Collecting and Managing Network Data

In: Social Network Analysis and Education: Theory, Methods & Applications

By: Brian V. Carolan

Pub. Date: 2013
Access Date: January 18, 2022
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412999472
Online ISBN: 9781452270104
DOI: https://dx.doi.org/10.4135/9781452270104
Print pages: 67-96

Collecting and Managing Network Data

Objectives

The previous chapter introduced the two different ways in which network data are represented (graphs and matrices) and the three different types of variables that are included in social network analyses (relational, structural, and attribute). This chapter takes a step back and addresses issues related to the collection and measurement of these network variables and how they can be stored and managed for subsequent analyses. Specifically, this chapter addresses issues related to boundary specification, sampling, measurement, collection, storage, and measurement; methodological issues that are generic across empirical social science disciplines. The key difference is that social network data are derived from relations in context; therefore, the methodological issues inherent to the collection and management of social network data are somewhat unique in several different ways. These issues are critical for a range of analytical decisions, which are addressed in the next three chapters. By the end of this chapter, you will understand how to measure, collect, and manage network data.

To highlight these issues, this chapter will rely on the Daly's Network of District School Leaders data set (“School Leaders” data set). As noted in Chapter 1, these data were collected from school leaders in two districts over 3 consecutive years. In addition to attribute data (e.g., sex, ethnicity, marital status, etc.), multiplex data on 11 different relations were collected, including collaboration, support, and expertise among others.

Boundaries

Chapter 1 noted that social networks can be analyzed at four different levels: ego, dyad, triad, and complete. Analyzing social network data at one or more of these levels, however, requires that you focus on either one of two objects of measurement: egocentric or complete-networks. Egocentric network studies have a more limited goal of describing a focal actor's (ego) social environment relative to one or more others (alters). On the other hand, complete-network studies measure the relations among actors in some bounded social group by collecting data on one or more relations among the group's actors. Complete-network studies can, therefore, subsume the three lower levels of analysis: Complete-network data include egocentric data for each actor as well as information about the network's dyads and triads. Recall, however, that most analyses of social networks are done at either the ego- or complete-network level. Given this information, the School Leaders data set can be considered an example of a complete-network study: one complete network of

school leaders from two separate school districts.

Boundary Specification

When collecting either ego- or complete-level network data, one of the first decisions you must make is to define the target population and construct a sampling frame that adequately provides access to that population of interest. On one level, this is no different from the standard sampling logic advocated in the quantitative social sciences. In studies of social networks, these decisions about populations and samples are referred to as the “boundary specification problem” (Laumann, Marsden, & Prensky, 1989).

In complete network studies, the issue of boundary specification is somewhat easier to negotiate. Formal or positional criteria typically offer a clear definition of who is a member of the group. Consider the school leaders data set; each leader who held a formal administrative position in the school district was considered to be within the network's boundary. Other examples of how formal criteria can be applied to determine the network's boundary include a classroom of students (Pittinsky & Carolan, 2008) or school staff within the same district (Spillane, Healey, & Kim, 2010).

However, boundaries in complete network studies may not be so clear cut. Delineating the network's boundary for actors that do not share some formal criteria is tricky. Examples of such groups may include students who regularly use the school library, adolescents who often socialize outside of school, or parents who run into each other and discuss school-related issues. Because there is no formal boundary demarcating who is “in” and “out” of the network, social network analysts use a few different ways to construct a bounded sample. These different means are discussed in the following section, but note that they also shape decisions regarding the collection of ego-network data.

Ego-network studies are typically conducted as part of a representative sample survey. Therefore, unlike complete-network studies, the collection of actors in an ego-network study represents some larger target population. As Marsden (2011) notes, the boundary specification of ego-network studies follows the definition of the study's target population. Once a sample of respondents (egos) has been identified, a second boundary-determination issue surfaces. This issue has to do with delimiting the set of alters within any given respondent's egocentric network. Because there is no predetermined network of potential alters as there is in complete network studies, how can social network analysts elicit information about ego's alters? This issue is discussed in the section on collection.

Strategies for Boundary Specification

Collecting network data for either egocentric or complete-network studies requires you to consider issues related to boundary specification, sampling, and measurement. This section directly addresses the first of these issues in more detail and sets up the subsequent sections on sampling and measurement. Specifying

the boundary of a network is analogous to asking: Where do you set the limits when collecting relational data when, in theory, there are no limits (Barnes, 1979; Knoke & Yang, 2008)? There are three generic approaches to addressing this issue: positional, relational, and event-based.

Positional

The first of these approaches, positional, has already been mentioned as the most common way to identify a complete-network population bounded by some common attribute (e.g., all school leaders within a district). But it can also be employed in ego-network studies in which respondents are not all drawn from the same bounded network as in the School Leaders data set, but, rather, they share some common attribute such as grade or achievement level. This positional approach generates a set of actors that occupy a similar position in some social structure. Each actor, however, need not be directly connected to every other actor.

There are a few issues to keep in mind when constructing networks through this positional approach (Knoke & Yang, 2008). First, network structures uncovered through this approach look different from one another. For example, a collaboration network among teachers in a school's English department may look very different from the teachers' collaboration network in the same school's math department. Even in a similar context, in this case the school, the teachers' collaboration networks may not exhibit the same social structure, and, therefore, making an inference from one network to another is unwise. A second issue is that relations among actors in bounded networks, particularly large ones, are likely to be sparse. Most likely there will be pockets of actors lacking connections to each other. While this may expose something interesting, the lack of overall connectivity may provide too little with which to work. A third issue when employing the positional approach is that you are obligated to justify the inclusion or exclusion of certain positions. For example, in the School Leaders data, Daly employed a nominalist criterion that included all administrative personnel in the two districts that have 1 of 14 possible administrative titles. This criterion was justified by the fact that leaders at these levels were charged with devising, implementing, or monitoring an array of school policies.

Relational

The second approach to specifying a network's boundary is relational and is most commonly employed in ego-level network studies (Knoke & Yang, 2008). This approach is either based on your knowledge about relations among a set of actors or relies on the actors themselves to nominate additional actors for inclusion. This approach leads to several different procedures that address the issue of boundary specification in slightly different ways. Five of these are introduced now and discussed in more detail later in this chapter's subsection on sampling: reputational methods, snowball sampling, fixed-list selections, expanding selection, and *K*-core methods.

The first procedure specifying a network's boundary using a relational approach is through a reputational method. This method is as straightforward as it sounds. You identify the most knowledgeable informants and asks them to list a set of actors for your study. For example, using this approach, you may investigate

the national network of charter school advocacy organizations by first compiling a master list of these organizations from archival records. You can then ask respondents—that is, people “in the know” about these organizations—to identify the small number of organizations from that list that are most important. These two steps, first compiling a list of actors and then asking key informants to rate them, can uncover the set of important organizations for future analyses. In addition, these informants can be asked to expand the master list by identifying others based on their inside knowledge.

While the benefits of this reputational approach are obvious, so too are three related shortcomings. First, it relies on key informants whose ability to provide accurate and complete information needs to be questioned. How can you be assured that the respondents can really identify those charter school advocacy organizations that are most important? In addition, there is no standard by which you can evaluate the accuracy of the boundary's specification. Finally, while you are obligated to provide a strong theoretical and empirical justification for choosing certain key informants, there is still the potential for the key informants to produce data that are systematically different than the data collected through other methods.

In particular, the data collected through reputational methods are likely to be much different than the data collected through snowball sampling, the second of the five relational approaches. Useful in finding members of hard-to-reach populations, snowball sampling begins with a small set of actors who are then asked to nominate others with whom they have a certain kind of relation. In turn, these actors nominate others, and the process goes on until few or no new names emerge. The obvious problem with this approach is the ethical concerns it raises, as securing informed consent, protecting anonymity, and assuring confidentiality become difficult as you ask for information about each ego's alters (Borgatti & Molina, 2005). Assuming that you can secure human subjects approval for this sampling strategy, it can be a useful one for educational researchers interested in studying topics such as bullying, cheating, or any behaviors that occur within “hidden” populations.

The third and fourth relational approaches can be distinguished by whether informants provide their own list of actors or you provide that list to them. For example, in fixed-list selection, a respondent is limited to reporting ties between actors that have already been identified by you (e.g., you provide students with a list of all the other students in their high school). On the contrary, in the expanding-selection approach, respondents can list as many actors as they want, which resembles snowball sampling, and therefore with it come the difficulties associated with implementing this strategy in practice (e.g., you ask the students to list the names of a fixed or undefined number of students in their school). The downside to the fixed-list selection approach is that it is subject to nonrandom sampling bias. Because it relies on your knowledge of the network's boundary, the tendency is to include only “central” actors and ignore those on the periphery who, according to your perception, may be less important.

The fifth and final relational approach to boundary specification is used infrequently but is worth noting. The *K*-core method is useful for finding subsets of connected actors within a much larger and sparser network (e.g., all school district superintendents in a given state). A *K*-core, discussed in more detail in the next chapter, is a subset of actors that has ties with at least *K* other actors. You determine the value of *K*, thereby

setting a more (high-*K*) or less (low-*K*) restrictive criterion for bounding a network. The recommendation is to first use a low *K* to establish network boundary and then increase *K* if you want to focus on a more exclusive subset of the network. This technique has superior potential for empirically locating a network's boundaries and has been made much easier by network analysis software developments. Despite these benefits, it is curious why it is not used more frequently.

Event-Based

The third broad approach to specifying a network's boundary falls under the heading of event-based strategies (Knoke & Yang, 2008). These strategies delineate a network's boundaries by selecting only those actors who participated in an “event” at a specific time and place. For example, you may be interested in observing participants at local board of education meetings, collecting relational and attribute data on only those participants who attended three consecutive meetings in the past 6 months. As opposed to positional or reputational approaches that focus on actors’ behaviors, the criterion for inclusion through this approach is an event rooted in time and place.

The success of this approach rests on your ability to select events that specifically address the questions that are motivating the study. For example, you may include an event that is insignificant or have an event (or set of events) that excludes important players. So, while you may be interested in studying how a community's interests are actualized in the context of local school district policies, the local board of education meetings may not be the right event to uncover these relational dynamics. Important players may be absent, or the actual “action” occurs outside the formal setting of the meeting. Therefore, it is unwise to focus on a single event and better practice to observe multiple events (e.g., parent–teacher organization meetings, nonpublic meetings among the district's leaders, collective bargaining sessions between employees and district leaders, etc.) that would produce a more comprehensive network. Every event, therefore, would produce a distinct network whose actors may only partially overlap with those attending other events. It is possible to then collapse participants across all events, thereby producing a more representative network that can address the study's guiding question(s).

Network Sampling

In light of these three approaches to specifying a network's boundaries, there are a few things to keep in mind when constructing samples for analysis at either the ego- or complete-network levels. The most significant distinction to be mindful of is whether a study is motivated by an interest in complete or egonetwork-level data. Complete network studies collect attribute and relational data on all actors in a population. As discussed in the previous section, there are few ways to delineate the complete network's boundary, most likely relying on a positional approach that has clear-cut criteria for inclusion. While these types of designs can produce rich accounts of both static and dynamic network properties, they are limited by their ability to make inferences to other populations of actors.

However, ego-level network data can be used to make inferences about larger populations. In these instances, ego-level network data are drawn from a larger target population, often by a sample survey instrument. Ideally, these egocentric-network data are representative of this larger population, thereby permitting statistical inferences to be drawn.

Frank (1981) describes several sampling schemes through which such inferences can be made. One approach relies on drawing a probability sample of actors using, for example, simple random selection. Then only those relationships among this sampled subset of actors are observed. A second means through which ego-level network data can be used to make statistical inferences is by drawing a probabilistic sample of ego actors and then observing all relationships incident on those actors. A final way is through snowball sampling (also known as link tracing). These three different means are most closely aligned with the reputational approach discussed above.

Marsden (2011) notes that different inferences about network properties are available through different sampling designs. Therefore, you must first consider what properties of a network are of interest when designing a social network sampling scheme. Once this decision has been made, you can then move on to issues involving the collection of relational data.

Network Data Collection

The collection of network data—specifically, relational data on a set of actors—is primarily informed by whether you are interested in a complete–or ego–level of analyses. Because ego-level analyses can be extracted from complete network data, this section first discusses the common strategies and instruments through which complete network data are collected. It then turns more specifically to ego-level data collection, with a special focus on name-generator instruments that produce extensive network data and are most likely shaped by a reputational approach to boundary specification. The section then concludes by briefly discussing a third type of network data collection referred to as partial network data.

In addition, it will be helpful to introduce some new notation that will be used to discuss the different strategies and instruments used for network data collection. Specifically, actors will be denoted as *i* or *j*, where *i* ? *j*. In the language of graph theory introduced in the previous chapter, *i* and *j* represent a pair of nodes in graph X. The relationship between *i* and *j* is represented as *a*ij, meaning that *a* can take a value ? 0. To further simplify this discussion, only single-mode networks, such as the School Leaders data set, are considered, in which the rows and columns of the data matrix consist of the same set of actors.

Complete-Network Data Collection

In complete-network studies, you select the set of actors that serves as the study's bounded population. Informed by either positional or event-based approaches to network boundary specification, these studies typically measure a small number of relations between each pair of the network's actors. There are two

primary sources of complete-network data: census and archives. Both, however, require information about the relations between actors within the network's boundary. With archives, these relations can be inferred from the documentary evidence, such as an organizational chart of who reports to whom (see, e.g., Carolan, 2008b). The most popular technique for gathering relational data from a census is doing so through a sociometric survey instrument: a tool that requires respondents to provide information about their relations with others. This section also revisits a special variant of compete network data collection, cognitive social structure (CSS), in which respondents are asked about their perceptions of the relationships among alters.

Network data collection through a census is the simplest manner through which relational data are collected on a well-defined population of interest. This method of data collection consists of gathering relational data from all (or almost all) actors in a population. Actors who either occupy certain positions (e.g., students in a classroom) or participate in certain events (e.g., attendance at student government meetings) represent common frames through which participants are sampled. This type of network data collection is preferred when you can enumerate all actors, *i* and *j*, of the network such as teachers within a school, school leaders within a district, or school-aged children in a delineated neighborhood (Valente, 2010). Because of advances in computing power, researchers can now study census networks that consist of thousands and even millions of actors.

Census

Census data collection requires that you obtain a master list of actors bounded by some context and then ask each actor, *i*, to rate various types of relations with all or some limited number of actors, *j*. These types of data provide a complete snapshot of the entire network so that it is possible to examine how attitudes or behaviors move across the network, assuming you have collected this information over time. This can be important for studies that seek to understand how network dynamics influence individual outcomes, including students' engagement or resistance.

To elicit complete network data, you may use a nomination (reputational) or roster (positional) method. For example, using the positional method, you may give a student a roster of all students in his or her grade and have the student rate the frequency with which he or she discusses school matters with every other student (this would be a directed and valued relation). Or, using the nomination method, you would ask the student to write down the names of those that the student discusses school matters with.

There are advantages to each method. Valente (2010) identifies six advantages of the nomination method: (1) It is an unassisted recall; (2) it is less demanding on the respondent; (3) the rank order in which a respondent lists alters can be preserved and used as a proxy for tie strength; (4) the number of alters that each respondent can list can be adjusted; (5) data entry and management are easier (discussed later in this chapter); and (6) it is more likely to identify actors outside of the network's imposed boundary. However, there are also advantages of the positional method. Namely, weak and strong ties are measured and the network's boundary is unambiguous, as everyone on the list is in the network. Often, however, researchers combine elements of both approaches. One common way to do this is to use a nomination format but also provide a

roster to which respondents can refer. This technique combines the strengths of both methods.

Archives

Complete network data can also be gathered through archival sources (Valente, 2010). The sources of such data are varied and include diaries, public records, email, participation in social networking platforms, and school transcripts. Using these sources, you need not directly ask, or cannot even ask, respondents to report on their network. Rather, the network is reconstructed from these records and a number of different relations can be captured. For example, using the public Facebook profiles of a set of students in the same school, it is possible to construct a network of who bullies whom, so long as you can precisely define how bullying is measured. These critical issues of measurement are discussed later in this chapter.

For example, those who have worked in the field of bibliometrics—the study or measurement of text—have been able to reveal the structure and dynamics of scholarly networks, often by examining reference lists from publications. These efforts construct networks of who has cited whom or who has collaborated with whom. A similar technique was employed by Carolan (2008) is his study of one prominent educational research journal. Using its electronic database, Carolan was able to create a large two-mode network of readers and articles that captured which readers read what articles. This information was then used to partition the network into sets of distinct communities that were surprisingly well–connected. As digital archival sources become more accessible, they represent a wonderful opportunity to yield insights into actors' behaviors and attitudes. For example, with the proliferation of online learning venues, you can easily imagine using the digital archives from these sources to examine an array of relational dynamics among students.

Sociometric Instrument

However, a majority of complete network studies require that you employ some type of sociometric instrument to capture relational data on some clearly delineated population of actors. Unlike network data that are harvested through archival sources, these types of instruments require that each *i* actor within a network report on some relationship, *a*, with some set of alters, *j*. This produces a value of a_{ij}, which represents *i*'s choice or nonchoice of *j*. Pioneered by Moreno, these sociometric "tests" measure relationships between pairs of actors, including relations such as support, friendship, and collaboration. They can also be adapted to collect two-mode network data in which *i* and *j* are not the same; in this case, respondents, *i*, would be asked to report on memberships or affiliations in or with *j*. For example, students could be asked to identify the extracurricular activities in which they regularly participate. Sociometric items—that is, items that elicit relational data—are typically administered through a standard survey instrument, either in person or self-administered. More frequently, however, this information is collected through computer-assisted means, which can simplify presentation and ease data-management concerns.

Social network analysts use different criteria to extract information about a respondent's sociometric choices. These criteria are primarily determined by both theoretical interest and the substantive questions they pose to respondents. In the School Leaders data set, Daly asked respondents to report on a number (11) of

different relations, thereby inducing what is referred to as multiplex network data. Asking school and district leaders, for example, to state how often they collaborated with other members of the administrative team and the frequency with which they discussed matters of a confidential nature with these other members helped generate Daly's sociometric data.

Most sociometric surveys, like Daly's, supply a list of possible alters in the network for respondents to consult; others, however, permit respondents to freely recall their ties to alters from memory. While both methods are acceptable, there are two reasons why it is recommended that analysts provide respondents with a list of possible alters (Marsden, 2011). First, providing a roster to respondents eases the respondents' reporting burden by reminding them of the network's eligible alters. In addition, a roster minimizes measurement error, as respondents are often prone to forget potential alters. But there are tradeoffs you must consider. For instance, asking a respondent to review and evaluate all the alters on a large roster is tedious and time consuming, even if assisted by a computer interface. Whatever technique you choose, however, the analyst must exercise caution with alters' names. Recall methods must somehow make sure that respondents' alters known by different names (e.g., changes in surnames due to marriage, nicknames, etc.) are correctly matched. Along these lines, rosters that provide respondents with a list of potential alters must use those names by which the alters are actually known.

Researchers have offered conflicting guidance as to how to best construct sociometric data-collection instruments. These different suggestions center on the number of alters that respondents are allowed to choose. Early guidelines suggested that respondents be given the opportunity to make an unlimited number of nominations, while others have suggested that the number of nominations be limited to three or four. Limiting the number of alters is currently the more widely exercised option; when administering a survey, it is simply more practical to specify a sociometric task for respondents and make things as straightforward as possible, which eases the respondents' burden. However, in doing so, you may also be increasing measurement error. Imposing a limit on the number of alters a respondent may nominate potentially induces what Marsden (2011) refers to as a false negative (the respondent's actual number of alters is higher than the imposed limit) or a false positive (the respondent's actual number of alters is lower than the imposed limit). The conclusion to be drawn from this is that, similar to standard survey-based research, bias is generated in many sociometric tasks, and it is your responsibility to mitigate these adverse effects as much as possible. These measurement issues are discussed later in this chapter.

Sociometric tasks capture relational data through a number of different possible response categories or formats. The simplest and one of the most widely used is binary measurement. Respondents identify those alters with whom they have a given relationship by making a separate yes/no (1/0) distinction for each of them. As noted in the previous chapter, binary measures such as this result in a sociomatrix that consists of a_{ij}, which represents i's choice ($a = 1$) or nonchoice of j ($a = 0$). This is the technique used by Penuel, Frank, and Krause (2010), in which teachers were asked to list their closest professional in-school colleagues, and this information was then used to identify subgroups within the network (these techniques to identify subgroups are discussed in the following chapter). Other sociometric tasks, however, require an ordinal

response. For example, in the School Leaders data set, Daly asked respondents to select the number of times they engaged in different types of relations from four different response categories: 1 (the least frequent) to 4 (1–2 times a week). Therefore, in the relationship between *i* and *j*, represented as a_{ij}, *a* can take a value of 1, 2, 3, or 4. These values become the cells of the single-mode sociomatrix. Finally, sociometric responses can be also be ranked. Newcomb's Fraternity Data illustrates this technique. Newcomb asked all 17 members of the fraternity to rank all other members from 1 to 16, with 1 representing the first friendship preference, and no ties were allowed. From this, it follows that in the relationship between any two members, represented as a_{ij}, *a* can take a value of 1, 2, … 16. The sociomatrix that reflects these relational data consists of cells ranging in value from 1 to 16 and a main diagonal consisting of 0s (one could not rank the relationship with oneself).

Table 4.1, adapted from Marsden (2011), provides several examples of the questions that researchers have asked to elicit sociometric data on complete networks. It is evident from this table that sociometric choices can be generated through many different means, but the means that you choose to employ are dictated by substantive issues in which you are interested. In addition, these questions do not reconcile the outstanding issue as to whether you should provide respondents with a list of possible alters or rely on respondent's free recall.

Table 4.1 Examples of Complete Network Sociometric Questions.

A. Single-criterion recognition question (Pittinsky & Carolan, 2008)

To collect student reports of friendships, students were given a class roster and asked to describe their relationship with each student in the class. Choices included best friend, friend, know-like, know, know-dislike, strongly dislike, and do not know (1–6, with 0 = don't know).

B. Multiple-criteria recognition questions (Moolenaar, Daly, & Sleegers, 2011)

"Whom do you go to for work-related advice?"

"Whom do you go to for guidance on more personal matters?"

A school-specific appendix was attached to each survey, which included the names of all the school's team members and a corresponding letter combination (e.g., Mrs. Erin Smith = AB). Respondents (teachers and principals) could answer the social network questions by indicating the letter combination of the intended colleague(s) (e.g., Mrs. Erin Smith = AB), and they could name all the colleagues with whom they interacted.

C. Free-recall questions (Cairns, Leung, Buchanan, & Cairns, 1995)

The interviewer asked, "Some people have a number of close friends, but others have just one 'best friend,' and still others don't have a best friend. What about you?" Children (in the fourth or seventh grades) were free to nominate any number of friends. No class lists or photographs were provided to assist the children: the method involved the free recall of persons. Similarly, children were free to name persons outside the classroom and outside the school.

D. Cognitive social structure task (Gest, 2006)

The teacher was asked, "Are there some children in your class who hang around together a lot? Please use the boxes below to list the names of the children in each group. For very large groups, you can continue listing names in the next box and indicate you are describing one large group." Nine boxes were provided, with lines for six names within each box. At the bottom of the page, teachers were asked, "Are there any children in your class who do not have a group? If so, please list."

Source: Marsden, P. V. (2011). Survey methods for network data. In J. Scott & P. J. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 370-388). Thousand Oaks, CA: Sage Publications.

Cognitive Social Structures (CSS)

Researchers (e.g., Krackhardt, 1987a) have long advocated that the role of cognition, as opposed to plain structure, be more strongly considered in social network analysis (de Lima, 2010). While de Lima (2010) considers this a third level of analysis, the fact is that it relies on complete-network data, so it will be treated as such in this section. CSS emphasize that actors' perceptions of relational patterns have implications for their own individual attitudes and actions. These implications outweigh the structural reality in which the actors are embedded. To identify these CSS, researchers rely on self-report data that reflect an actor's perception of who is linked to whom and in what ways. Rather than focus on an actor's self-report of his of her own relations with others, the analytic focus is shifted toward asking respondents about how they perceive the relations between any two alters. In these instances, the respondent acts as an informant about the social ties between pairs of actors. For example, in the Peer Groups data set referred to in the previous chapter, the teacher was

asked to rate the friendship between each pair of students on a six-point scale (1 = best friend, 2 = friend, 3 = know-like, 4 = know, 5 = know-dislike, 6 = strongly dislike). While this approach yields interesting insights, the obvious drawback is that it places a significant burden on the respondent. In addition to being time consuming and memory intensive, it is also limited to networks that are relatively small in size, as collecting these data on even moderate-sized networks is not feasible.

Egocentric Network Data Collection

While studies that rely on complete-network data may resonate more clearly with what you may typically think of as social network analysis, many network studies operate at a different analytical level. Egocentric network analysis, also referred to as local network analysis, views a social network as a particular actor's set of connections (de Lima, 2010). Oftentimes these types of data are collected as part of a representative sample survey and are therefore drawn from some target population (Marsden, 2011). There are two advantages associated with working at this analytical level (Butts, 2008). First, the data requirements are fairly modest; you do not need to collect relational data from a complete network, which often results in thorny problems with missing data (discussed below). Second, egocentric data collection can be incorporated into large-scale survey research and potentially be used to make inferences about its target population. Therefore, within this framework, researchers usually examine a sample of egocentric networks within a population. The key assumption, and one that is very reasonable if selected from a large enough population, is that each ego's network is independent from those of other respondents.

While egocentric network data are typically part of large-scale surveys, these types of data differ from the simple attribute data that are elicited through standard survey efforts. Egocentric data-collection instruments collect data on the characteristics of the respondent's (ego) alters. In addition to collecting information on the relationship between the respondent and alters, egocentric network data-collection instruments also sometimes go further by asking respondents to provide information on the relations between alters. Valente (2010) notes that this additional information is often useful in order to examine the relationship between personal network characteristics and the possible influence on one's behavior and attitudes. Marsden (2011) delineates the four different ways in which these egocentric network data are collected, often as part of a standard survey instrument: name generators, position generators, resource generators, and social support scales.

Name Generators

These types of questions are the most common way in which egocentric relational data are elicited. Name generators are questions, most often employed in a free-recall format, that ask respondents themselves to identify members of their network. These types of questions, therefore, require that a “focal” ego generate a list of his or her alters. They are referred to as name generators for the simple reason that they elicit a roster of alters within an egocentric network and, in doing so, establish the ego network's boundaries.

These types of questions share many of the same characteristics as sociometric items used in complete

network studies. However, one key difference is that name generators rely exclusively on respondent recall, as a roster of possible alters is typically not made available. Like sociometric items, name-generator questions focus on a particular type of relationship. These relationships can be based on role, relational content, or specific types of exchanges. For example, the name generator used in the Educational Longitudinal Survey of 2002 (ELS: 02; Ingels, Pratt, Rogers, Siegel, & Stutts, 2004) asked respondents (10th-grade students) to identify alters based on role; specifically, the prompt asked respondents to list their three closest friends in school. Other name generators focus on the relational content or resource exchange, asking respondents, for example, to identify those that they turn to for advice or with whom they discuss family, home, and personal issues (e.g., Cole & Weinbaum, 2010).

In addition to specifying a particular type of relationship, there are several other issues you should consider when collecting egocentric data through name generators. One issue is the number of probes you should use in order to induce additional alters. This can substantially increase network size as the respondents list additional alters that may have been previously forgotten. Marsden (2011) recommends that these probes be used judiciously, as respondents may interpret their use as an expectation that they are to name more alters, leading them to bend the definition of the generator's role or relational content. A second issue is whether to use more than one generator to identify an egocentric network. For example, you could ask about different kinds of relational content (e.g., help with schoolwork, frequency of contact, etc.), which could possibly produce different sets of alters and, therefore, different networks (Burt, 1997). In addition, when using multiple name generators, you should be aware of how the ordering of the name-generator questions affects the number of alters given in response. A final issue to keep in mind when collecting egocentric data through name generators is whether to provide a fixed number of possible alters. Whereas the ELS: 02 limited the possible number of alters to three and Cole and Weinbaum (2010) in their study of teachers' attitude formation toward reform limited their number of alters to five, many other egocentric instruments do not impose a limitation.

After a list of alters has been produced through one or more name-generator questions (and possibly probes, too), name interpreters are then used to gather information about those named alters and ego's relations with them. Because of this second step, egocentric data collection must be mindful of respondents' time and efforts, as providing this information asks much of respondents. Marsden (2011) identifies three types of name-interpreter items. The first type of name-interpreter questions ask the respondent to provide information about the alters' attributes, including basic demographic characteristics such as race/ethnicity and gender, or other interesting characteristics such as the importance that the alter attaches to grades in school. A second type of name-interpreter item asks about ego's relations with alters. These types of questions are likely the focus of the study and can include questions about the duration, frequency, or intensity of ego's relationship with each alter. For example, after generating a list of friends through a name generator, a name interpreter could be used to ask the respondent to rate the intensity of his or her friendship with each alter. Or the name interpreter could be used to ask the respondent to state the frequency with which the respondent discusses school-related matters with that alter. The third and final type of name interpreter poses the biggest burden on the respondent but yields valuable information about the egocentric network structure. These types of

questions follow the preceding two types and ask about the relationships among the alters themselves. So these questions, for example, would ask the respondent to rate the intensity of the friendship between each pair of alters or the frequency with which the two alters discuss school-related matters.

Egocentric data collected through name generators and interpreters can be done through a variety of means (Marsden, 2011). These means include in-person interviews, telephone interviews, or written questionnaire instruments, the standard means through which quantitative social science data are collected. Because name-generator instruments, especially those with multiple name generators, can become difficult for respondents to follow, one of the more promising avenues of egocentric data collection involves self-administered and adaptive computer interfaces. Whether administered online or on a stand-alone computer (e.g., Maroulis & Gomez, 2008), this technique often results in a more streamlined instrument that is easier for respondents to complete. Because these are likely to be self-administered, this also promotes higher levels of disclosure and data quality. But the absence of an interviewer may also reduce motivation and respondents' attentiveness. However, if opting to employ a self-administered egocentric data-collection instrument through a computer interface, it is critically important to be mindful of the interface's visual design. For instance, something as trivial as the amount of space provided to the respondent will influence the content of the response. Table 4.2, adapted from Marsden (2011), provides examples of both name generators and interpreters.

Table 4.2 Examples of Name Generators and Interpreters Used to Elicit Egocentric Network Data.

A. Single name generators

1. "Please write down the names of your best friends at your present school. Please fill in up to three names. If you have fewer close friends, provide less than three names." Respondents asked to provide first name and last initial of each named friend (from 2002 Educational Longitudinal Study [ELS: 02], Ingels et al., 2004).

   Interpreters (Alter attributes and Ego-Alter ties)

   1. Is this friend male or female?
   2. How important is getting good grades to this friend?
   3. Do you know either or both of this friend's parents?
   4. Does your mother or father know either or both of this friend's parents?

2. In the friendship section of the Add Health in-school questionnaire, the respondent was asked to nominate up to five male and five female friends from the roster of all students enrolled in the respondent's school and in the sister school. Once friends were nominated, the respondent entered each friend's identification number on the questionnaire (from The National Longitudinal Study of Adolescent Health [Add Health], Harris, et al. 2009).

   Interpreters (Ego-Alter ties)

   1. Did you go to this friend's house in the last 7 days?
   2. Did you go with this friend after school to hang out or go somewhere in the last 7 days?
   3. Did you talk with this friend about a problem in the last 7 days?
   4. Did you spend time with this friend last weekend?

B. Multiple name generator (Kogovšek & Ferligoj, 2005)

1. From time to time, people borrow something from other people, for instance, a piece of equipment, or ask for help with small jobs in or around the house. Who are the people you usually ask for this kind of help (material support)?

   From time to time, people ask other people for advice when a major change occurs in their life, for instance, changing jobs or a rather serious accident. Who are the people you usually ask for advice when such a major change occurs in your life (informational support)?

   From time to time, people socialize with other people, for instance, they visit each other, go together on a trip or to a dinner. Who are the people with whom you usually do these things (social companionship)?

   Interpreters (Ego-Alter ties)

   1. Indicate how frequently you have contact with this person. Frequency of contact is measured in five or six ordered categories, with possible responses ranging from "rarely" to "more than once a week." This is asked for each named alter.
   2. Indicate whether each friend named was an acquaintance (coded 1), a good friend (2), or a very close friend (3). This was also asked for each alter.

Source: Marsden, P. V. (2011). Survey methods for network data. In J. Scott & P. J. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 370-388). Thousand Oaks, CA: Sage Publications.

Position Generators

Another means through which egocentric network data are collected is through position generators (Lin, Fu, & Hsung, 2001), which share several characteristics with the more frequently used name generators. Developed out of the literature on social capital, position generators, however, focus on the collection of relational data about actors' resources by asking ego to report on whether they have contacts (alters) in certain social positions (Knoke & Yang, 2008). Investigating an ego's personal contacts with alters that occupy those positions reveals information about the types of social resources to which they have access and, just as importantly, how they gain access to those resources.

The success of position generators to capture an ego's access to a wide array of social resources is dependent on your choice of positions. For example, the position generator in Table 4.3 asks respondents to indicate whether they have a certain type of relationship with any alter in a specific position. These positions are predetermined by you. This is an important step that should be carefully considered and cover the range of variation within the dimensions that are of interest; for example, if examining relations among actors in certain occupations within a school district, you might consider listing all relevant occupations as they vary by prestige. Ego is then asked a series of follow-up questions on the strength of the relationship with each position with which ego has identified having a relationship.

Position generators have three advantages. First, responses can be collapsed into summary measures that reflect the composition and range of an egocentric network. Common summary measures include extensivity, upper reachability, and range (Lin, Fu, & Hsung, 2001). In addition, the position-generator instrument is efficient, requiring less interview time than the name generators described earlier (Marsden, 2011). However, as the number of positions listed by you increases along with the follow-up questions, the demands on the respondent also increase. The advice offered by Van der Gaag, Snijders, and Flap (2004) is to keep the number of positions made available respondents around 15 to 20. The final advantage of position generators is that they capture respondents' access to those occupying different positions and, therefore, their access to varying social resources.

Table 4.3 Examples of Position Generators Used to Elicit Egocentric Network Data. Developed out of the literature on social capital, position generators focus on the collection of relational data about actors' resources by asking each ego to report on whether they have contacts (alters) in certain social positions. This is a promising but underutilized technique that should be considered by educational researchers operating from a social capital framework.

A. Do you know anyone well enough to talk to in each of the occupations?
   1. Physician
   2. Elementary school teacher
   3. Police officer

(etc.)

Twelve occupations were presented to respondents in random order. Two measures of individual social capital were derived from the position generator. Reach is the number of occupations in which participants said they knew someone in the three highest-ranked positions. Diversity is the number of different occupations in which participants said they knew people (Johnson, 2010).

B. Do you know anyone who is a . . .
   1. Hairdresser
   2. Cook
   3. Nurse

(etc.)

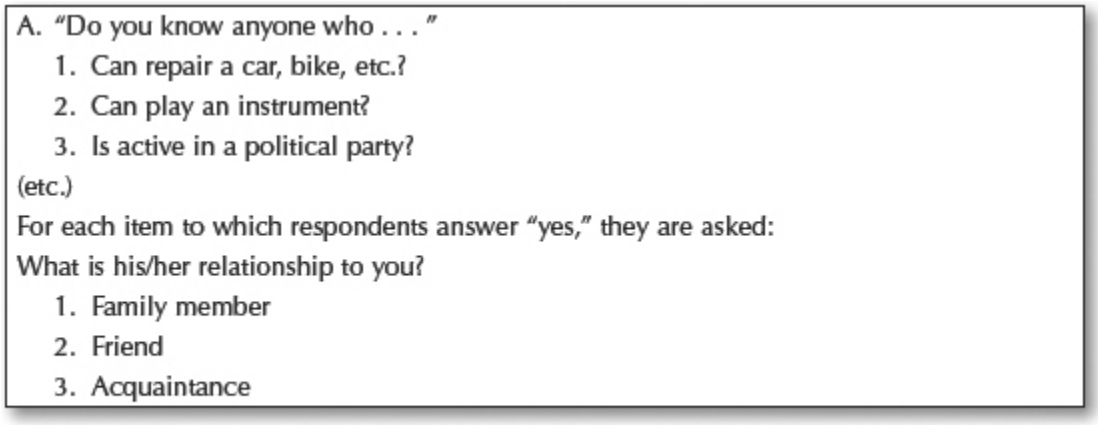
Thirty occupations were presented. As a minimum criterion of "knowing" a person who could give access to each of them was asked to imagine that when accidentally met on the street, he or she would know the name of that person, and both could start a conversation with each other. Then the respondent was asked whether he or she knew an acquaintance, a friend, or a family member in that position. Assuming an order of increasing tie strength, answers are coded into four categories: no person at all (0), an acquaintance (1), a friend (2), a family member (3); the interpretation of the distinction between these answer categories to label the relationship was left up to the respondent (Van Der Gaag, & Snijders, 2005).

Source: Marsden, P. V. (2011). Survey methods for network data. In J. Scott & P. J. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 370-388). Thousand Oaks, CA: Sage Publications.

Resource Generators

A third means through which egocentric network data are collected is through resource generators (Van der Gaag & Snijders, 2005). Resource generators attend to a significant shortcoming of position generators. This shortcoming is that ego actors often receive help from alters beyond those in the positions listed by you. Therefore, unlike position generators that ask egos whether they have contact with alters in certain positions, a resource generator asks egos whether they know alters that are useful for any specific purpose. Table 4.4 provides some examples from resource generators. It is evident from these examples that, like the position generator, the resource generator does not list an ego's alters individually or measure egocentric network structure (Marsden, 2011). Rather, the focus is on the resources that comprise one's network.

Table 4.4 Examples of Resource Generators Used to Elicit Egocentric Network Data. Resource generators attend to a significant shortcoming of position generators. This shortcoming is that ego actors often receive help from alters beyond those in the positions listed by the researcher. Therefore, unlike position generators that ask egos whether they have contact with alters in certain positions, a resource generator asks egos whether they know alters that are useful for any specific purpose. However, there are few, if any, examples of their use in educational research.


A. "Do you know anyone who . . ."
1. Can repair a car, bike, etc.?
2. Can play an instrument?
3. Is active in a political party?

(etc.)

For each item to which respondents answer "yes," they are asked:

What is his/her relationship to you?
1. Family member
2. Friend
3. Acquaintance

Source: Marsden, P. V. (2011). Survey methods for network data. In J. Scott & P. J. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 370–388). Thousand Oaks, CA: Sage Publications.

Social Support Scales

A fourth means through which egocentric network data are collected is social support scales. Marsden (2011) points out that there is a vast literature on social support that details the different instruments that ask respondents to report on the support perceived to be available or the support actually received. Some of these measures look like name-generator instruments that ask respondents about the support they perceive or receive from named alters. Other measures, however, employ a format similar to resource generators and ask respondents whether they have access to others who could provide support. A final type of social support instrument asks separate questions about the various types of support available from different types of alters, including friends, family, neighbors, etc. Table 4.5 presents these three different examples of social support scales.

Partial Network Data Collection

Table 4.5 Examples of Social Support Scales Used to Elicit Egocentric Network Data. In addition to name generators/interpreters, position generators, and resource generators, these types of scales represent a fourth possible means through which egocentric data can be collected.

	Name generator instrument that associates support with named alters (Burt, 1984)
	From time to time, people borrow something from other people, for instance, a small sum of money or a piece of equipment, or ask for help with small jobs in or around the house. Who are the people you usually ask for

this kind of help (instrumental support)?

Ego access to a given type of support (from The Early Childhood Longitudinal Study of 1999, Kindergarten Class [ECLS-K], Tourangeau, Nord, Lê, Sorongon, & Najarian, 2009)

What adult do you to talk to when you need …

Someone to cheer you up?

Help with schoolwork?

Advice about important decisions?

For each type of support, respondents could select one or more of the following: parent, adult relative, adult at school, other adult, or no one.

Forms of support available from different classes of alters (Turner & Marino, 1994)

To measure ego's experience or perception of being supported by their spouse/partner, relatives, friends, and coworkers, respondents completed a shortened version of the Provisions of Social Relations Scale. Using subsets of items from this scale, Turner and Marino separately assessed the level of support each respondent experienced from these four different classes of alters.

There are occasions when network data collection cannot be easily classified as being at either the complete- or egocentric–level. In this instance, partial network data collection (also referred to as sequenced data) consists of you collecting information from all or some portion of ego's alters (Valente, 2010). This type of relational network collection addresses the boundary specification issue through a relational approach in which a respondent's named alters guide subsequent data-collection efforts. The most common type of partial network data is collected through snowball sampling. While there are several drawbacks to this type of respondent-driven approach (noted earlier in this chapter), it is a cost-effective means of contacting a hard-to-reach population (Salganick & Heckathorn, 2004).

Another way in which partial network data are collected is through a general sequenced approach. In this approach, you take a random sample of those alters that have been nominated by ego (the respondent) and collect data only from that sample. This approach has three advantages over snowball sampling (Klovdahl, et al., 1994): (1) they are less likely to end in a social dead end; (2) they provide more entry points into a community of interest; and (3) they provide better estimates of social structure, which can then be employed in inferential models. In relation to this last point, both egocentric and partial network data-collection techniques can work within a random sampling framework and, therefore, can be used to make population parameter estimates.

Quality of Relational Data

Having addressed the twin issues of boundary specification and sampling and various complete, egocentric, and partial data-collection instruments, it is now appropriate to turn to issues regarding the measurement quality of relational data. Recall from the previous chapter that social network analysis can employ three types of variables: attribute, relational, and structural. The first of these will not be discussed here, as the measurement of these types of standard social science variables is discussed widely in introductory research methods texts. In addition, this section will not address structural variables, as they are derived from relational data. Rather, the focus is on the relational variables that are the cornerstone of social network analysis. Similar to standard attribute variables, the quality of relational measures hinges on their reliability, validity, accuracy, and measurement error. Moreover, these concerns cut across the different levels of network analysis. It should be noted, however, that available studies on the quality of network data have not produced a consensus about the quality of network measures (Marsden, 2011). Therefore, the ideas offered below are more suggestive than perscriptive.

Reliability

The first threat to the quality of relational data concerns their reliability, generally defined as the extent to which a particular instrument yields a similar result every time when applied repeatedly to the same participant. Social network researchers can rely on several common reliability measures, including interobserver reliability, internal consistency reliability (which includes split-half reliability and Cronbach's alpha), and test–retest reliability (Knoke & Yang, 2008). The latter, however, is the most commonly used means through which network researchers check informant reliability. Consider the following hypothetical example. A teacher nominates all those other teachers from whom he or she seeks advice about professional matters. The retest repeats the same (or similar) request at a later time. Then a comparison between the two sets of responses reflects the teacher's reliability; a perfect correlation between them indicates high reliability. If the retest, however, yields a different set of alters, then the comparison would indicate little reliability. The important issue for a network researcher to consider is the time between the initial test and retest. Differences between the two may reflect an authentic change in respondents' networks; after all, respondents' relations with others change for a variety of legitimate reasons. Reducing the time between the test and retest, however, can mitigate this turnover. Researchers that have examined test–retest reliability in network studies include Morgan, Neal, and Carder (1997), White and Watkins (2000), and Bignami-VanAssche (2005). The general conclusion from this work is that closer ties tend to be reported more reliably than weaker ones, a point that should be kept in mind when designing a network study.

Numerous studies have examined other issues related to the reliability of relational items that constitute network studies (Marsden, 2011). Focusing on the reciprocity of ties—two different respondents nominate each other—these studies have operated on the wholly reasonable assumption that an undirected tie should be reported by both respondents. For example, Feld and Carter (2002) reported a reciprocation rate of 58% for college students who were asked to report on the issue of with whom they spend time. Gest, Farmer, Cairns, and Xie (2003), in their study of peer relations among elementary school children, report that observed interaction is higher in reciprocated pairs of children. These points lend evidence to the claim that reciprocated

relations may, in fact, be more valid.

Validity and Accuracy

Validity is the second issue you must consider when collecting and analyzing relational data, which is often correlated with informant reliability (Romney & Weller, 1984). Validity in the case of network studies refers to the extent to which a measure actually measures what it is intended to measure (Wasserman & Faust, 1994). Often, network researchers assume that they are working with valid measures, an assumption that needs to be checked. For example, in eliciting egocentric network data from students through a name generator that asks, “List the five classmates you regularly study with,” you can assume that it has face validity in the sense that this prompt will provide a set of alters related to ego through studying behavior. But the validity of this measure would need to be tested in a rigorous manner. This is what is referred to as construct validity: A concept's measure behaves as expected according to theoretical predictions (Wasserman & Faust, 1994).

Since Wassermann and Faust's (1994, p. 58) contention that “there has been very little research on the construct validity of measures of network concepts,” surprisingly few studies have sought to address this issue that was brought to the field's attention by the classic Bernard and Killworth (1977) and Bernard, Killworth, and Sailer studies (1980). The issue is whether a respondent's reported behavior reflects his or her actual behavior. For example, one may list those five classmates with whom one studies, but actual observed behavior may indicate something totally different. These behaviors can also be captured, perhaps more accurately, through diaries, logs, or preferably systematic observation. Survey reports and “behavioral” measures of interaction tend to exhibit moderate, at best, levels of agreement (Marsden, 2011). Therefore, the quality of cognitive reports of social ties obtained through surveys is questionable.

Similar to the standard survey instruments, there are ways in which social network researchers might be able to induce more valid responses. The first way is to focus on more stable relations, as opposed to ones that are time specific (Freeman & Romney, 1987). So asking a respondent to identify who one regularly studies with as opposed to asking who one studied with this past month might be the better option. A second way is to limit the number of alters one can nominate in order to counteract “expansiveness bias,” the tendency to under/overreport the number of alters in one's network (Feld & Carter, 2002). The number of options made available to the respondent in this “fixed choice” should be theoretically and empirically justified; that is, if the average size of an adolescent's peer group consists of three to five members, then it only follows that there should be a maximum of five fixed options. A third way is to check respondents’ reports (Feld & Carter, 2002). Even where it is entirely impractical to collect information from all the others that a respondent nominates, you might reduce the extent of expansiveness bias merely by explicitly asking respondents to try to answer what the others would say as well as what they themselves think and by telling respondents that some information may be confirmed with others.

Measurement Error

Related to concerns about the reliability and validity of network data is the issue of measurement error. Wasserman and Faust (1994) define this as the difference between the “true” score of a concept and the observed score of that same concept. For example, in the School Leaders data set, respondents were asked to rate the frequency with which they had certain types of relations on a scale of 1 to 4, with 4 being the most frequent (1–2 times per week). The assumption on which researchers rely is that the observed score is equivalent to the “true” score plus error. This measurement error, then, is the difference between the “true” and observed scores. The best way to reduce this error is to design the measurement instrument in ways that are most likely to result in observed scores that most closely approximate the “true” score.

There are three issues that researchers should consider in order to best address measurement error. The first is the debate over how to best elicit respondents’ recollections: free-recall versus recognition (roster). The former refers to the technique in which you ask respondents to indicate, by recall, those alters with whom they have a specific kind of relation. The latter, which is the technique employed by Daly to generate the School Leaders data set, relies on a roster that is provided to respondents. While both techniques have advantages and disadvantages, it is recommended whenever possible to use a roster technique, as it helps remind respondents about relations that otherwise would have been forgotten (de Lima, 2010; Ferligoj & Hlebec, 1999). This is especially relevant when the number of actors in a complete network study is large. The obvious tradeoff with the roster technique is that it can become quite large and is only generally useful when there are fewer than 50 possible alters (Butts, 2008).

A second issue related to measurement error is the number of alters respondents can nominate. Fixed-choice designs, those preferenced by Feld and Carter (2002), limit the number of alters that a respondent can nominate. Free-choice designs, on the other hand, do not provide respondents with any limits as to how many people they can include. Both designs have strengths and weaknesses that you must carefully consider when designing the instrument. Fixed-choice designs limit the burden on the respondent but may artificially distort the number of alters present in one's network (Holland & Leinhardt, 1973). However, there is little agreement as to which design best minimizes measurement error, so the best advice is to employ a combination of both. For example, the “discuss important matters with” name generator on the General Social Survey does not limit the number of people a respondent can nominate, but it does have a follow-up probe if a respondent lists fewer than five people.

A third issue regarding measurement error involves measuring the intensity of relationships. Here the question is one of ratings versus rankings (de Lima, 2010). Designs that require ratings ask that respondents assign a value to each of their ties with alters. These ratings reflect dimensions such as strength, frequency, or intensity and are often captured through a scale that ranges from low to high values. The School Leaders data set was constructed through an instrument that asked respondents to assess the frequency of different types of relations on a four-point scale. In contrast, ranking designs require that respondents rank order all the other actors in their network. The Peer Groups data set referenced in the previous chapter relied on this method: Students were asked to rank their friendship preference for every other student in the class.

In smaller networks, this may be possible, but rank ordering becomes problematic in larger systems ($n > 20$)

where there is a bigger burden on respondents and issues related to missing data become more apparent. In addition, ratings are easier to administer and quicker for respondents to complete. The final and perhaps most important reason is that ratings provide a more precise measure of the relationship, as the score need not be relative to others as in a rank-order design.

While ratings are the generally preferred method for capturing a relationship's intensity, it is also preferable to have scores for this intensity as opposed to a score that reflects the presence or absence of a tie. This is the difference between valued and binary relational data that was introduced in the previous chapter. Binary relational data provide you with the least information; therefore it is recommended that you collect some measure of tie strength (that is, collect “valued” relational data). Valued data, in addition to being more useful, is important, as ties that vary in strength have been shown to perform different expressive and instrumental functions in networks. Finally, when trying to induce a measure of tie strength, it is advisable to provide a clear scale to respondents. For example, rather than provide a four-point scale that simply asks respondents to rate the frequency of ties (with higher scores indicating more frequency), it is best to attach a more precise meaning to each possible value. Therefore, similar to the School Leaders data set, it is best to avoid ambiguity in the interpretation of participants’ responses by indicting what a score of 4 means. In this instance, “very frequent” is denoted as 1 to 2 times per week.

Missing Data

In addition to the measurement issues discussed above, another concern about the quality of relational data is that network analysis may be hypersensitive to missing data (Kossinets, 2006). Costenbader and Valente's (2003) study confirms this point. Using social network data from eight studies and 58 different networks, they randomly removed an increasing percentage of actors from each network and calculated, then recalculated, a number of different network measures. They concluded that measures that reflect a network's centrality are most sensitive to missing data, whereas simple measures that capture other characteristics remain fairly stable. Borgatti, Carley, and Krackhardt's (2006) analysis found similar results.

The influence of missing data is dependent on whether you are conducting a complete or egocentric network study. For example, if conducting a complete network study on a school district's teachers who have been identified through a census sample, each teacher contributes $N - 1$ pieces of information, the ties or nonties to other teachers. A response rate of, say, 75 to 85% would be high by most any social scientific standard but might not be good enough to accurately calculate most network measures. Recent studies have further investigated these effects and have concluded that it is possible to estimate networks from just parts of them (Rhodes & Keefe, 2007), but more research is needed to model networks that are robust when faced with missing data (Carley, 2004).

Because, however, egocentric network data are sampled from some larger target population, this threshold is different. Here the important consideration is whether the data are missing at random (MAR) or completely at random (MCAR). If the data are not MAR or MCAR, the problem is likely serious, as there may be some

systematic error in data collection that “missed” some part of the target sample. The question you need to consider is whether the exclusion of some set of egos due to missing data affects the representativeness of the sample. Another related problem regarding missing data in egocentric network studies is that respondents do not report ties among their alters, which prevents certain egocentric measures from being calculated. Burt (1987) reports that missing relations among alters in an egocentric network tend to be weak ones, thereby indicating a systemic bias in missing data that may affect one's analyses. The way in which network researchers calculate response rates differs according to whether you are collecting egocentric or complete network data. For a brief review of these calculations, see Knoke and Yang (2008).

Despite one's best efforts, the curse of missing data is likely to affect one's network study. There are, however, several different strategies for dealing with missing data in network studies. The six strategies that de Lima (2010) delineates provide some clues as to how researchers might mitigate the adverse effects of missing data: (1) respecification of the network's boundary; (2) imputation; (3) reconstruction; (4) dichotomization; (5) symmetrization; and (6) triangulation. While none of these can be considered “solutions” to missing data, some combination of these strategies will likely allow the analysis to proceed without inducing a tremendous amount of error into the analysis. Of course, the best solution is to prevent missing data as best as possible. In addition to good sampling techniques and appropriate instrumentation, this requires a combination of persuasion techniques, including personal letters, monetary incentives, and phone contacts. This is no different than what you must go through when implementing a standard survey.

Managing Relational Data

The management of relational data is determined by the study's analytical level: egocentric versus complete network (Valente, 2010). In other words, how you organize the data in preparation for subsequent analysis is contingent on the study's analytical level. Data that are collected as part of an egocentric network study are typically analyzed using standard statistical packages such as SPSS, SAS, or STATA. The analysis of these relational data requires that the analyst construct various network measures for each case in the data file. These measures, discussed in Chapter 7, include egocentric network properties such as size, density, and centrality. In order to calculate these measures, egocentric data first have to be reshaped so that each tie is treated as a separate case in the data set, thereby creating what is referred to as dyadic data. That is, the data file has to be reshaped from “wide” to “long.” For example, a respondent who nominates two alters contributes two new cases to the data file, while someone who nominates four alters contributes four new cases. The analysis then follows the creation of these dyadic cases. Valente (2010) cautions that any statistical inferential tests that are performed using dyadic data must be carefully interpreted, as each dyad is not necessarily independent of the other (a violation of the usual assumption of independence). Fortunately, multilevel models can address this violation, so the analysis of egocentric network data should be done in this framework (Snijders & Bosker, 1999). In addition, each individual and dyad can also have attribute data that can be included in the analysis. Reshaping the data from wide to long is fairly simple in standard statistical software packages, and you can easily consult the application's documentation.

Complete network studies—those types of studies that most closely correspond to what you think of as social network analysis—require that the data be managed in a different manner. In these types of studies, the data can be managed in either a node-list or edge-list format. The node-list format is consistent with typical survey data storage, with each row representing a respondent and the columns representing the alters listed by that respondent. The node-list format, however, does not allow for valued data; that is, the data are binary, with a tie being either absent or present. The edge-list format is slightly different in which each row is the respondent and one alter. Therefore, each row is a dyad, which can also contain other information regarding the strength or duration of the relationship between the two actors constituting the dyad. It follows, therefore, that the number of cases in the edge-list format is equivalent to the number of ties present in the entire network. These complete network data can then be read (or directly entered) into specialized network computer programs and labels can be added to the each node's ID number. Table 4.6 shows the same relational data managed in both the node-list (top panel) and edge-list (bottom panel) formats. The programs used to analyze network data such as these are reviewed in the book's final chapter.

Regardless of whether you manage your data using a node-list or edge-list format, these specialized network programs read these data into a matrix, the data representation discussed in the previous chapter. A matrix simply consists of rows and columns that represent the respondents and alters in a network study. In a directed network study, the rows would be the respondents who send ties and the columns would be the alters that receive ties. You always have the option of entering network data into this matrix format directly, which is advisable if you are to new social network analysis. This is feasible so long as the network has a limited number of actors and you are interested in a small number of relations. Once the data are in this matrix format, the different computer programs calculate a variety of network measures and also provide useful graphical displays. Table 4.7 shows how the valued data from the edge-list format in Table 4.6 are represented in an adjacency matrix. Similar to the edge-list format in Table 4.6, this adjacency matrix shows that node 4 is connected to node 1 with a tie-strength of 2.

Table 4.6 Example Relational Data in Node-List (Top Panel) and Edge-List (Bottom Panel) Formats. In the node-list format, the first node in each row is ego, and the remaining nodes in that row are the nodes to which ego is connected (alters). So node 1 is tied to nodes 2 and 3. These relations are directed and binary (nonvalued).

1	2	3	
2	1		
3	1		
4	1	2	5
5	1	2	

1	2	2
1	3	1
2	1	2
3	1	2
4	1	1
4	2	2
4	5	2
5	1	1
5	2	2

In the above edge-list format (bottom), the fourth line of data says that node 3 is connected to node 1 with a tie-strength of 2. If a dyad is omitted, for example (5, 3), that indicates that there is no relationship between the pair. This format is appropriate when your data are directed and valued.

Table 4.7 Example Valued and Directed Relational Data From Table 4.6 in an Adjacency Matrix Format. Regardless of whether you manage your data using a node-list or edge-list format, specialized network programs read these data into a matrix. Alternatively, you can enter your data directly into a matrix format, which is possible so long as your network is relatively small.

	1	2	3	4	5
1	0	2	1	0	0
2	2	0	0	0	0
3	2	0	0	0	0
4	1	2	0	0	2
5	1	2	0	0	0

Summary

This chapter covered a set of important issues in the preanalytical stages of any network study. In some respect, this chapter can be thought of as unfolding in a series of steps that you should consider in order to successfully design a network study. These steps require that you first decide whether you are interested in an egocentric study drawn from some randomly sampled population or a complete network study that aims to explains structural properties of an entire bounded network. Determining this boundary sets up a number of critical decisions that influence data collection, measurement, and management. There are three general ways in which the network's boundary can be specified (positional, relational, or event-based). These sampling issues influence the instruments through which you can collect relational data. For example, in complete network studies in which the actors have been identified through a census, a sociometric instrument is designed to capture information about the relations among actors that constitute that network. On the other hand, in egocentric studies, you must employ some form of name generator and interpreter to induce information about each ego's network. Regardless of the analytic level and instrument that are used, there are three elements that you must keep in mind in order to ensure the highest-quality network data. Issues of reliability and validity are critically important, just as they are in standard social scientific work that operates from a quantitative framework. However, while missing data are the bane of just about any researcher, this issue is particularly problematic for network researchers, as certain relational properties are very sensitive to missing data. Therefore, every effort should be made to minimize missing data through sound sampling procedures and good instrumentation. Finally, this chapter briefly addressed how you manage relational data after they have been collected. These management issues vary according to the study's analytical level. The following chapter discusses how data collected at the complete level can be analyzed to reveal important structural properties that have historically been of interest to network researchers across various disciplines.

Chapter Follow-Up

Select a peer-reviewed empirical research article that employs social network analysis and is related to a topic of interest to you. You may consider searching various education-related databases (e.g., ERIC) or searching directly in a small number of influential journals (e.g., *Social Networks, American Educational Research Journal*, *Sociology of Education*, among others).

Please use that article to respond to the following questions.

Describe whether the study employs a positional, relational, or event-based approach to specify the network's boundary.
Were the data collected on the complete, ego, or partial network? Describe the sources of these network data.
What relations were measured and what instruments were used to measure them? Evaluate the quality of these relational data in terms of validity, reliability, error, and patterns of missingness.

Essential Reading

Bernard, H. R.Killworth, P. D.Sailer, L. Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks,*(1980).2,191–218.

Feld, S.Carter, W. C. Detecting measurement bias in respondent reports of personal networks. *Social Networks,*(2002).24,365–383.

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

http://dx.doi.org/10.4135/9781452270104.n4