arcangela, who has joined me in drawing the utmost joy from seeing how much our father is favored by persons of such caliber. May it please the Lord to grant you the robust health you will need to fulfill your desire to visit His Holiness, so that you can be even more greatly esteemed by him; and, seeing how many promises he makes you in his letters, we can entertain the hope that the Pope will readily grant you some sort of assistance for our brother.

This excerpt clearly reveals relationships among Galileo's family as well as other relationships, even providing information on the possible strength of Galileo's relationship to the Pope. The coding of relations from the 124 letters Maria wrote to her father might describe much about aspects of Galileo's familial and political networks from the period 1623–1633.

Another example comes from the same Galileo Project database. Relations among scientists and the structure of the scientific community of Galileo's time can be derived using archival sources that include scientists' university attendance, scientific disciplinary training, patronage (often a major source of support for an academic of the time), correspondence among scientists, and membership in scientific societies. From these sources, two-mode data can be constructed from patronage (scientist-by-patron), university attendance (scientist-by-university) and scientific societies (scientist-by-scientific societies). One-mode data can be derived from the correspondence among scientists – the equivalent of emails today. This can be coupled with attribute data in records and narratives, such as

[1] Galileo Project: http://galileo.rice.edu/fam/letters/10aug1623.html.

date of birth and death, father's status and occupation, nationality, aspects of education, religion, and means of support (e.g., inherited wealth), to test any number of hypotheses about power or the dominance of scientific thought in Galileo's time.[2]

There are a number of classic studies that have extracted social network data from archival sources. In a study of social change, Bearman (1993) looked at local elite social networks in Norfolk, England, between 1540 and 1640. The network data for the study were derived from various archival records on kinship relations over that time period and were related to various attributes such as status (i.e., class of gentry), occupation and religion. Similarly, Padgett and Ansell (1993) coded data from a major historical work on social dynamics in fifteenth-century Florence, with a particular focus on the rise of the Medici (Kent, 1978), to build a multiplex network dataset (intermarriage ties, business ties, joint ownerships, partnerships, bank employment, real estate ties, patronage, personal loans, friendships, and what they call 'surety ties' – actors who put up bond for someone in exile). Attribute data were also coded from the various historical accounts and included economic wealth obtained from tax records (*catasto*), a family status measure based on 'date of first Prior (a monastic superior)', neighborhood residence, and tax assessments for the 600 richest households in Florence in 1403. In these two examples the authors were able to build datasets that included dynamic networks involving multiple relations and modes (both one- and two-mode) and a variety of attributes that could be used to test hypotheses.

One of the real advantages of archival sources for the study of social networks is that archival data are often longitudinal in nature. This allows for the study of network dynamics and evolution and facilitates the study of social change. Longitudinal research that involves the collection of primary data must use a prospective design in which data are collected periodically over some time period. This can be very costly and time-consuming. Imagine having to prospectively collect network data over a 100-year period as in the Bearman example above (1540–1640).

It is important to make one final comment about the validity of archival sources. As discussed in the chapter on research design, secondary or archival sources can suffer from a number of reliability and validity issues. Archival records can document non-events (e.g., Congressional Record) or represent a reconstruction of the past or an event to meet some agenda where narratives or numbers are constructed to make a group or a single actor look good in light of poor outcomes. Often these records include elements of scapegoating and false attribution. Thus, records may

[2] http://galileo.rice.edu/Catalog/Docs/categories.html.

be biased in that they are constructed to fit some agenda or reflect actor biases (e.g., state-owned newspaper reports). For example, in the South Pole research Johnson reviewed 15 years of managers' end-of-year reports and found them to contain inaccuracies (he compared ethnographic historical interviews with the reports). This was understandable in that these reports were an attempt to put a more positive spin on the winter events to make the station manager look good and to place the blame for any problems on others. Had these reports been used to construct some valid historical account of the social dynamics and life at the stations, any resulting conclusions drawn from an analysis would have had some likelihood of being wrong. It is always best to use triangulation of multiple independent sources so that the data can be verified and validated.

4.7 Data from electronic sources

The collection of data from electronic sources is very similar to the collection of network data in archival or historical research. Many sites on the Internet contain information that is inherently network-oriented. There is a large amount of existing data on – or data that can be mined from – email communications, social networking sites, movie, music and book databases, scientific citation databases, wikis, Web pages, digital news sources, and so on. Many of these already have information available in a one-mode or two-mode network format, while others require the writing of programs for data mining in order to put it into data formats that can be more readily analyzed. Twitter readily affords network data in the form of follower and followee ties, while social networking sites, such as Facebook, consist of literally millions of ego networks. Electronic sources offer almost endless opportunities to collect and analyze network data of one kind or another.

An example of a useful electronic data source is the Internet Movie Database (IMDb) which has a tremendous amount of data on virtually every movie ever made. For example, it has – in machine-readable form – the cast and crew, storyline and plot summaries, news articles about the movie, trivia, quotes, references, movies that reference a given movie, company credits, technical specs and so on. Some of this information can be used to construct two-mode data matrices, such as actor-by-movie, movie-by-keyword, movie-by-news article and so on, which can then be converted into one-mode networks (see Chapters 5 and 13 on this). As an illustration of the use of IMDb data, we examined the following research question: do conservatives and liberals in Hollywood work together on films? We obtained a list of the top 20 most liberal and the top 20 most conservative actors in Hollywood from a (now defunct) website called celepolitics.com. Looking at

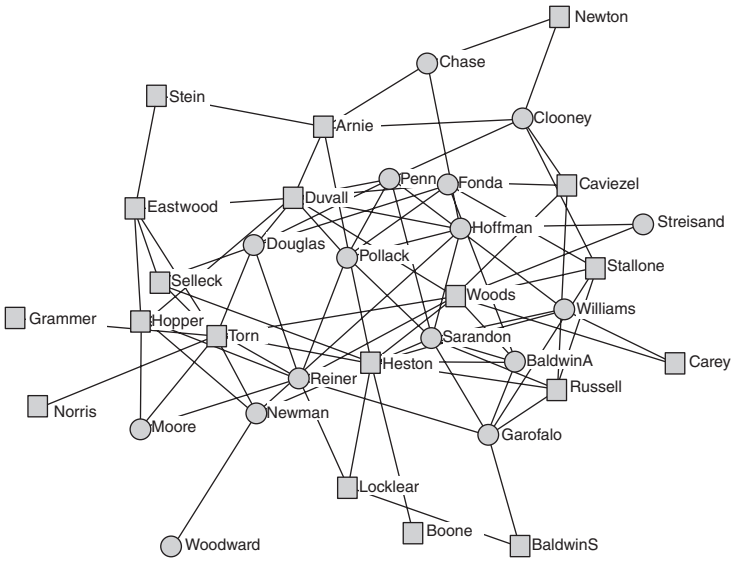


Figure 4.5 Collaboration ties among Hollywood actors. Squares are conservatives and circles are liberals.

only movies involving these 40 actors yielded a 40 × 96 two-mode network. Figure 4.5 shows the affiliation network for the 40 actors. Despite having different ideologies, at first glance it appears that conservative and liberal actors do in fact work together in Hollywood. In looking at the network visualization, however, there appears to be some segregation by ideological stance, with conservatives co-occurring more to the left and liberals more to the right of the graph.

The above analysis is just a simple example but does show how an online database can be studied from a network perspective. However, like archival data, electronic sources of data can have reliability and validity problems. The Web of Science database, for example, provides some important lessons on potential errors in electronic sources. This source contains data on citations between scientific papers which can be analyzed in terms of networks among scientists. One obvious problem with many databases of this kind is that some of the data are copied from original sources by hand or via optical character recognition software, and therefore contain errors. In addition, the sources themselves may contain errors: the authors might use different initials than in other papers, and they may misspell the names of the authors they cite. Electronic data need to be cleaned and checked just like primary data. These databases are sometimes so large that the data cleaning task can be daunting, but it is an important one nevertheless.

Although electronic sources afford almost unlimited opportunities for the collection of network data, caution should be exercised in inferring the meaning of social ties in such sources. Although Facebook ties are called 'friends', these need not correspond to the usual meaning of friends and can in fact include a wide variety of types and strengths. On the other hand, with sufficient creativity, effort and access to data it is possible to add quite a bit of richness to Facebook data. For example, we might declare a tie from A to B to be strong to the extent that A tags B in A's pictures. This can also be used to establish directionality, which is otherwise absent in the Facebook friend tie.

The same is true for micro-blogging sites such as Twitter. Although clear directed ties exist between followers and the followed, there is no direct indicator of strength of tie, and it is difficult to know what the followership ties entail. We can be sure that information is flowing from the followed to follower, at least in the case where the follower retweets a message, but can we infer an emotional bond between follower and followed? Should we expect structural hole theory to apply to the follower relation?

4.7.1 Social media

Probably the largest source of network data in recent years has been from social media sites. Unfortunately access to these data is controlled by the social media companies and they are able to determine the level of access for any researcher. As an example, Facebook until relatively recently allowed members to download a network of all their friends including ties between friends. This ceased in 2014 but a number of Facebook networks are still publicly available and can be found on the web. In addition, Facebook does still allow some access for example to Fan Page networks. Some providers such as snapchat have never allowed access whereas sites such as Reddit and even YouTube do allow data access. It would not be possible in this chapter to discuss all the potential sites that allow access so we restrict our discussion to Twitter as an example.

There are essentially three ways to get Twitter data. All these methods use an API (Application Programming Interface) that is a means by which a developer can access the Twitter data by writing a computer program using the open source APIs provided by Twitter. As this is a highly technical means to get data we shall discuss a simple-to-use program, NodeXL (available at http://www.smrfoundation.org/nodexl/), which allows the non-expert access to two of these. NodeXL is an add-on to Excel and can be downloaded and allows some functionality for free. These two methods provide access to two different sorts of data. The first is the search API. This gives potential access to all Tweets that have occurred based on a search criterion. This criterion can be a hashtag, username, location, etc. However, Twitter limits the amount of data that can be accessed in a number of

ways. For an individual user, you can only have access to the last 3,200 Tweets over the last week. For a keyword, the limit is 5,000. But the biggest restriction is that you can only make 180 queries in any 15-minute period. Hence if you wish to build a network of any size this has to be done over a period of time.

The second API in NodeXL is the streaming API. In this instance Tweets are collected in real time. In essence, rather than pulling the data down these are pushed to the researcher. Again, this is restricted by Twitter and the percentage of data received is dependent upon the volume of activity. This has been estimated at between 1% and 40% and is compounded by the fact that the sampling strategy is not published.

To obtain data simply download NodeXL and then click on the import button. Some Twitter data can be imported directly using the free basic edition but the Pro addition allows the user to download some Facebook data, Flickr and YouTube. Once in NodeXL it provides a number of features to analyze and draw the network or it can be exported for use in other network analysis packages. A good introduction to using NodeXL for social media data collection and analysis is the book by Hansen et al. (2010).

As a simple example Figure 4.6 shows the Twitter network for just 100 Tweets using the search term Brexit on 4 April 2017. The relation is 'replies to' or 'mentions' and the isolates have been deleted. Clearly this is a very fragmented network but it represents a very small sample of Tweets which used that term over the previous week.

It is possible to get all Tweets that satisfy a specified criterion in real time by using the Twitter Firehose API. Access to the Firehose is controlled by two commercial companies – GNIP and Datasift – and to gain access requires payment.

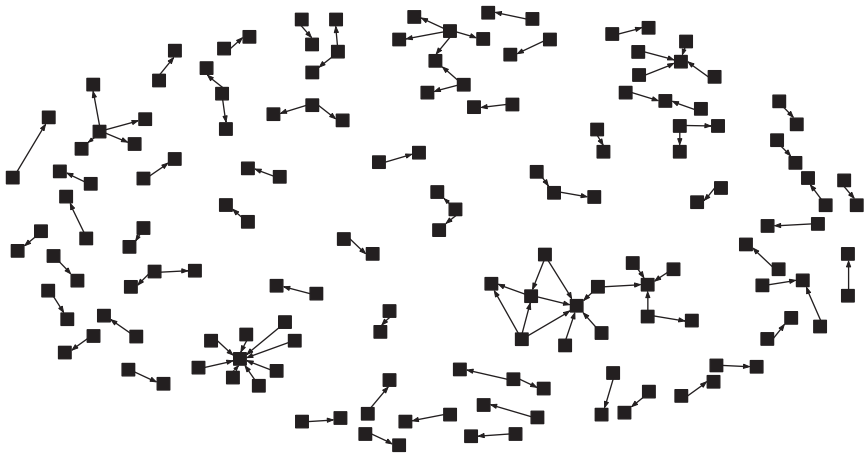


Figure 4.6 The Twitter network of just 100 tweets using the search term "Brexit" on 4 April 2017. The edges are 'replies to' or 'mentions'.

4.8 Summary

The collection of primary social network data via questionnaires or interviews is very different from that of standard survey data. Problems of recall make it difficult for a respondent to name others in their network on an unaided basis, so the free-list method should only be used when it is not possible to use a roster. It is important to use precise terms when asking respondents about any association. For example, a question such as 'who did you socialize with last month?' is preferable to a vaguer one such as 'who is your friend?'. Likert scales are less demanding on the respondent than absolute scales and therefore are often the preferred method for collecting valued data. However, they do have the drawback that different respondents can interpret the questions differently. These effects can be lessened by normalization, but they cannot be eliminated. Pre-testing questions and using ethnographic methods to help develop questions and scales will help ensure question relevance and validity. Historical sources rarely contain social network data, so associations between actors usually have to be deduced from attendance at events or meetings or inferred indirectly from narratives. Data from electronic sources do not usually suffer from these issues, since they are often already in network form. However, sources such as Twitter or Facebook present challenges of interpretation, since the connections made do not always reflect those in the offline world.

4.9 Problems and Exercises

1. For the network boundary problem (Problem 5) in Chapter 3, provide examples of questions that a researcher might ask to collect network data. In addition, discuss the advantages and disadvantages for open- versus closed-ended question formats in each case.
2. Network data can be extracted from both archival and electronic data sources. Provide an example of social network data that can be collected from each source.
3. When designing a social network survey instrument, what are some of the ways you can reduce respondent burden?
4. John, a social network researcher, is interested in the relationship between people's frequency of recreational interactions and their political attitudes. Provide examples of the kinds of questions John might ask, using both an absolute and relative format for eliciting tie frequency.
5. Jennifer studies organizational behavior and is interested in understanding the relationship between frequency of interaction among employees in a corporate headquarters with 355 employees and their attitudes about corporate culture. She wants to use a closed-ended multigrid question format in her study. Discuss the advantages and disadvantages of this approach in this case, and provide some suggestions of ways to reduce respondent burden.