2

Social Network Data: Collection and Applications

This chapter discusses characteristics of social network data, with an emphasis on how to collect such data sets. We categorize network data in a variety of ways, and illustrate these categories with examples. We also describe the data sets that we use throughout the book. As noted in Chapter 1, the most important difference between social network data and standard social and behavioral science data is that network data include measurements on the relationships between social entities. Most of the standard data collection procedures known to every social scientist are appropriate for collecting network data (if properly applied), but there are a few techniques that are specific to the investigation of social networks. We highlight these similarities and differences in this chapter.

2.1 Introduction: What Are Network Data?

Social network data consist of at least one structural variable measured on a set of actors. The substantive concerns and theories motivating a specific network study usually determine which variables to measure, and often which techniques are most appropriate for their measurement. For example, if one is studying economic transactions between countries, one cannot (easily) rely on observational techniques; one would probably use archival records to obtain information on such transactions. On the other hand, friendships among people are most likely studied using questionnaires or interviews, rather than using archival or historical records. In addition, the nature of the study determines whether the entire set of actors can be surveyed or whether a sample of the actors must be taken.

The nature of the structural variables also determines which analytic methods are appropriate for their study. Thus, it is crucial to understand the nature of these variables. The data collection techniques described here determine, to some degree, the characteristics of the relations.

2.1.1 Structural and Composition Variables

There are two types of variables that can be included in a network data set: *structural* and *composition*. Structural variables are measured on pairs of actors (subsets of actors of size 2) and are the cornerstone of social network data sets. Structural variables measure ties of a specific kind between pairs of actors. For example, structural variables can measure business transactions between corporations, friendships between people, or trade between nations. Actors comprising these pairs usually belong to a single set of actors.

Composition variables are measurements of actor attributes. Composition variables, or *actor attribute* variables, are of the standard social and behavioral science variety, and are defined at the level of individual actors. For example, we might record gender, race, or ethnicity for people, or geographical location, after-tax profits, or number of employees for corporations. Some of the methods we discuss allow for simultaneous analyses of structural and composition variables.

2.1.2 Modes

We will use the term "mode" to refer to a distinct set of entities on which the structural variables are measured (Tucker 1963, 1964, 1966; Kroonenberg 1983; Arabie, Carroll, and DeSarbo 1987). Structural variables measured on a single set of actors (for example, friendships among residents of a neighborhood) give rise to one-mode networks. The most common type of network is a *one-mode* network, since all actors come from one set.

There are types of structural variables that are measured on two (or even more) sets of entities. For example, we might study actors from two different sets, one set consisting of corporations and a second set consisting of non-profit organizations. We could then measure the flows of financial support flows from corporations to non-profit actors. A network data set containing two sets of actors is referred to as a *two-mode* network, to reflect the fact that there are two sets of actors. A two-mode network data set contains measurements on which actors from

one of the sets have ties to actors in the other set. Usually, not all actors can initiate ties. Actors in one of the sets are "senders," while those in the other are "receivers" (although the relation itself need not be directional). We will consider one-mode and two-mode, and even mention higher-mode, social networks in this book.

2.1.3 Affiliation Variables

A special type of two-mode network that arises in social network studies is an *affiliation* network. Affiliation networks are two-mode, but have only one set of actors. The second mode in an affiliation network is a set of *events* (such as clubs or voluntary organizations) to which the actors belong. Thus, in affiliation network data the two modes are the actors and the events. In such data, the events are defined not on pairs of actors, but on subsets of actors. These subsets can be of any size. A subset of actors affiliated with an affiliation variable is that collection of actors who participate in a specific event, belong to a given club, and so forth. Each affiliation variable is defined on a specific subset of actors.

For example, consider a set of actors, and three elite clubs in some city. We can define an affiliation variable for each of these three clubs. Each of these variables gives us a subset of actors — those actors belonging to one of the clubs.

The collections of individuals affiliated with the events can be found in a number of ways, depending on the substantive application. When events are clubs, boards of directors of corporations, or committees, the membership lists or rosters give the actors affiliated with each event. Often events are informal social occasions, such as parties or other gatherings, and observations or attendance or interactions among people provide the affiliations of the actors (Bernard, Killworth, and Sailer 1980, 1982; Freeman and Romney 1987). One of the earliest, and now classic, examples of an empirical application is the study of Davis, Gardner, and Gardner (1941) of the cohesive subgroups apparent in the social activities of women in a Southern city. Using newspaper records and interviews, they recorded the attendance of eighteen women at fourteen social events.

2.2 Boundary Specification and Sampling

A number of concerns arise in network studies that must be addressed prior to gathering any network data. Typically, a researcher must first identify the population to be studied, and if sampling is necessary, worry about how to sample actors and relations. These issues are considered here.

2.2.1 What Is Your Population?

A very important concern in a social network study is which actors to include. That is, who are the relevant actors? Which actors are in the population? In the case of small, closed sets of actors (such as all employees at a service station, faculty in an academic department, or corporations headquartered in a major metropolitan area), this issue is relatively easy to deal with. For other studies, the *boundary* of the set of actors may be difficult (if not impossible) to determine. The boundary of a set of actors allows a researcher to describe and identify the population under study.

Actors may come and go, may be many in number and hard to enumerate, or it may be difficult even to determine whether a specific actor belongs in a set of actors. For example, consider the study of elites in a community. The boundary of the set, including all, and only, the elites within the community, may be difficult, or impossible, to determine. However, frequently there will be a clear "external" definition of the boundary of the set which enables the researcher to determine which actors belong in it.

In some instances it is quite plausible to argue that a set of actors is relatively bounded, as for example, when there is a fairly complete membership roster. In such a case, the entire set of members can make up the actor set. However, there are other instances when drawing boundaries around a set is somewhat arbitrary. In practice, while network researchers recognize that the social world consists of many (perhaps infinite) links of connection, they also find that effective and reasonable limits can be placed on inclusion. Network researchers often define actor set boundaries based on the relative frequency of interaction, or intensity of ties among members as contrasted with non-members.

Laumann, Marsden, and Prensky (1989) describe two different approaches to boundary specification in social network studies. The first way, which they refer to as the *realist* approach, focuses on actor set boundaries and membership as perceived by the actors themselves. For example, a street-corner gang is acknowledged as a social entity by its members (it may even have a name — "Jets" or "Sharks") and the membership of the gang is the collection of people the members acknowledge

as belonging to the gang. The second way of specifying network boundaries, which Laumann, Marsden, and Prensky refer to as the *nominalist* approach, is based on the theoretical concerns of the researcher. For example, a researcher might be interested in studying the flow computer messages among researchers in a scientific specialty. In such a study, the list of actors might be the collection of people who published papers on the topic in the previous five years. This list is constructed for the analytical purposes of the researcher, even though the scientists themselves might not perceive the list of people as constituting a distinctive social entity. Both of these approaches to boundary specification have been used in social network studies.

Consider now two specific examples of how researchers have defined network boundaries. The first example illustrating the problem of identifying the relevant population of actors comes from a study of how information or new ideas diffuse through a community. Coleman, Katz, and Menzel (1957) studied how a new drug was adopted by physicians. Their solution to the problem of boundary identification is as follows:

> It was decided to include in the sample, as nearly as possible, *all* the local doctors in whose specialities the new drug was of major potential significance. This assured that the "others" named by each doctor in answer to the sociometric questions were included in the sample. (page 254)

The second example comes from the study of community leaders by Laumann and Pappi (1973). They asked community leaders to define the boundary by identifying the elite actors in the community of Altneustadt. These leaders were asked to

> ... name all persons [who] are now in general very influential in Altneustadt.

From these lists, each of which can be considered a sample of the relevant actors in the elite network, the actor set was enumerated.

Many naturally occurring groups of actors do not have well-defined boundaries. However, all methods must be applied to a specific set of data which assumes not only finite actor set size(s), but also enumerable set(s) of actors. Somehow, in order to study the network, we must enumerate a finite set of actors to study.

For our purposes, the set of actors consists of all social units on which we have measurements (either structural variables, or structural and compositional variables). Social network analysis begins with measurements on a set of actors. Researchers using methods described here must be able

to make such an assumption. We assume, prior to any data gathering, that we can obtain relevant information on all substantively important actors; such actors will be included in the actor set. However, some actors may be left out unintentionally or for other reasons. Thus, the constitution of the actor set (that is, its size and composition) depends on both practical and theoretical concerns. The reason for the assumption that the actor set consists of all social units on which we have measurements is quite simple — the methods we discuss here cannot handle amorphous set boundaries. We will always start our analyses with a set (or sets) of actors, and we must be able to enumerate (or label) all members.

Many network studies focus on small collectivities, such as classrooms, offices, social clubs, villages, and even, occasionally, artificially created and manipulated laboratory groups. All of these examples have clearly defined actor set boundaries; however, recent network studies of actors such as elite business leaders in a community (Laumann and Pappi 1976), interorganizational networks in a community (Galaskiewicz 1979, 1985; Knoke 1983; Knoke and Wood 1981; Knoke and Kuklinski 1982), and interorganizational networks across an entire nation (Levine 1972) have less well-defined boundaries.

In several applications, when the boundary is unknown, special sampling techniques such as *snowball sampling* (Goodman 1949, 1961; Erickson 1978) and *random nets* (first proposed by Rapoport 1949a, 1949b, 1950, and especially 1963; recently resurrected by Fararo 1981, 1983, and Fararo and Skvoretz 1984) can be used to define actor set boundaries. Examples of social network studies using snowball sampling include: Johnson (1990) and Johnson, Boster, and Holbert (1989) on commercial fishermen; Moore (1979) and Alba and Moore (1978) on elite networks. Such sampling techniques are discussed in the next section.

2.2.2 *Sampling*

Sometimes, it may not be possible to take measurements on all the actors in the relevant actor set. In such situations, a sample of actors may be taken from the set, and inferences made about the "population" of actors from the sample. Typically, the sampling mechanism is known, and the sample is a good, probability sample (with known selection probabilities).

We will not assume in this book that the actors in the actor set(s) are samples from some population. Most network studies focus on well-defined, completely enumerated sets, rather than on samples of actors from larger populations. Methodology for the latter situation is

considerably different from methods for the former. With a sample, one usually views the sample as representative of the larger, theoretically interesting population (which must have a well-defined boundary and hence, a known size), and uses the sampled actors and data to make inferences about the population. For example, in a study of major corporate actors in a national economy, a sample of corporations may be taken in order to keep the size of the problem manageable; that is, it might take too much time and/or too many resources actually to take a census of this quite large population.

There is a large literature on network sampling, both applied and theoretical. The primary focus of this literature is on the estimation of network properties, such as the average number of ties per actor (see Chapter 4), the degree of reciprocity present (see Chapter 13), the level of transitivity (see Chapters 6 and 14), the density of the relation under study (see Chapter 5), or the frequencies of ties between subgroups of actors (see Chapter 7) based on the sampled units.

Frank (1977a, 1977b, 1977c, 1978b, 1979a, 1979b, 1980, 1985) is the most widely known and most important researcher of sampling for social networks. His classic work (Frank 1971) and more recent review papers (Frank 1981, 1988) present the basic solutions to the problems that arise when the entire actor set is not sampled. Erickson and Nosanchuk (1983) review the problems that can arise with network sampling based on a large-scale application of the standard procedures to a network of over 700 actors. Various other sampling models are discussed by Hayashi (1958), Goodman (1961), Bloemena (1964), Proctor (1967, 1969, 1979), Capobianco (1970), Sheardon (1970), and Cabobianco and Frank (1982).

One very clever network sampling idea originated with Goodman (1961). A snowball network sample begins when the actors in a set of sampled respondents report on the actors to whom they have ties of a specific kind. All of these nominated actors constitute the "first-order" zone of the network. The researcher then will sample all the actors in this zone, and gather all the additional actors (those nominated by the actors in the first-order zone who are not among the original respondents or those in this zone). These additional actors constitute the "second-order" zone. This snowballing proceeds through several zones. Erickson (1978) and Frank (1979b) review snowball sampling, with the goal of understanding how other "chain methods" (methods designed to trace ties through a network from a source to an end; see, for example, Granovetter 1974, and Useem 1973, for applications) can be used in practice. Chain methods include snowball sampling and the small world technique discussed below. Erickson also discusses at length the differences between standard network sampling and chain methods.

In some network sampling situations, it is not clear what the relevant sampling unit should be. Should one sample actors, pairs of actors, triples of actors, or perhaps even subsets of actors? Granovetter (1977a, 1977b) and Morgan and Rytina (1977) have sensitized the network community to these issues (see also Erickson, Nosanchuk, and Lee 1981, and Erickson and Nosanchuk 1983). In other situations, one might sample actors, and have them report on their ties and the ties that might exist among the actors they choose or nominate. Such samples give rise to "ego-centered" networks (defined later in this chapter). With a sample of ego-centered networks, one usually wants to make inferences about the entire population of such networks (see for example, the epidemiological networks discussed by Klovdahl 1985; Laumann, Gagnon, Michaels, Michael, and Coleman 1989; and Morris 1989, 1990). Statistically, sampling dyads or ego-centered networks leads to sampling designs which are not simple; the sampling is actually clustered, and one must adjust the standard statistical summaries to allow for possible biases (Reitz and Dow 1989).

2.3 Types of Networks

There are many different types of social networks that can be studied. We will categorize networks by the nature of the sets of actors and the properties of the ties among them. As mentioned earlier in this chapter, we define the *mode* of a network as the number of sets of entities on which structural variables are measured. One-mode networks, the predominate type of network, study just a single set of actors, while two-mode networks focus on two sets of actors, or one set of actors and one set of events. One could even consider three- (and higher) mode networks, but rarely have social network methods been designed for such complicated data structures. Our discussion in this section is organized by the number of modes in the network. We will first discuss one-mode networks (with a single set of actors), then discuss two-mode networks, first with two sets of actors and then with one set of actors and one set of events. Applications of these three types of networks are the focus for methods presented in this book.

The number of modes in a network refers to the number of distinct kinds of social entities in the network. This usage is slightly different from the use of the term "mode" in the psychometric literature (Tucker 1964;

Carroll and Arabie 1980). In that literature, mode refers to a "particular class of entities" (Carroll and Arabie 1980, page 610). Thus, a study in which subjects respond to a set of stimuli (such as questionnaire items) gives rise to two modes: the subjects and the stimulus items. In the standard sociometric data design, a number of actors are presented with a list of the names of other people in the actor set, and asked to rate each other person in terms of how much they "like" that person. In a non-network context one could view these data as two-mode: the people as respondents are the first mode, and the names of the people as stimulus (questionnaire) items are the second mode. However, as a social network, these data contain only a single set of actors, and thus, in our terminology, it is a one-mode network in which the relation of friendship is measured on a single set of people. One might very well be interested in studying the set of respondents making evaluations of the other people, in addition to studying the people as the "stimuli" that are being evaluated. In that case one would consider respondents and stimuli as two different modes (Feger and Bien 1982; Noma 1982b; Kumbasar, Romney, and Batchelder n.d.).

We first categorize networks by how many modes the network has (one or two), and by whether affiliational variables are measured. There are, however, other kinds of relational data that are not one of these types. One example is data arising from an ego-centered network design. Data on such networks are gathered using special sampling strategies that allow the researcher to focus on a specific set of respondents, and the ties that these respondents have to particular others. We briefly describe special ego-centered networks and special dyadic designs at the end of this section.

We turn now to a discussion of one-mode, two-mode, and then affiliational, and egocentric and special networks.

2.3.1 One-Mode Networks

Suppose the network under study is one-mode, and thus involves measurements on just a single set of actors. Consider first the nature of the actors involved in such networks.

Actors. The actors themselves can be of a variety of types. Specifically, the actors may be

- People
- Subgroups
- Organizations
- Collectives/Aggregates:
  - Communities
  - Nation-states

Note that subgroups usually consist of people, organizations usually consist of subgroups of people, while communities and nation-states are larger entities, containing many organizations and subgroups. Thus, there is a natural progression of types of actors from sets of people, to collections or aggregates. Throughout this book, we will illustrate methodology with examples consisting of social network data on different types of actors.

Relations. The relations measured on the single set of actors in a one-mode network are usually viewed as representing specific substantive connections, or "relational contents" (Knoke and Kuklinski 1982). These connections, measured at the level of pairs of actors, can be of many types. Barnes (1972) distinguishes, quite generally, between attitudes, roles, and transactions. Knoke and Kuklinski (1982) give a more extensive list of general kinds of relations. Specifically, the kinds of relations that we might study include:

- Individual evaluations: friendship, liking, respect, and so forth
- Transactions or transfer of material resources: lending or borrowing; buying or selling
- Transfer of non-material resources: communications, sending/receiving information
- Interactions
- Movement: physical (migration from place-to-place), social (movement between occupations or statuses)
- Formal roles
- Kinship: marriage, descent

One or more of these types of relations might be measured for a single set of actors.

Individual evaluations are usually measurements of positive or negative affect of one person for another. Sometimes, these relations are labeled *sentiment*, and classically were the focus of the early sociometricians (see Moreno 1934; Davis 1970; Davis and Leinhardt 1972). Without question, such relations historically have been the most studied.

Transactions, or transfers of material resources, include business transactions, imports and exports of goods, specific forms of social support, such as lending and borrowing, contacts made by one actor of another in order to secure valuable resources, and transfer of goods. Such relations include exchange of gifts, borrowing or lending items, and sales or purchases (Galaskiewicz and Marsden 1978; Galaskiewicz 1979; Laumann, Galaskiewicz, and Marsden 1978). Social support ties are also examples of transactions (Wellman 1992b).

Transfers of non-material resources are frequently communications between actors, where ties represent messages transmitted or information received. These ties involve sending or receiving messages, giving or receiving advice, passing on gossip, and providing novel information (Lin 1975; Rogers and Kincaid 1981; Granovetter 1974). Information about innovations is frequently diffused over such communication channels (Coleman, Katz, and Menzel 1966; Rogers 1979; Michaelson 1990).

Interactions involve the physical interaction of actors or their presence in the same place at the same time. Examples of interactions include: sitting next to each other, attending the same party, visting a person's home, hitting, hugging, disciplining, conversing, and so on.

Movement can also be studied using network data and processes. Individuals moving between communities can be counted, as well as workers changing jobs or people changing statuses (see, for example, Breiger 1981c).

Formal roles, such as those dictated by power and authority, are also relational. Ties can represent authority of one actor over others, especially in a management setting (White 1961). Example of formal roles include boss/employee, teacher/student, doctor/patient, and so on.

Lastly, kinship relations have been studied using network methods for many years. Ties can be based on marriage or descent relationships and marriage or family relationships can be described using social network methods (for example, see White 1963; Boyd 1969).

Actor Attributes. In addition to relational information, social network data sets can contain measurements on the characteristics of the actors. Such measurements of actor attribute variables constitute the composition of the social network.

These variables have the same nature as those measured in non-network studies. People can be queried about their age, gender, race, socioeconomic status, place of residence, grade in school, and so on. For corporate actors, one can measure their profitability, revenues, geographical location, purpose of business, and so on. The "size, shape, and flavor" of the actors constituting the network can be measured in many ways.

2.3.2 *Two-Mode Networks*

Suppose now that the network under study is two-mode, and thus involves measurements on two sets of actors, or on a set of actors and a set of events. We will first consider the case in which relations are measured on pairs of actors from two different actor sets. We will then discuss a special kind of two-mode network in which measurements are taken on subsets of actors.

Two Sets of Actors. Relations in a two-mode network measure ties between the actors in one set and actors in a second set. We call such networks *dyadic* two-mode networks, since these relations are functions of dyads in which the first actor and the second actor in the dyad are from different sets.

With respect to the different types of actors, the types of relations, and the types of actor attribute variables, all of our discussion about one-mode networks is relevant. Note, however, that there can be multiple types of actors, and we can have a unique collection of attribute variables for each set of actors.

Actors. In a two-mode network that contains two sets of actors, these actors can be of the general types as described for one-mode networks. However, the two sets of actors may be of different types.

Relations. In a two-mode network with two sets of actors, at least one relation is measured between actors in the two sets. In a more extensive two-mode network data set, relations can also be defined on actors within a set. However, for the network to be truly two-mode with two sets of actors, at least one relation must be defined between the two sets of actors.

An example of such a network can be found in Galaskiewicz and Wasserman (1989). The data analyzed there consisted of two sets of actors: a collection of corporations headquartered in the Minneapolis/St. Paul metropolitan area, and the non-profit organizations (such as the Red Cross, United Way, public radio and television stations) which rely on contributions from the public sector for their operating budgets. The

primary relation was the flow of donations from the corporations to the non-profit organizations, clearly a two-mode relation. Also, it is important to note that this relation is *unidirectional* since it flows from actors in one set to actors in the other set, but not the reverse. In addition, the analysis by Galaskiewicz and Wasserman considered a number of relations defined just for the corporations (such as shared country club memberships among the chief executive officers) and several just for the non-profits (such as interlocking boards of directors). A part of this data set will be discussed in more detail later in this chapter.

One Set of Actors and One Set of Events. The next type of two-mode social network, which we refer to as an *affiliation network*, arises when one set of actors is measured with respect to attendance at, or affiliation with, a set of events or activities. The first mode in an affiliation network is a set of actors, and the second is a set of events which affiliates the actors.

An example comes from Davis, Gardner, and Gardner (1941), as described and analyzed by Homans (1950) and Breiger (1974). A set of women attended a variety of social functions, and this attendance was recorded over a period of several months. Each social function can be viewed as a variable, and a binary measurement made as to whether a specific actor attended the specific function. These variables are termed *affiliational*. Such data and networks are called *affiliation* networks, or sometimes, *membership* networks. And since the affiliations are measured on subsets of actors, such networks are non-dyadic, two-mode networks.

Actors. In an affiliation network, we have a first set of actors, and a second set of events or activities to which the actors in the first set attend or belong. The types of actors in affiliation networks can be exactly the same as those in one-mode and two-mode networks. The only requirement is that the actors must be affiliated with one or more events.

Events. In affiliation networks, actors (the first mode) are related to each other through their joint affiliation with events (the second mode). The events are often defined on the basis of membership in clubs or voluntary organizations (McPherson 1982), attendance at social events (Davis, Gardner, and Gardner 1941), sitting on a board of directors, or socializing in a small group (Bernard, Killworth, and Sailer 1980, 1982; Wilson 1982).

The nature of the events, which affiliate the actors, depends on the type of actors involved. People may attend social functions or belong to athletic clubs, subgroups of people may attend various committee meetings (for example, departments at a major university send representatives to college committee meetings), organizations may be represented on various boards of directors in a community, or countries might belong to treaty organizations, and so on.

Attributes. We can have actor attribute variables that are of the same types as those for one-mode and two-mode networks. In addition, the events themselves may have characteristics associated with them which can be measured and included in the network data set. For example, clubs will be of a particular size or located in a specific geographical area. Events usually occur at discrete points in time, as well as in particular geographical places. Thus, there can be two sets of attribute variables in an affiliation network data set: attributes of the actors, and attributes of the events.

Methods for analyzing affiliation network data are described in Chapter 8, and are applied to a network data set giving the memberships of a set of chief executive officers of major corporations in Minneapolis/St. Paul in a set of exclusive clubs.

2.3.3 Ego-centered and Special Dyadic Networks

Not all structural data give rise to standard social network data sets. With standard network data (regardless of how many modes the network has), one enumerates not only the actors, but the relevant pairs as well. All actors (theoretically) can relate to each other in one-mode networks. In two-mode networks with two sets of actors, all actors in the first mode can (theoretically) relate to all in the second. However, some data collection designs gather structural information on some pairs but not others. An example of such data arises in studies of couples. Each partner in the couple can interact with the other but with no other person during counseling sessions. Interactions during these sessions are then recorded. When interest centers on a collection of pairs (husband-wife, father-son, and so forth), one frequently samples from a large population of such pairs. We will refer to these non-network relational data as *special dyadic* designs.

An actor may also relate to a limited number of "special" other actors. For example, one might observe mothers interacting with their

own children in an experimental situation. In this case, mothers only interact with their own children, and children only interact with their own mother. Thus, the partners for one person (either mother or child) are different from the partners for another. In this situation, the design of the experiment constrains the interactions among the set of people so that all people cannot, theoretically, interact with all others.

Another related design is an *ego-centered* network. An ego-centered network consists of a focal actor, termed *ego*, as set of alters who have ties to ego, and measurements on the ties among these alters. For example, when studying people, one samples respondents, and each respondent reports on a set of alters to whom they are tied, and on the ties among these alters. Such data are often referred to as *personal network* data. Clearly these data are relational, but limited, since ties from each actor are measured only to some (usually only a few) alters. For example, in 1985 the General Social Survey conducted by the National Opinion Research Center (see Burt 1984, 1985) asked respondents:

> Looking back over the last six months — who are the people with whom you discussed matters important to you? (1984, page 119)

Respondents also reported on the ties between the people they listed. Bernard, Johnsen, Killworth, McCarty, Shelley, and Robinson (1990), Killworth, Johnsen, Bernard, Shelley, and McCarty (1990), Huang and Tausig (1990), Burt (1984, 1985), Marsden (1987, 1990b), Wellman (1993), as well as Campbell, Marsden, and Hurlbert (1986) discuss measurement of such personal, ego-centered networks.

Ego-centered networks have been widely used by anthropologists to study the social environment surrounding individuals (Boissevain 1973) or families (Bott 1957). Ego-centered networks are also used quite often in the study of social support. The term "social support" has been used to refer to social relationships that aid the health or well-being of an individual. The emphasis on relationships has allowed researchers to study support using social networks. Such networks are of great interest in clinical and community psychology, as well as in sociology. A variety of hypotheses (see Hammer 1983; Cohen and Syme 1985) have been offered to explain how personal relationships, as reflected by such ego-centered networks, can affect the emotional and physical well-being of an individual.

The methods described in this book assume that there are no theoretical limitations on interactions among actors. A social network arises when all actors can, theoretically, have ties to all relevant actors. The primary object of study for methods discussed in this book is this complete collection of actors (one or more sets) and the ties among them.

2.4 Network Data, Measurement and Collection

We now turn to issues concerning the measurement and collection of network data, the accuracy, validity, and error associated with these data, and particular design considerations that can arise in network studies.

2.4.1 Measurement

Social network data differ from standard social and behavioral science data in a number of important ways. Most importantly, social network data consist of one (or more) relations measured among a set of actors. The presence of relations has implications for a number of measurement issues, including the unit of observation (actor, pair of actors, relational tie, or event), the modeling unit (the actor, dyad, triad, subset of actors, or network), and the quantification of the relations (directional vs. nondirectional; dichotomous vs. valued). We will discuss each of these issues in turn.

Social network data can be studied at a number of different levels: the individual actor, the pair of actors or dyad, the triple of actors or triad, a subset of actors, or the network as a whole. We will refer to the level at which network data are studied as the *modeling unit*. However, social network data often are gathered at a level that is different from the level at which they are modeled. We discuss the unit of observation and the modeling unit in the next two sections.

Unit of Observation. The unit of observation is the entity on which measurements are taken. Most often social network data are collected by observing, interviewing, or questioning individual actors about the ties from these actors to other actors in the set. Thus, the unit of observation is an actor, from whom we elicit information about ties. The dyad is the unit of observation when one measures ties among pairs of actors directly. For example, one could record instances of aggression among pairs of children on a playground. When affiliation network data are collected, the unit of observation is often the event. The researcher selects events or social occasions, and for each event, records the actors who are affiliated with it.

Modeling Unit. Just as social network data can be observed at a number of levels, there are several levels at which network data can be modeled or summarized. These levels are the:

- Actor
- Dyad
- Triad
- Subgroup
- Set of actors or network

In categorizing network methods, it is useful to consider the level to which a model or network property applies. Some network properties pertain to actors (for example the number of "choices" that an individual actor receives from others in the network). Other properties pertain to pairs of actors (for example, if one person "chooses" another as a friend, is the "choice" returned by the second person?). Models at the level of the triad consider triples of actors and the ties among them. Many methods pertain to subgroups of actors; for example, one could study whether there are subsets of actors in the network who interact frequently with each other. Finally many properties pertain to the network as a whole, for example, the proportion of ties that are present in the network.

Relational Quantification. There are two properties of relations that are important for understanding their measurement, and for categorizing the methods described here: whether the relation is *directional* or *nondirectional*, and whether it is *dichotomous* or *valued*. In a directional relation, the relational tie between a pair of actors has an origin and a destination; that is, the tie is directed from one actor in the pair to the other actor in a pair. For example, one country exports manufactured goods to a second country; the first country is the source of the manufactured goods, and the second country is the destination. In a nondirectional relation the tie between a pair of actors does not have a direction. For example, we could define a tie as present between two countries if they share a border.

A second important property of a relation is whether it is dichotomous or valued. Dichotomous relations are coded as either present or absent, for each pair of actors. For example one could record whether one country sends an ambassador to a second country; thus giving rise to a dichotomous relation that can only take on two values: "send" or "not send." On the other hand, valued relations can take on a range of values, indicating the strength, intensity, or frequency of the tie between each pair of actors. For example, we could record the dollar value of manufactured goods that are exported from one country to a second country, thus giving rise to a valued relation.

2.4.2 Collection

There are a variety of ways in which social network data can be gathered. These techniques are:

- Questionnaires
- Interviews
- Observations
- Archival records
- Experiments
- Other techniques, including ego-centered, small world, and diaries

Each of these techniques will be discussed and illustrated with examples.

Questionnaire. This data collection method is the most commonly used (especially when actors are people). The questionnaire usually contains questions about the respondent's ties to the other actors. Questionnaires are most useful when the actors are people, and the relation(s) that are being studied are ones that the respondent can report on. For example, people can report on who they like, respect, or go to for advice. Questionnaires can also be used when the actor in a study is a collective entity, such as a corporation, but an individual person representing the collective reports on the collective's ties. For example, Galaskiewicz (1985) asked officers in charge of corporate giving whether or not the corporation had made a donation to a non- profit agency.

There are three different question formats that can be used in a questionnaire:

- Roster vs. free recall
- Free vs. fixed choice
- Ratings vs. complete rankings

In the following sections we will discuss each of these formats and describe examples of their use.

Roster vs. Free Recall. One issue in the design of a questionnaire to gather network data is whether each actor should be presented with a complete list, or *roster*, of the other actors in the actor set. Rosters can be constructed only when the researcher knows the members in the set prior to data gathering. For example, Krackhardt and Stern (1988) collected information on friendships among members of a university class as part of their study of "simulated" corporations. They had each person rate their friendship with every member of the class on a five point scale:

> Everyone in the class completed a questionnaire which asked them to rate every other person in the class as to how close a friend he or she was. The directions for this questionnaire included the following: "Please place a check in the space that best describes your relationship with each person on the list." The names of everyone participating in the game were listed below, with five categories from which the respondent could choose: "trust as a friend", "know well", "acquaintance", "associate name with face", and "do not know". (page 131)

For some network designs, the researcher does not present a complete list of the actors in the network to the respondent on the questionnaire. In such instances, it is common simply to ask respondents to "name those people with whom you (*fill in specific tie*)". Such a format, where respondents generate the list of names, is called *free recall.* For example, Rapoport and Horvath (1961) studied friendships in two junior high schools. Students were asked to list their best friends, but were not presented with a roster. Specifically,

> Each pupil in both schools was asked to write his name, age, grade, and home room number on a card and to fill in the blanks in the statements:
>
> - "My best friend in (name of school) Junior High School is ..."
> - "My second best friend is ..."
> - ...
> - "My eighth best friend is. ..." (page 281)

Note here how the network membership is known beforehand (all students in a school are the set of actors) but students listed their friends using free recall.

In some settings, the researcher might not even have a list; that is, the actors within the actor set might not even be known in advance. In this situation, sampling or enumeration techniques are necessary (as we have discussed earlier in this chapter). For example, in studies of community elites (Friedkin 1984; Moore 1979; Alba and Moore 1978), selected actors are asked to name other actors they believe to be influential in the community.

Free vs. Fixed Choice. If actors are told how many other actors to nominate on a questionnaire (for example, to name a specific number of "best friends"), then each person has a fixed number of "choices" to make. Such designs are termed *fixed choice.* In a fixed choice design each actor has a fixed maximum number of ties to the other actors in the set of actors. For example, Coleman, Katz, and Menzel (1957), in a study of diffusion of a medical innovation among physicians, interviewed all physicians in a community. Specifically,

> Each doctor interviewed was asked three sociometric questions:
>
> (i) "To whom did he most often turn for advice and information?"
> (ii) "With whom did he most often discuss his cases in the course of an ordinary week?"
> (iii) "Who were the friends, among his colleagues, whom he saw most often socially?"
>
> In response to each of these questions, the names of three doctors were requested. (page 254)

In this study, each person was constrained to have no more than three ties for each of the three relations.

On the other hand, if actors are not given any such constraints on how many nominations to make, the data are *free choice.* For example, Carley and Wendt (1988) studied the ties among people in an "invisible college" of users of a computer program at a variety of universities.

> Each individual was asked to denote for each member of the user group whether or not they:
>
> - Had an office next to each other
> - Attended the same school at the same time
> - Shared an office
> - Lived in the same living group or apartment
> - Were at the same school at the same time
> - Were in the same academic department at the same time

Note that there is no constraint on the number of people that an individual respondent can choose on these six relations.

The study of a university class by Krackhardt and Stern (1988) was a free choice design, since respondents were not limited in the number of friends they could choose. The Rapoport and Horvath design allowed each student to make eight choices; however, as Rapoport and Horvath note, students did not always fill in all of the 8 choices. Similarly, in a study of 384 sociograms that were collected using a fixed choice procedure, Holland and Leinhardt (1973) found that in fewer than 20

percent of the data sets did all respondents conform to the fixed number of choices.

Later in this chapter, we discuss limitations of social network data collected using fixed choice designs.

Ratings vs. Complete Ranking. In some network designs, actors are asked to rate or rank order all the other actors in the set for each measured relation. Such measurements reflect the intensity of strength of ties. Ratings require each respondent to assign a value or rating to each tie. Complete rankings require each respondent to rank their ties to all other actors.

An example of a complete rank order design is the study by Bernard, Killworth, and Sailer (1980). They asked each of forty members of a social science research office to report the amount of communication with each other member of the office using the following procedure:

> ... each participant was given the familiar deck of cards containing the names of the other participants. They arranged (that is, ranked) the cards from most to least on how often they talked to others in the office during a normal working day. (page 194)

Such data are *complete rankings* or *complete rank orders.* This questionnaire design is quite different from that employing *ratings* of the ties.

Alternatively, one can gather ratings from each actor about their ties to other members on every relation. These ratings can be dichotomous, as in the Carley and Wendt (1988) study (ties are either present or absent), or valued, as in the Krackhardt and Stern (1988) study where ratings were made by choosing one of five possible categories for the strength of each tie.

Full rank-orders and rating scales with multiple response categories produce *valued* relations. Response formats where respondents either nominate a person or not on a given relation produce *dichotomous* relations. In either case, when "choices" are directed from respondents to the people they name, the resulting relations are *directional.*

Interview. Interviews, either face-to-face or over the telephone, are occasionally used to gather network data in instances where questionnaires are not feasible. For example, Galaskiewicz (1985) interviewed the chief executive officers of the largest corporations in the Minneapolis/St. Paul metropolitan area. Chief executive officers were much more willing to participate in face-to-face interviews than via an impersonal questionnaire.

Interviews have been used to gather data from respondents in ego-centered networks, such as the 1985 NORC General Social Survey (Burt 1984, 1985), Wellman's study of social support in East York, Ontario (Wellman 1979; Wellman, Carrington, and Hall 1988; Wellman and Wortley 1990, and references therein), and Fischer's study of friendships in a community in Northern California (Fischer 1982).

Observation. Observing interactions among actors is another way to collect network data. This method has been widely used in field research to study relatively small groups of people who have face-to-face interactions (Roethlisberger and Dickson 1961; Kapferer 1969; Hammer, Polgar, and Salzinger 1969; Thurman 1980; Bernard and Killworth 1977; Killworth and Bernard 1976; Bernard, Killworth, and Sailer 1980, 1982; Freeman and Romney 1987; Freeman, Romney, and Freeman 1987; Freeman, Freeman, and Michaelson 1988, 1989). For example, Freeman, Freeman, and Michaelson (1988, 1989) observed a collection of fifty-four windsurfers on a beach in Southern California.

> Observations on the subjects' interaction patterns were made for two half-hour periods on each day of 31 consecutive days. (Freeman, Freeman, and Michaelson 1989, page 234)

The information recorded was the number of minutes of interaction between pairs of people.

Observational methods have been used extensively in the studies of Bernard, Killworth, and Sailer (Bernard and Killworth 1977; Killworth and Bernard 1976; Bernard, Killworth, and Sailer 1980, 1982). These researchers systematically observed interactions among people in a variety of social settings, such as a social science research office, faculty, staff, and graduate students in a university department, and members of a college fraternity. Their research focused on the relationship between these observed interactions and actors' recollections of their own interactions. Since data are collected by observing interactions, without requiring verbal responses from the people, this method is quite useful with people who are not able to respond to questionnaires or interviews.

Observational methods are widely used in the study of interactions among non-human primates (Dunbar and Dunbar 1975; Sade 1965). For instance, Wolfe (see MacEvoy and Freeman n.d.) observed a colony of monkeys, and recorded which monkeys visited a river together. Sailer

and Gaulin (1984) present data collected on interactions among members of a colony of mantled howler monkeys.

Observational methods are also useful for collecting affiliation network data. The researcher can record who attends each of a number of social events. For example, Freeman, Romney, and Freeman (1987) recorded which faculty members and graduate students attended a weekly departmental colloquium over the course of a semester. Each colloquium is an event in this affiliation network.

In some studies, the researcher observes a set of actors for an extended period of time, and then summarizes his or her impressions of the ties among all pairs of actors in the set (Roethlisberger and Dickson 1961; Kapferer 1969; Thurman 1980). The ties are based on the researcher's impressions.

Archival Records. Some network researchers measure ties by examining measurements taken from records of interactions. Such records can take many forms, such as measurements on past political interactions among nations, previously published citations of one scholar by another, and so on. Burt and Lin (1977) discuss how social networks can be obtained from archival data, such as journal articles, newspapers, court records, minutes of executive meetings, and the like. Frequently, as noted by Burt and Lin, such data give rise to longitudinal relations and can be used to reconstruct ties that existed in the past. For example, Burt (1975, 1983) obtained information on interactions among corporate actors from the front pages of previously published issues of *The New York Times*.

Rosenthal, Fingrutd, Ethier, Karant, and McDonald (1985) used biographical records to study the organizational affiliations of women reformers in the 19th century in New York. These researchers were interested in the overlaps among the organizations. The list of women and their affiliations was compiled from biographical dictionaries which included information about organizational affiliations of 202 women, and 1015 organizations. These data are thus affiliation data compiled from archival sources.

Galaskiewicz (1985) obtained information on memberships of the chief executive officers of corporations in Minneapolis/St. Paul in elite country clubs by examining the membership rosters of the clubs. Other researchers have conducted similar elite studies by looking at volumes such as *Who's Who*, and social registers.

Another common use of archival records is for the study of sociology of science, specifically, patterns of citations among scholars. One

can examine "who cites whom" in order to understand diffusion of a scientific innovation (Burt, 1978/1979a; Breiger 1976; McCann 1978; Noma 1982a, 1982b; Doreian and Fararo 1985; White and McCann 1988; Michaelson 1991; Carley and Hummon 1993). In these studies, the unit of observation is a citation, but since a given article usually contains many citations, the actor can be the article containing the citation, or the journal containing the article, or even the authors of the cited articles.

All of the data collection methods discussed above attempt to measure the ties among all the actors in the set. Many network studies employ a variety of data collection methods for recording ties, in addition to gathering actor attribute information. These data collection methods (questionnaires, observations, interviews, experiments, and so forth) are common social and behavioral science procedures.

Other. Here, we focus on other designs for collecting relational data. These include the cognitive social structure design (which is an extension of sociometric data to include actor perceptions of the network), experimental studies (in which network data are collected under controlled situations), and studies in which information is collected on ties among just some actors. Often these studies are used to estimate the size (de Sola Pool and Kochen 1978; Freeman and Thompson 1989; Bernard, Johnsen, Killworth, and Robinson 1989; Wellman 1992b) or composition (Verbrugge 1977; Wellman 1979; Marsden 1988; Wellman and Wortley 1990, and references therein) of an individual's ego-centered network. Perhaps only a few actors are chosen as respondents. Or, the actors might not even be members of a well-defined set of actors. Clearly in these instances, we are not studying a network with a boundary. We refer to such studies as special network designs.

In the next paragraphs, we discuss data collection procedures for cognitive social structure designs, experimental, ego-centered networks, and small- and reverse small-world techniques.

Cognitive Social Structure. In a standard sociometric questionnaire, one asks respondents about their own ties. A variation of this design is to ask respondents to give information on their perceptions of other actors' network ties. Such designs are called *cognitive social structures* because they measure perceived relations (Krackhardt 1987a; Kumbasar, Romney, and Batchelder n.d.).

As an example, Krackhardt and Porter (1985) studied turnover in several fast food restaurants. They were interested in the employees'

perceptions of friendships among all other employees in the restaurant. Thus, they had to gather information from each person not only about their own friendships, but also about their perceptions of the friendships among all other pairs of employees. They collected network data at two points in time.

Their procedure is described as follows:

> In the first questionnaire, each person in the work group was asked to record who they perceived to be a friend of whom. While simple on the surface, this substantial task required that employees consider all possible pairs of friends in the restaurant. To accomplish this, the respondent was told to check the names of all those listed whom he or she thought would be considered a friend by employee # 1 (for example, "Henry"). Then, the same list was repeated on the next page, and the respondent was asked to check all names of those whom he or she thought would be considered a friend of employee # 2 ("Rita"). This process was repeated a total of N times (for N employees). In this way, we could assess each person's perception of everyone's friends, their own as well as their coworkers. (page 250)

Alternatively, one can ask respondents to report subgroups of people who form relatively tightly knit subgroups within the larger collection of people (Freeman, Freeman, and Michaelson 1988, 1989).

Data collected using a cognitive social structure design gives considerably more information than the usual sociometric design, since actors report not only on their own ties, but also on their perceptions of ties among all pairs of actors.

Experimental. Social network data can be collected using experimental designs. There are (at least) two basic ways to conduct such experiments. First, one can choose a set of actors and observe their interactions in an experimentally controlled situation. The researcher then records interactions or communications between pairs of actors. Ties may be observed between all pairs of actors. Second, one can not only choose actors but also specify which pairs of actors are permitted to communicate with each other during the experiment. One only records the frequency or content of communications between those pairs of actors who are permitted to interact.

Group problem-solving experiments (Bavelas 1950; Leavitt 1949, 1951) in which actors are assigned to positions within the network defined by the experimenter and allowed to communicate only with specific others are an example of the second type of experiment. The experimenter manipulates both group members and their ties. Power and exchange experiments are also of the second type (Cook, Emerson, Gilmore, and Yamagishi 1983; Bonacich 1987; Markovsky, Willer, and Patton 1988; and Friedkin and Cook 1990). The experimenter assigns actors to positions, and allows certain pairs of actors to negotiate the exchange of resources.

Ego-centered. An ego-centered, or *local*, network consists of a focal person or respondent (ego), a set of alters who have ties to ego, and measurements on the ties from ego to alters and on the ties between alters. One begins by asking a collection of respondents about their ties to other people to elicit the set of alters. In 1985 the NORC General Social Survey (see Burt 1984, 1985) asked a sample of 1531 people

> From time to time, most people discuss important matters with other people. Looking back over the past six months, who are the people with whom you discussed matters important to you? (page 119)

One also asks respondents information about the ties among the people that the respondent has named. The 1985 General Social Survey contained a question about the ties among all pairs of people named by the respondent. If we label two of the people named by a particular respondent "Alter 1" and "Alter 2," then the question can be worded

> Please think about the relations between the people you just mentioned. Some of them may be total strangers, in the sense that they would not recognize each other if they bumped into each other on the street. Others might be especially close, as close to each other as they are to you. First think about [Alter 1] and [Alter 2]. Are these people total strangers? (Burt 1985, page 120)

Such measurements give rise to ego-centered networks.

Small World. Special network designs are also used in small world and reverse small world studies. A small world study is an attempt to determine how many actors a respondent is removed from a target individual based on acquaintanceship. Of primary interest is not only how long these "chains" are, but also the characteristics of the intermediate actors in the chain. This data collection design was pioneered by Milgram (Milgram 1967; Travers and Milgram 1969). Korte and Milgram (1970) describe the typical small world study as follows:

> The small world method consists of presenting each of the persons in a "starting population" with the description of a given "target person" — his name, address, occupation, and other selected information. The task of a starter is to advance a booklet toward the target person by sending

> the booklet to a personal acquaintance whom he considers more likely than himself to know the target. Each person in turn advances the booklet in this manner until the chain reaches the target. (page 101)

Often the intermediaries are asked to return a postcard to the researcher reporting some basic demographic characteristics. The researcher can then compare characteristics of successful and unsuccessful chains. Korte and Milgram (1970), Erickson and Kringas (1975), and Shotland (1976) have also used this design, as discussed by Lin (1989), and by papers in the volume edited by Kochen (1989).

A reverse small world study focuses on the ties from a specific respondent to a variety of hypothetical targets (Killworth and Bernard 1978; Cuthbert 1989). Cuthbert (1989) states:

> ... individuals are asked to imagine that they will pass something to someone who is to eventually reach a target person they do not know. They are instructed to think of someone they know, who might be a first link in a chain to the target person. ... The respondent is given a list of possible targets who are located geographically and socially in different parts of the society. In this way the reverse small world method clearly maps the outgoing network of the people who complete the questionnaire. (page 212)

White (1970) discusses the possible biases that can arise by using the small world technique. Many of these biases arise because response rates are typically much lower with this form of network data collection. Better estimation strategies of network properties are discussed by White (1970) and by Hunter and Shotland (1974).

Diary. Another way to gather social network data is to ask each respondent to keep a continuous record of the other people with whom they interact (for example, Gurevich 1961; de Sola Pool and Kochen 1978). Such methods have been used in the study of personal networks among people. For example, see Cubbitt (1973), Mitchell (1974), and Higgins, McClean, and Conrath (1985).

Social support researchers sometimes ask respondents to keep daily records of all people with whom they come into contact. In addition to generating a list of people in every respondent's personal network, these data sets frequently include information on the type of relation and characteristics of the alters in each ego-centered network (see Reis, Wheeler, Kernix, Spiegel, and Nezlek 1985; Pagel, Erdly, and Becker 1987).

2.4.3 *Longitudinal Data Collection*

Occasionally, a researcher is interested in how ties in a network change over time. In studies of such processes, one measures one or more relations at fixed intervals of time. Such designs allow one to study how stable ties are and whether such ties ever reach an equilibrium state. There are (usually) two research questions to be answered when studying longitudinal network data. The first is how the process has changed over time, while the second question asks how well the past, or the history of the process, can predict the future. Some comments on how to gather longitudinal social network data can be found in Wasserman (1979).

Longitudinal social network data can be collected using any of the methods described above (questionnaire, interview, observation, and so on). There have been some important longitudinal studies, primarily of sociometric relations, such as friendship. Other researchers have looked at communications throughout a network over time.

Nordlie (1958) and Newcomb (1961) studied two 1956 University of Michigan fraternities, each containing seventeen men housed together, for a period of fifteen weeks. All students were incoming transfer students who were initially unknown to each other. Each person was asked to rank each of his fellow fraternity members on the basis of positive feeling. Rankings were recorded each week, except for week 9. These data were studied in depth by Nordlie (1958), White, Boorman, and Breiger (1976), Boorman and White (1976), and Wasserman (1980).

Bernard, Killworth, and Sailer (1980, 1982) studied another fraternity over time, this one existing in the late 1970's in Morgantown, West Virginia. The fifty-eight fraternity members had been living together at least three months. Interactions among members within the fraternity were recorded by an outside observer every fifteen minutes, twenty-one hours per day, for five days. This observation process was conducted three times during the year. The observer noted every group in conversation, yielding a very rich set of longitudinal interaction ties. In addition, the researchers asked the fraternity members both about their "friendships" within the fraternity and about their recollections of their interactions with other fraternity members at the end of each of the three observation periods. To measure the interaction relation, the students were asked to give a rating of their interactions with each of the other actors on an ordinal scale of 1 (no communication) to 5 (great deal of communication). Thus, three longitudinal relations were studied: interaction (measured

almost continuously for three different five-day periods), friendship, and recalled communication (measured at three points in time).

Another classic example is Freeman's *EIES* data, which consist of measurements of computer mail interactions, over the course of an eighteen month period, among a set of quantitative researchers studying social networks. These data are described at the end of this chapter. Yet another example comes from Katz and Proctor's (1959) study of ties in an eighth-grade classroom of twenty-five boys and girls. These data consist of friendship choices made four times during the school year. The data were gathered by Taba (1955), who focused on the differences and similarities between boy-boy and girl-girl choices, and "mixed gender" ties.

2.4.4 Measurement Validity, Reliability, Accuracy, Error

As we noted in Chapter 1, social network research is concerned with studying patterns of social structure. As Freeman and Romney (1987) note, "social structure refers to a relatively prolonged and stable pattern of interpersonal relations" (1987, pages 330–331). In their discussion of measurement error in sociometry, Holland and Leinhardt (1973) refer to this pattern as the *true structure*, in contrast to the *observed structure* contained in the measured network data, which might contain error. Important concerns in social network measurement are the validity, reliability, and measurement error in these data. In addition, since social network data are often collected by having people report on their own interactions, the accuracy of these self-report data is also a concern. Surprisingly little work has been done on the issues of validity, reliability, and measurement error in social network data. A recent paper by Marsden (1990b) reviews this work; we summarize this and other research briefly here.

"Accuracy". Often sociometric data are collected by having people report on their interactions with other people. For example, a researcher might ask each actor to report "With whom did you talk last week?", or "What other people were at the party with you last Saturday?" In either case, the respondent is asked to recall his or her interactions. An important issue is the relationship between information collected using verbal reports and information collected by observing the peoples' interactions.

Considerable research has been done on the question of *informant accuracy* in social network data. Much of this research was conducted by Bernard, Killworth, and Sailer using very clever data collection designs in which they observed interactions among people in several different communities (for example, a fraternity, a research office, and ham radio operators) and also asked the same people to report on their interactions (Bernard and Killworth 1977, 1979; Killworth and Bernard 1976, 1979; Bernard, Killworth, and Sailer 1980, 1982; Bernard, Killworth, Kronenfeld, and Sailer 1985). They concluded that about half of what people report about their own interactions is incorrect in one way or another. Thus, people are not very good at reporting on their interactions in particular situations.

However, recent studies by Freeman, Romney, and colleagues (Romney and Faust 1982; Romney and Weller 1984; Freeman and Romney 1987; Freeman, Romney, and Freeman 1987; Freeman, Freeman, and Michaelson 1988) and by Hammer (1980, 1985) argue that particular interactions are not of primary concern to social network researchers. Rather, as we noted above, the "true" structure of the network, relatively stable patterns of interaction, are of most interest. Thus it is these long-term patterns the researcher should be studying and estimating, not the particular interactions of individuals. Freeman, Romney, and Freeman (1987) argue that verbal reports (recall of interactions) should be understood using principles of memory and cognition. They found that what people report about their interactions is in fact related to the long-range social structure, rather than to particular instances.

Another issue related to the accuracy of network data occurs when the actors in the network are organizations (for example corporations) but information on ties is collected from individuals as representatives of the organization. For example, Galaskiewicz (1985) measured donations from corporations to non-profit agencies by interviewing the officer in charge of corporate giving. One must be able to assume that the individual who is interviewed in fact has knowledge of the information being sought.

Validity. A measure of a concept is *valid* to the extent that it actually measures what it is intended to measure. Often, a researcher assumes that the measurements of a concept are indeed valid. For example, one might assume that asking people "Which people in this group are your friends?" has face validity as a measure of friendship, in the sense that the answer to the question gives a set of actors who are related to the respondent through friendship ties.

However, the validity of a measure of a concept is seldom tested in a rigorous way. A more formal notion of validity, *construct validity*, arises when measures of concepts behave as expected in theoretical predictions. Thus, the construct validity of social network measures can be studied by examining how these measures behave in a range of theoretical propositions (Mouton, Blake, and Fruchter 1955b; Burt, Marsden, and Rossi 1985).

Very little research on the construct validity of measures of network concepts has been conducted. In one study of this important idea, Mouton, Blake, and Fruchter (1955b) reviewed dozens of sociometric studies and found that sociometric measures, such as number of choices received by an actor, were related to a number of actor characteristics, such as leadership and effectiveness, thus demonstrating the construct validity of those sociometric measures.

Reliability. A measure of a variable or concept is *reliable* if repeated measurements give the same estimates of the variable. In a standard psychometric test-theoretic framework (see Lord and Novick 1968; Messick 1989), the reliability of a measure can be assessed by comparing measurements taken at two points in time (test-retest reliability), or by comparing measurements based on subsets of test items (split-halves or alternative forms). For the test-retest assessment of reliability to be appropriate, one must assume that the "true" value of a variable has not changed over time. This assumption is likely to be inappropriate for social network properties, since social phenomena can not be assumed to remain in stasis over any but the shortest spans of time. Assessing reliability of social network measurements using the test-retest approach is therefore problematic. Three approaches that have been used to assess the reliability of social network data are: test-retest comparison, comparison of alternative question formats, and the reciprocity of sociometric choices (Conrath, Higgins, and McClean 1983; Hammer 1985; Laumann 1969; Tracy, Catalano, Whittaker, and Fine 1990).

Reliability of sociometric data can also be assessed at different levels. One can study the reliability of the "choices" made by individual actors, or one can study the reliability of measures aggregated over a number of individual responses (for example, the popularity of an actor measured as the total number of choices it received) (Mouton, Blake, and Fruchter 1955a; Burt, Marsden, and Rossi 1985).

Although it is difficult to draw general conclusions from the research on the reliability of social network data collected from interviews or questionnaires, several findings are noteworthy. Sociometric questions using ratings or full rank orders are more reliable (have higher test-restest reliability) than fixed choice designs in which just a few responses are allowed (Mouton, Blake, and Fruchter 1955a). Responses to sociometric questions about more intense or intimate relations have higher rates of reciprocation than sociometric questions about less intense or intimate relations (see Marsden 1990b; Hammer 1985). Lastly, the reliability of aggregate measures (such as popularity) is higher than the reliability of "choices" made by individual actors (Burt, Marsden, and Rossi 1985).

Measurement Error. Measurement error occurs when there is a discrepancy between the "true" score or value of a concept and the observed (measured) value of that concept. It is common to assume that the observations or measurements of a concept are an additive combination of the "true" score plus error (or noise). This error, the difference between the true and observed values, is referred to as *measurement error*.

Holland and Leinhardt (1973) present a thorough discussion of measurement error and its implications in social network research. As they note, in social network research the measurements are the collection of ties among actors in the network, represented in the sociomatrix or sociogram. These measurements may differ from the "true" structure of the network. Since there are several levels at which we can study social networks (for example, one can look at properties of actors, pairs of actors, subsets of actors, or the network as a whole), it is important to understand the implications of measurement error at each of these levels.

Of particular importance in the discussion presented by Holland and Leinhardt is the error that arises in fixed choice data collection designs. Recall that in a fixed choice design, the respondent is instructed to nominate or name some fixed number of others for each relation. For example, each person may be asked to "List your three best friends." This design introduces error since it is quite unlikely that all people have exactly three best friends. The restriction of the nomination process also introduces error into the measurement of other network properties, such as properties of triads (triples of actors and their ties) and of subgroups.

2.5 Data Sets Found in These Pages

We now turn our attention to the network data sets that we focus on throughout this book. Each is described in detail, with attention given to the issues mentioned earlier in the chapter. All of these data sets,

including measurements on all relations and actor attributes (if included) can be found in Appendix B. As the reader will see, these data are quite diverse, coming from a variety of disciplines and theoretical concerns. There are five primary data sets we discuss below.

2.5.1 Krackhardt's High-tech Managers

This is a one-mode network, with three relations measured on a set of people. These data were gathered by Krackhardt (1987a) in a small manufacturing organization on the west coast of the U.S. This organization had been in existence for ten years and produced high-tech machinery for other companies. The firm employed approximately one hundred people, and had twenty-one managers. These twenty-one managers are the set of actors for this data set. Throughout the book, we will refer to this example as "Krackhardt's high-tech managers." Krackhardt's interest in these data focused on the managers' perceptions of the entire network of informal advice and friendship relations. Specifically, he was interested in the perceptions held by the managers of the structure of the entire network. As we note later, he gathered much more extensive data than we will use. Here, we are interested only in the reports made by each manager of his or her own advice seeking and friendships.

Each manager was given a questionnaire and asked two questions: "Who would [you] go to for advice at work?" and "Who are your friends?" Each manager was given a roster of the names of the other managers, and asked (in a free choice setting) to check the other managers to whom they would go for advice at work, and with whom they were friends. Krackhardt also gathered a third relation based on the official organizational chart. He recorded "who reports to whom" for all twenty-one managers.

Thus, this is a multirelational data set, with three relations: "advice," "friendship," and "reports to." All three are dichotomous and directional. The first two were gathered from questionnaires, and the third, from organizational records. These relations were measured for a single point in time. The friendship relation clearly is an individual evaluation, while the advice relation is a verbal report of an interaction between actors. The third relation is a measurement of the formal bureaucratic structure within the organization. So, this data set has three very different types of relations.

The network is one-mode, since we have just a single set of twenty-one actors. The actors are people. This data set also includes four actor attributes: age; length of time employed by the organization (tenure); level in the corporate hierarchy; and the department. The first two are measured in years. There are four departments in the firm. All but the president of the firm have a department attribute coded as an integer from 1 to 4. The level attribute is measured on an integer scale from 1 to 3: 1 = CEO, 2 = vice president, and 3 = manager.

Of primary interest to Krackhardt were the perceptions held by each actor of the friendships and advice seeking within the firm. Each actor was asked to evaluate all the ties between all actors, not just the ties involving the respondent. In this way, Krackhardt was able to study perceptions of network structure. For example, how were an actor's actual reported friendships perceived by all the other actors? Krackhardt (1987a) categorized actors by their importance (as measured by centrality indices) and found that more important actors had better perceptions than those less important.

2.5.2 Padgett's Florentine Families

This is a one-mode network with two relations measured among a set of families. These multirelational network data, compiled by Padgett, consist of the marriage and business ties among 16 families in 15th century Florence, Italy. These data were compiled from the history of this period given by Kent (1978). The 16 families were chosen for analysis from a much larger collection of 116 leading Florentine families because of their historical prominence. Padgett (1987), Padgett and Ansell (1989, 1993), and Breiger and Pattison (1986) have extensively analyzed these data. Throughout, we will refer to this example as "Padgett's Florentine families."

The actors in this network are families. As noted by Breiger and Pattison, the family was an important economic and political unit, so the history of 15th century Florence can be well understood by focusing on families, rather than individual people. In the early 1430's, a political battle was waged in Florence for control of the government, primarily between the Medicis and the Strozzis, two of the families included in this data set. An excellent account of this history can be found in Padgett (1987). We note that Padgett and Ansell (1989) studied seventy-one families, and were interested in how the Medici family rose to dominate Florence between 1427 and 1434. Of primary interest to them was the association between the two relations, marriage and business.

The two measured relations are marriage and business. Both are nondirectional and dichotomous, and are transactional, since the business relation as well as the marital ties were used to solidify political and economic alliances. A marital tie exists between a pair of families if a member of one family marries a member of the other. A business tie exists if, for example, a member of one family grants credits, makes a loan, or has a joint business partnership with a member of another family (Breiger and Pattison 1986).

For these data, Padgett was not able to determine how families married each other or how families did business with each other. This nondirectionality is proper for marital ties, but perhaps not for business dealings. A variety of authors (including Breiger and Pattison 1986) have remarked that the nondirectionality of the business relation is unfortunate, since loans and credits are clearly directed from one family to another. More recent research by Padgett and Ansell (1993) contains an updated coding of the marriage relation that records both the family for the bride and the family for the groom, so that a directional marital relation can be studied. Both relations reflect activities occurring during this time period, but are not longitudinal.

The actors are families, 16 in number. There are three actor attributes: net wealth in 1427 (as taken from government records); number of priors (seats on the city council) from 1282–1344; and number of business or marriage ties in the total network (consisting of all 116 families).

2.5.3 *Freeman's EIES Network*

This is a one-mode network with two relations measured on a set of people. These data come from a computer conference among researchers working in the emerging scientific specialty of social network research, organized by Freeman, and sponsored by the National Science Foundation. These data were collected as part of a study of the impact of the Electronic Information Exchange System (*EIES*) housed at the New Jersey Institute of Technology. Fifty researchers interested in social network research participated. We focus here on the thirty-two people who completed the study. These researchers included sociologists, anthropologists, and statisticians/mathematicians. As part of the conference, a computer network was set up and participants were given computer terminals and access to a network for sending electronic mail messages to other participants. We note that this study was done prior to the widespread use of *BITNET*, *INTERNET*, and other popular computer networks that are widely available to academics today; consequently, this study involved a novel way for researchers to communicate. For more details of this study, see S. Freeman and L. Freeman (1979), L. Freeman and S. Freeman (1980), and Freeman (1986). A more detailed description of the design of this study can be found in Bernard, Killworth, and Sailer (1982). Here, we will refer to this example as "Freeman's *EIES* network."

Of particular interest are the network data arising from this study. Two relations, messages sent and acquaintanceships, were recorded. As part of this project, the computer system recorded all message transactions, specifically the origin and destination of the message, the day and time, and the number of lines in the message. Records were kept for several months. We therefore have a record of the number of messages sent from each participant to every other participant. We restrict our attention to the total number of messages sent from one actor to another; however, this message-sending relation can be defined for any time interval, for example, the number of messages sent in a given month. A second relation is acquaintanceship, and was gathered by a questionnaire. At the beginning and at the end of the project, participants were asked to fill out a questionnaire that included, among other things, a network question. Each participant was asked to indicate, for every other participant, whether she/he: (1) did not know the other, (2) had heard of the other but had not met him/her, (3) had met the other, (4) was a friend of the other, or (5) was a close personal friend of the other. This acquaintanceship relation is longitudinal, measured at two points in time: at the beginning of the study (January 1978), and at the end (September 1978) (S. Freeman and L. Freeman 1979).

There are two attribute variables in this data set: Primary disciplinary affiliation of the person; and Number of citations of the researcher's work in the *Social Science Citation Index* for the year 1978 (when the research started). The disciplinary affiliation variable has four categories: (1) sociology, (2) anthropology, (3) mathematics or statistics, and (4) other. The citation variable is coded as the number of citations.

These data are a part of a more comprehensive data set gathered by Bernard (who, along with Freeman, supplied us with these data) to study the accuracy of informants' reports of communications (see Bernard, Killworth, and Sailer 1982). Freeman (1986) studied the impact of this newly formed computer network on the acquaintanceships and friendships among the network researchers. Wasserman and Faust (1989) used these data to demonstrate the application of correspondence and canonical analysis to social network data.

2.5.4 Countries Trade Data

This is a one-mode network with five relations measured on countries. These data were gathered by us for use in this book. The actors are countries, selected from a list of sixty-three countries given in Smith and White (1988). We chose countries representing different categories from across several developmental classifications: Snyder and Kick's (1979) core/periphery status, Nemeth and Smith's (1985) alternative world system classification and level of industrialization, and a historical economic base from Lenski (as reported in Breedlove and Nolan 1988). We also chose countries both to span the globe and to represent politically and economically interesting characteristics. Only countries for which data were reported in 1984 commodity trade statistics were eligible for inclusion. We also attempted to reduce the number of shared borders between countries; however, some politically interesting countries are included even though they share borders (Israel and Syria, for example). Because of data availability, less-developed nations (African nations in particular) are probably under-represented in this set.

The final twenty-four countries represented as actors in this network are a geographically, economically, and politically diverse set, chosen to represent a range of interesting features and to span the categories of existing world system/development typologies. We will refer to these data as the *countries trade network*. Because of the selection mechanism, we will assume that this set of actors is representative of all possible countries.

Five relations were measured. Four of them are economic and one is political. The relations are:

- Imports of food and live animals
- Imports of crude materials, excluding fuel
- Imports of mineral fuels
- Imports of basic manufactured goods
- Diplomatic exchange

The first four relations are taken from the United Nations Commodity Trade Statistics (1984). We chose these four types of commodities (with single digit section codes 0, 2, 3, and 6 from the commodity trade statistics) since these commodities were studied originally by Breiger (1981a). The last relation comes from *The Europa Year Book* (Europa Publications 1984), which lists for each country those countries that have embassies or high commissions in the host country.

All five relations are dichotomous and directional. The four economic relations were reported on a continuous US$ scale. The reported values indicate the amount of goods (of the specified type) in 100,000 US$ imported by one country from the other (the UN does not list trade amounts under 100,000 US$). In order to standardize the imports to control for the vastly different economy sizes across countries, we first standardized each value by dividing by the country's total imports on that commodity. If the realized proportion was less than 0.01%, we coded the tie as absent. Otherwise, the tie was coded as present. This standardization actually had very little impact. Most of the ties that were changed from "trade present" to "trade absent" were large countries (US, Japan, UK) importing small amounts from very small countries (Madagascar, Liberia, Ethiopia).

The diplomatic relation records a tie as present if one country has an embassy or a high commission in another country. These data are taken from the 1984 *Europa Year Book* (Europa Publications 1984).

The data set includes four attribute variables reflecting the economic and social characteristics of the countries. The first two attribute variables measure annual rates of change between 1970 and 1981. They are: Annual population growth rate between 1970 and 1981, and Annual growth rate in GNP per capita between 1970 and 1981. The second two attribute variables measure rates of education and energy consumption. These variables are: Secondary school enrollment ratio in 1980, and Energy consumption per capita in 1980 (measured in kilo coal equivalent). Researchers have argued that these variables are related either to level of national development (industrialization) or to world system status. Measurements on these four variables were taken from The World Bank (1983).

Numerous social scientists have used network methods and data to study the world political and economic system (Snyder and Kick 1979; Nemeth and Smith 1985; Breiger 1981c). These researchers are primarily interested in whether location in a network "system" affects the rates of industrialization and development.

2.5.5 Galaskiewicz's CEOs and Clubs Network

This data set is a two-mode, affiliation network. The first mode consists of twenty-six chief executive officers (and spouses) of the major corporations, banks, and insurance companies headquartered in the Minneapolis/St. Paul metropolitan area. These data were gathered by Galaskiewicz

through interviews with the CEOs and records of the clubs and boards. Thus, the first mode is a set of corporate CEOs as actors. The second mode is a collection of fifteen clubs, cultural boards, and corporate boards of directors to which the CEOs belong. There are two country clubs (Woodhill Country Club and Somerset Country Club), three metroplitan clubs (Minnesota Club, Minneapolis Club, and the Womens Club), four prestigious cultural organizations (such as Guthrie Theater, Minnesota Orchestra Society, Walker Art Center, St. Paul Chamber Orchestra, Minnesota Public Radio), and the six corporate boards of the *FORTUNE* 500 manufacturing firms and *FORTUNE* 50 banks headquartered in the area. These data record which CEO belongs to each of the clubs and boards. These memberships are for 1978–1981 (as discussed by Galaskiewicz 1985). We will refer to these data as *Galaskiewicz's CEOs and clubs.*

All data are dichotomous, indicating presence or absence of a membership. The first mode is a set of people, and the second, a set of organizations. The data are affiliational, and represent memberships. There are a number of attributes that are measured for both modes. For the first mode, we can categorize the actors by the nature of the corporations they head. For the second, we can categorize the organizations by their nature (clubs or corporate boards).

2.5.6 Other Data

In addition, we analyze a hypothetical data set throughout the book. This data set is used mostly to illustrate calculations, and consists of six second-grade children. It has measurements on four relations, three measured for the first mode (a set of six children) and one for actors in the first mode choosing actors in the second mode (a set of four teachers). One of the relations is longitudinal — friendship at the beginning and end of the school year. In addition, we have a single affiliation relation (party attendance). There are also a number of attributes that are recorded for both children and teachers, which will be introduced as needed.

Part II

Mathematical Representations of Social Networks