3 Data

In this chapter you will learn what is special about network data and its collection. You will be able to differentiate between data about actors and data about relations between them, and thus be able to identify the kind of data required for an empirical study using any of the network types introduced in the previous chapter. We will also discuss straightforward ways of organizing various types of data.

The distinctive characteristic of network studies is their relational perspective. This perspective is reflected most clearly in the type of data used. Attributes are collected and analyzed not only for actors, such as persons or organizations, but also, and even more importantly, for the relations between them. A dyadic attribute is an attribute associated with pairs of actors, not just individual actors, and because each actor appears in several of these pairs, there are inherent and inevitable dependencies. Unlike in population studies, in which dependencies among actors are often considered a nuisance, they are the primary focus of interest in network studies.

In the simplest case, dyadic data only differentiate between the existence or non-existence of a relationship. In more complex scenarios, however, ties can be comprised of multiple relationships, each of which may require more detailed representation than a binary variable. For instance, if you are studying contact networks, you would be expecting relevant variation across media (face-to-face, phone calls, emails, etc.) and intensities of communication. It would be careless, therefore, to record only the pair of actors that communicates sufficiently to match some definition of "being in contact."

In the following section we will differentiate between the various types of network data and then focus on their collection. We will also discuss

quality issues and some ethical considerations for studies involving humans at the end of this chapter.

3.1 Kinds of Data

Network studies are grounded in theories about the antecedents and consequences of social relations between actors. Hence, they make use of data that differ from those in population studies. In this section, we will clarify what exactly this difference entails. In fact, the type of data studied represents the most precise distinction between population studies and network studies.

3.1.1 Units and Levels

Data generally refers to the *values* of a collection of *variables*. In an empirical study, variables are associated with the *entities* of interest, and usually describe their properties. Depending on the possible expressions of a property, variables have values in different *ranges*.

For example, it is convenient to describe the height of a person in a *numerical range* such as decimal numbers because height is a measurable quantity. In contrast, hair color appears to be more appropriately described by a category label, and is thus a *nominal variable*. The range of labels is either standardized or crafted for the study. Numerical variables are often referred to as *quantitative* while nominal variables are called *qualitative*, however this distinction is neither useful nor unanimous. For instance, if height is not expressed in centimeters but by classification into short, medium, and tall, we obtain an *ordinal* variable. Ordinal variables can be meaningfully compared by rank but not by differences or ratios. While mathematically inclined analysts would probably classify ordinal variables as qualitative because they do not represent quantities in a formal sense, more qualitatively-oriented researchers might still refer to them as them quantitative because they admit formal treatment. We shall not take sides in this debate.

A collection of variables representing the same kind of property for a group of entities is often called an *attribute* or a *feature* of that group.

The entities, for which we collect (or someone else has collected) data are called the *units of observation*. The reason for this is that, even when

such entities, like companies, are aggregates, they are the finest level of detail considered during data collection. Otherwise, it would also be necessary to obtain data on the constituents, which would make the constituents the units of observation.

The *level of analysis* is the level, for which we would like to obtain results. Like the unit of observation, the *unit of analysis* is defined by the entities that are characterized by the data used in an analysis.

The units of observation and analysis need not be the same, however. Assume for example, we are conducting an ego-centered network study and collect data about the ties between egos and their alteri. Then the variables representing this data are indexed with ego-alter dyads, which are, therefore, the unit of observation. If we then compare egos by the size of their networks, we are actually comparing numbers associated to the egos. Hence, the size attribute is indexed by the egos which are, therefore, the unit of analysis.

Network studies differ from more common population studies, most notably by their units of observation. In addition to actor attributes, i.e., variables indexed by actors, they crucially involve tie attributes. While actor attributes are the same type of data as those used in population studies, the units of observation for relational attributes are overlapping dyads. Dyads are composite entities, however the main differences arise because actors participate in multiple dyads.

The reasons why the process of deciding on the units may not be as obvious as it seems are outlined briefly below. For a more detailed treatment of the choice of representation, see Butts (2009).

Actors

Actors are an elementary unit of observation because, in most cases, you will be interested in having at least some attribute information about them, even if the core interest is the relations between them.

However, clarifying the precise entities that form the set of actors may be less obvious than it seems. As in any socio-empirical study, actors can be *individual* (or *personal*), such as humans or animals, or they can be *aggregate* (or *corporate*), such as couples, flocks, or organizations. The decision for a particular unit of observation is a theoretical one because it should match the acting units in the social system being studied. In a study involving the analysis of economic exchange processes among pastoral nomads, it may be appropriate to conceptualize households as the

interacting units. In a study of social support among villagers in an industrialized society, on the other hand, households are probably too coarse a granularity. The important distinction need not be between individual and aggregate actors but could be between different levels of aggregation. This is illustrated by the choice between business units or entire companies as actors.

In addition, the theoretically informed selection of aggregation level may be complicated by pragmatic considerations such as cost or access. For example, it may be desirable for a particular research question to survey the male and female members of households separately. However, due to societal and religious factors, this may be virtually impossible on a practical level. In such cases it may be necessary to compromise on the definition of actors, and to make up for it by modifying the kind of data analyzed and the means used to collect it.

Ties

The defining characteristic of a network study is the focus on one or more relations between actors. Since the terms relation, relationship, tie, and dyad are often used with shifting or overlapping meanings, let us settle on one definition first.

We say that a pair of actors forms a *dyad*. A dyad, like an actor, is an entity, a unit of observation, defined so that we can associate variables with it. Hence, dyads serve as indices of variables but have no values themselves.

Interestingly, this is already sufficient terminology to delineate the scope of network studies: they differ from other empirical studies in that: (i) Essential units of observation are dyads; and (ii) some dyads overlap by design.

A *relationship* is a variable associated with a dyad. Such dyadic variables have three aspects: a *content*, a *direction*, and a *value*. By way of a concrete example, let us assume that we are interested in cell phone communication among a group of school children. Each pair of distinct pupils i and j forms a dyad, and we associate with each dyad the variables x_{ij} and x_{ji}, representing the number of times that i called j and j called i over a certain period of time. The variable x_{ij} represents a calling relationship (content) from i to j (direction) and has a numerical value.

The totality of all pairwise relationships that represent the same type of content makes up a *relation*. In the aforementioned example, the set

of phone-call relationships defines a valued relation. A Boolean relation consists of relationships that are either present or absent but know no other values. We say that i and j "*are in* (a certain) *relation*," if the corresponding relationship is present or has a non-zero value. If, in the above example, we are only interested in who talked to whom on the phone, irrespective of who called whom, we are studying a relation that is *symmetric* by design, i.e. $x_{ij} = x_{ji}$, and therefore also referred to as *undirected*. Note that these are features of relations and not individual relationships.

A *tie*, on the other hand, is the union of all present or non-zero relationships of a particular ordered pair i and j. Hence, a tie summarizes all relationships of a dyad in one direction. Consequently, the concept of "being tied" coincides with the concept of "being related" in the case of a single relation. In the case of several relations, a tie may consist of several relationships and is then referred to as *multiplex*. Multiplex ties can be difficult to compare if they are composed of different relationships.

Based on this terminology, the statement that a network consists of actors and the ties between them has a reasonably precise meaning.

The units of observation are dyads not ties. Ties are data on dyads. Since dyads are pairs of actors, the difficulty does not lie in finding the right level of aggregation because this is already determined by the choice of actors. Instead, it lies in the selection of dyads (all pairs? which pairs?) and, most importantly, in the definition of what constitutes a relationship.

Borgatti et al. (2009) provide a classification of example relations studied in the social sciences. We reproduce this classification in Figure 5. The collection of dyadic data shares most of its challenges with the collection of data on actors alone or any attribute data, for that matter. It should be noted, however, that some aspects are more likely to be relevant for relations. Let us discuss two such examples.

First, consider the directionality of relations. For example, we may be interested in the provision of social support among individual actors. When we conceptualize this from either the angle of giving or receiving support, we must be aware that the two relations we end up with need not be the reverse of each other. An example involving directed vs. undirected relations is friendship. While we may conceive friendship as symmetric, its expression or interpretation by interviewees may be asymmetric.

Second, some relations conceived as pair-wise are best observed indirectly via proxy data. A particular case of such indirect relations is encountered when contacts or interactions are facilitated by social settings

Similarities			Flows
Location	*Membership*	*Attribute*	
Same	Same	Same	Information
spatial	clubs	gender	Beliefs
and			Personnel
temporal	Same	Same	Resources
space	events	attitude	

Social Relations				Interactions
Kinship	*Other role*	*Affective*	*Cognitive*	
Mother of	Friend of	Likes	Knows	Sex with
	Boss of			Talked to
Sibling of	Student of	Hates	Knows about	Advice to
	Competitor of			Helped
	Student of		Sees as happy	Harmed

Figure 5: Typology of exemplary relations (Borgatti et al. 2009)

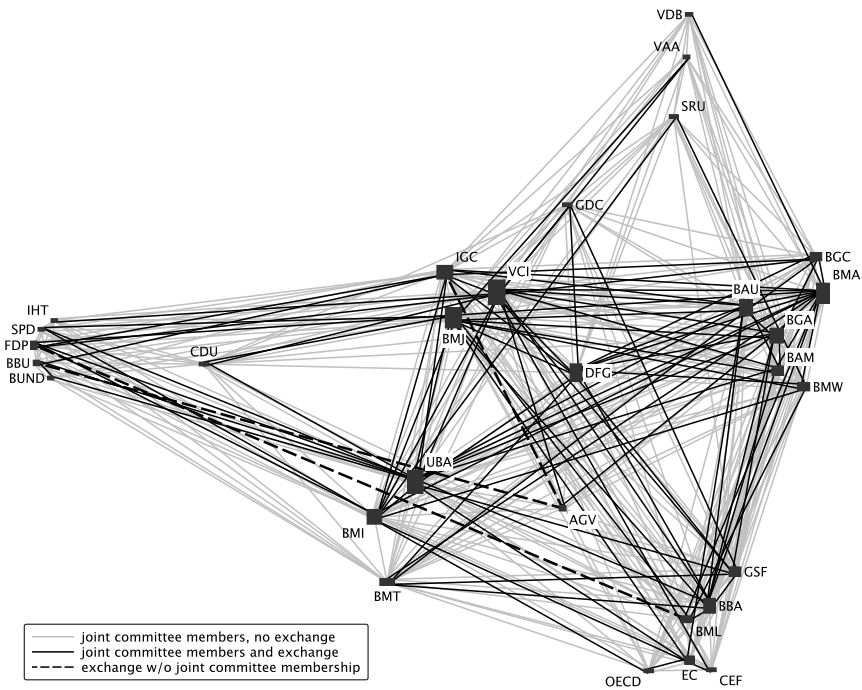


Figure 6: Joint committee membership as an infrastructure for information exchange among actors involved in a procedure for the regulation of chemicals. Redrawn from Brandes and Schneider (2009).

Box 12: Schweizer's Javanese Village

In his study of the social structure of a Javanese village, Thomas Schweizer (1996) made use of indirect relationships. So-called slametan rituals are regularly staged in the village to mark certain events in the life cycle and special occasions (e.g., harvest). Up to 200 people participate in these rituals which are organized by households. Over the course of two months, which were characterized by a particularly large number of activities (July and August 1979), Schweizer recorded who had taken part in which rituals through observation and subsequent surveys of the individual households. From these "attendance lists" he was able to reconstruct the relationships between the 98 households examined. These relationships provided information regarding who participated in a ritual, and with whom and how often they participated. He analyzed this network for the purpose of testing hypotheses on social stratification, the importance of kinship and relationships with neighbors, and the formation of social cohesion in village communities.

such as organizations or events. Figure 6 shows an example, in which this is actually a theoretical argument, and an example from an ethnographic study is provided in Box 12.

3.1.2 Organization

As we have just outlined, networks are represented by

- *actor attributes*, consisting of variables indexed by actors, and
- *tie attributes*, consisting of variables indexed by dyads.

Depending on the type of network, there will be differences in the composition of the sets of actors and dyads that index your variables, however before we describe these in more detail in the next section, let us first take a look at how attributes can be organized in data files.

While many software tools for network data analysis use their own formats for storing data in files, it is usually not necessary to read, let alone write them. The single most important format, and the only concept you really need to understand, is the data table. There are variations in the ways, in which data are arranged in tables, and we will discuss these

id	name	wealth	priors	ties
1	Acciaiuoli	10	53	2
2	Albizzi	36	65	3
3	Barbadori	55	?	14
4	Bischeri	44	12	9
5	Castellani	20	22	18
6	Ginori	32	?	9
7	Guadagni	8	21	14
8	Lamberteschi	42	0	14
9	Medici	103	53	54
10	Pazzi	48	0	7
11	Peruzzi	49	42	32
12	Pucci	3	0	1
13	Ridolfi	27	38	4
14	Salviati	10	35	5
15	Strozzi	146	74	29
16	Tornabuoni	48	?	7

Figure 7: Actor attributes represented in an actor-by-attribute table. Data provided by John Padgett for Breiger and Pattison (1986).

variations now, however, you can always think of your data as being represented in a spreadsheet.

Actor Attributes

Actor attributes are no different from attributes in other empirical studies. Each attribute is a collection of variables, one per actor, with values in a common range representing the same type of information. The conventional organization of actor attributes is a table with one row per actor and one column per attribute.

In the example given in Figure 7, actors are identified by a special attribute, ID, that has no other purpose than to disambiguate them. While it may generally be possible to identify actors via their names or some other label, this may quickly become inconvenient or even change over the course of your research. It is thus advisable to create an extra identification. As a general rule, every entity that serves as an index for variables should have a unique identifier associate with it.

Each row in Figure 7 represents a noble family in Renaissance Flo-

rence and thus an aggregate actor composed of many individuals over time. Each column represents an attribute and each cell contains the value of one variable. For instance, attribute wealth consists of the variables $\mathsf{wealth}_1, \mathsf{wealth}_2, \ldots, \mathsf{wealth}_{16}$ which have values $10, 36, \ldots, 3$.

If you find yourself asking what these values represent, you have just experienced the need for a *codebook*, in which everything that it is necessary to know about what this data represents is documented. In the case of the attribute wealth, we would learn that the number represents the total wealth of a family rounded to thousands of lira in the year 1427, and that the numbers have been extracted from that year's original tax reports. The codebook would also explain that the question marks represent missing values for priors, the number of seats in the municipal council that a family occupied during the years 1282–1344. Moreover, it would point to the fact that the Pazzi have zero priors because of another role that rendered them ineligible.

The last column is interesting because it contains an index that is not directly observed but derived from tie attributes (in this case counting the number of marital and business relations, in which a family is involved). Many of the analytic techniques discussed in Chapter 4 yield such indices.

In general, all actor attributes can be stored in one table and, therefore, in one file. If actors are partitioned into groups, for which different attributes are available, however, it may be better to create one table per group of actors rather than leaving the cells that correspond to unavailable attributes empty. This is particularly common for ego-centered networks and other two-mode data, and corresponds to horizontal cuts through the table.

Vertical cuts may also be useful, for instance when the number of attributes is excessive, or different researchers work on the same data but not all of them may access all attributes. In this case, some columns may be repeated and the id-column must be repeated to avoid reliance on the order of the rows.

Tie Attributes

The focus on relational data implies that some of our attribute data is associated with composite entities, namely dyads. The index of a relational variable is thus two-dimensional, which leaves us with several alternatives as to how to organize them. In all but exceptional circumstances one of the following three alternatives will be suitable.

marital	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
5	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
14	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
15	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0
16	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0

Figure 8: Marriage relation represented in an actor-by-actor table.

participation	E1	E2	E3	E4	E5	E6	E7	E8	E9	...
A1	0	0	0	0	0	0	0	0	0	
A2	0	0	0	0	1	0	0	0	0	
A3	0	0	0	0	1	0	0	0	1	
A4	0	0	0	0	0	0	0	0	0	
A5	0	1	0	0	0	0	0	0	0	
A6	0	1	0	0	0	0	0	0	0	
A7	1	1	0	0	0	0	0	0	0	
⋮										

Figure 9: Participation relation represented in an actor-by-event table.

The first mode of organization is in actor-by-actor tables as illustrated by Figure 8. This actor-by-actor table represents a single relation, marital, among families. Rows and columns are indexed by the id-values defined for each family. A relationship between the family in row i and column j exists and is coded as $\mathsf{marital}_{ij} = 1$, if there is at least one marriage of members from the two families. Note that this is a nominal attribute, even though it is represented by numbers. Moreover, the corresponding relation is symmetric by design. It is usually more convenient, however, to represent each pair of relationships $\mathsf{marital}_{ij} = \mathsf{marital}_{ji}$ explicitly.

If the actors are partitioned into two groups, such that a relationship can only exist between actors belonging to different groups (as in two-

id	id	marital	business
1	9	1	0
2	6	1	0
2	7	1	0
2	9	1	0
3	5	1	1
3	6	0	1
3	9	1	1
3	11	0	1
4	7	1	1
4	8	0	1
4	11	1	1
4	13	1	0
5	8	0	1
5	11	1	1
5	15	1	0
6	9	0	1
⋮	⋮	⋮	⋮

(a) tie-by-attribute table (generalized edge list)

id	(id,marital,business)											
1	9	1	0									
2	6	1	0	7	1	0	9	1	0			
3	5	1	1	6	0	1	9	1	1	11	0	1
4	7	1	1	8	0	1	11	1	1	13	1	0
5	8	0	1	11	1	1	15	1	0			
6	9	0	1									
⋮	⋮											

(b) actor-ties list (generalized adjacency list)

Figure 10: Compact formats for tie attributes.

mode networks), the representation in Figure 9 is more efficient. Two-mode matrices are called *incidence matrices* when columns are interpreted as edges in a (hyper)graph representation (see Chapter 4).

The convenient actor-by-actor table form of organization has at least two shortcomings. Since table cells are usually restricted to simple numerical or textual data types, one table is required for each relation. In addition, if the number of ties is low compared to the number of dyads, matrices waste a lot of storage space.

Two alternative forms of organization that do not suffer from these shortcomings are illustrated in Tables 10(a) and 10(b).

The tie-by-attribute table is organized in exactly the same way as the actor-by-attribute table, namely by assigning one row to each tie and one column to each attribute. Since ties are identified by an ordered pair of actors, however, two columns with id-values are required. If all relations are symmetric, each dyad is listed only once. Pairs of actors, for which no relationship exists are always omitted so that much space is saved. On the other hand, it is more difficult to look up a value because the rows

have to be scanned until the right index pair is found. If a sought tie does not exist, we will not find this out until we have inspected all of the ties that do exist. In the actor-by-actor table, we know exactly which cell, in which to look. This format is also called a (generalized) *edge list* because the first two columns form the list of edges in a graph representation (see Chapter 4).

The actor-ties list contains one entry for each actor. Each entry has an associated list that consists of all ties, in which the actor is involved in and all values associated with these ties. This representation is called a (generalized) *adjacency list* because each row lists the adjacencies of an actor in a graph representation (see Chapter 4). The list format summarizes all rows of the tie-by-attribute table pertaining to the same actor. It is the standard graph data structure used in software tools because it also only represents existing ties and provides more efficient access to them. If ties are symmetric and only one direction is listed, the lists of both actors must be scanned to find the attributes of a tie, or to be sure that the tie does not exist.

Since software for network analysis will be able to convert between these formats, you should select a format that is most convenient for your particular situation. When data is typed in by hand, for example, actor-by-actor tables are the most convenient format if there are only a few relations between a small number of actors. If there are too many actors to fit the entire actor-by-actor table on the screen, tie-by-attribute tables are preferable. In some cases, in which relations are few and binary and particularly when data was collected without asking the subjects about their ties, it may even be easier to list the actors' ties.

3.1.3 Which Data for Which Type of Network?

The main types of social networks have been distinguished in Chapter 2 based on the type of research question to be addressed. For these we can now state more precisely which kind of data you can expect to be dealing with in your study. Figure 11 provides an overview.

Complete Networks

Complete networks are studied when the research interest is focused on dependencies manifest in direct and indirect relations among a fixed set of actors and, possibly also, their attributes.

network/attributes	actors	ties
complete	all actors	all dyads
cognitive social structures	all actors	all dyads multiple times
two-mode	all actors in both modes	two-mode dyads only
ego	egos and alteri	ego-alter dyads
personal	egos and alteri	ego-alter and alter-alter dyads

Figure 11: Commonly used network data.

In principle, therefore, the dyads of interest are all pairs of actors. A common exception are dyads involving the same actor at both ends of a tie because this may be impossible in relations such as "child-of."

Cognitive Social Structures

Cognitive social structures are a generalization of complete networks and a special case of triadic data. Since every actor provides data for every dyad in the network, the variables are actually indexed by three actors: the one from which the value is solicited and the two forming the dyad.

Since each actor essentially provides a complete network and analyses often create aggregate dyadic data from the dyadic data provided by each actor, the usual format is as a collection of complete networks.

Similar statements apply to longitudinal networks that are available as panel data created, for example, in several rounds of data collection. Discrete observation times assume the role of actors providing their perception of a network.

Two-mode Networks

In two-mode networks, there are two kinds of actors and the relation studied is such that actors of the same kind cannot be related.

The two kinds of actors may require different attributes, hence, two actor-by-attribute tables are needed with different columns.

Note that indirect relations such as joint participation in events in the above example from Schweizer (1996) are derived from a two-mode affiliation relation. A *one-mode projection* is one of the two possible single-mode

networks obtained by summarizing for a dyad of actors of the same mode their common affiliations with actors of the other mode. In the case of families participating in rituals, this is the number of jointly attended rituals or the number of families attending both rituals respectively. While there are more complicated types of summarization, one-mode projections typically correspond to multiplications of an actor-by-actor table with its transpose, or vice versa for the other mode. It is convenient, therefore, to organize a two-mode network into an actor-by-actor table with different sets of row and column actors, although software tools will also be able to work with the more compact representations.

In any event, even if your analysis is based on its one-mode projection, you should keep the original two-mode data. While you can also obtain one-mode projections, reversing the operation is not generally feasible.

Ego Networks

The purpose of an ego network study is to compare members of a population (the egos) who are characterized in part by their social environment. The attributes of interest are, therefore, both individual and relational.

The individual ego actor attributes are represented in an actor-by-attribute table as usual. Attributes of alteri, however, are typically not stored in such a table because they are not of interest independently of the respective ego. Since there is a unique ego-alter dyad for each alter, alter attributes are represented, therefore, as a a tie attribute instead.

In a typical study, in which ego-alter dyads are evaluated individually, there are hence two tables linked by ego-ids: one containing ego attributes, the other containing ego-alter tie attributes.

If only summative aspects such as a diversity index of alter attributes or the total strength of ego-alter ties are relevant, the representation can be simplified. Values aggregated over the alteri of each ego can be respresented simply as another ego attribute. The entire data may then consist of ego attributes only. Although the format is then the same as in other population studies, it still originates from relational data and is analyzed from a relational perspective.

Personal Networks

Personal network studies have the same general focus of comparing a population of egos with respect to their social embeddedness. Their data, however, subsume ego network data. In addition to ego and alter at-

tributes and ego-alter tie attributes, they include attributes on alter-alter dyads. Since each alter is affiliated with exactly one ego, and we are only studying dyads between alteri affiliated with the same ego, we are actually studying a collection of complete networks whose boundaries are defined by the alteri of each ego.

To be useful for the comparison of egos, alter attributes will typically be the same for the alteri of different egos. The data are therefore organized into a (global) ego attribute table, a (global) alter attribute table, and one complete network data set per ego.

Hence, and in contrast to ego networks, alter attributes are not represented as attributes of ego-alter ties but rather the other way around. Since alteri in personal network studies are actors in a set of complete networks, they are of interest beyond their affiliation with ego. In fact, the network in which an ego is embedded may contain more actors than are related directly to ego.

3.2 Data Collection

Overlapping dyads as a unit of observation are particular to network studies and require procedures that differ from the collection of attribute data on population samples. Hence, this section focuses on methods for determining the set of dyads on which tie attribute data is collected. For more general introductions to data collection see Lohr (2009), Babbie (2009), or Bernard (2012).

3.2.1 Sources

The initial decision to be made is whether it is necessary to collect data yourself (*primary data*), or whether you can use existing data already compiled by others or yourself, albeit in all but rare cases for a different purpose (*secondary data*).

Primary data can be collected in two ways: *actively* by probing sources of information, i.e., by providing a stimulus to initiate a response from the desired information is inferred, or *passively* by observing a situation without interference.

The main category of active data collection is the survey, and the main categories for passive data collection are observation and archival work.

The delineation is not sharp, since several data collection strategies fall into multiple categories simultaneously. An example would be the compilation of a tailored bibliographic data set using specific queries to multiple databases.

Surveys

Surveys are a form of active data collection and are typically conducted using questionnaires or interviews, where questionnaires can be paper-based or Web-based and interviews can be face-to-face or via the telephone (Fowler 2009).

The most commonly used survey method is the questionnaire. While traditional, paper-based questionnaires are increasingly replaced by Web-based forms, the core aspects relating to networks are still the same. They are used to obtain information on relationships of the respondents themselves or of actors they are assumed to know about. For example, respondents are asked to report on who they give advice to or with whom they share information. Questionnaires can also be used for aggregate actors, such as organizations, by having individual actors representing the collective to provide information on the collective's ties.

The use of questionnaires is not bound to any of the network types, however their design may need to differ accordingly (Marsden 2005). A corresponding questionnaire item usually addresses a specific relation, for example by asking one of the following questions in a free recall design.

With whom do you …	relation
…talk about problems at work?	professional support
…talk about problems with your kids?	social support
…only exchange professional information?	instrumental

In addition to the mere existence of a relationship, a value indicating a quality or of frequency of interaction may be of interest. Instead of asking which actors in a focal network are connected to each other, researchers might want to find out how intense these relationships are using question such as the following in a free choice design.

How often do you talk to the following?	daily [3]	at least once a week [2]	once or twice per month [1]	less [0]
Alice	☐	☐	☐	☐
Bob	☐	☐	☐	☐
Charlie	☐	☐	☐	☐
⋮				

A most intricate, network-specific decision, however, is the way in which to select dyads that respondents are supposed to evaluate. If they are asked about their own ties, the following are typical design choices.

free choice	alteri are chosen from a given list
rating/ranking	actors from a given list have to be rank-ordered or rated on a given scale
free recall	alteri are chosen unrestrictedly

Each of these designs can be modified by restricting the number of answers sought.

Studies based on choice or rating/ranking need to compile a roster of eligible actors beforehand. This process, during which it is decided who can be a member of the network and who can not, is called *boundary specification* and outlined in more detail in the next section.

A prime example of clearly bounded networks is school classes. Network membership is given, and students can be queried on their friendship ties by asking "Who are your best friends in class?" (e.g., Knecht 2008). This free choice design becomes a *fixed* choice design when the item is rephrased, for instance, as "Who are your five best friends in class?" Fixing the number of answers may provide orientation to respondents as to how to interpret "best," but may also cause them to artificially inflate their response. In a *limited* choice design, the number of answers is not fixed, but an upper or lower bound is set. However, the principle problem remains.

Without a-priori criteria for network membership, the decision as to whom to include may be better left to the respondents. In complete-network studies this raises the question as to whether newly named actors shall also be questioned and if so, where to stop. In a *snowball sample* design, newly named actors are included up to a certain point. This is also addressed in the next section.

Box 13: Newcomb's Fraternity

Social psychologist Theodore M. Newcomb conducted a study on emerging friendship networks in the years 1954–1956 (Newcomb 1961; Nordlie 1958). Its setting is often referred to as "Newcomb's Fraternity." In two consecutive years, 17 male students each were selected from a group of applicants. The selected students were given free lodging in a fraternity. In return they had to participate in 4–5 hours of surveys and discussions every week. During the selection process it was made sure that none of these students had known any of the other students before the experiment. During both years data collection was carried out over the course of 16 weeks with a view to studying the process whereby the students became acquainted with each other. The data included attitudes, popularity, and estimations of attitudes of the other students. Each student had to rate each other student with respect to their "favorableness" once a week on a zero to hundred. They were not allowed to assign a the same score twice in the same week, and responses were actually coded as complete, ordered preference lists.

A classic example combining the above designs is the study by Rapoport and Horvath (1961), in which 861 students of two junior high schools were surveyed. The interest was in connectivity, defined as the fraction of the total population that can be contacted by tracing friendship choices from an arbitrary starting population of nine individuals. Thus the authors asked the students to rank-name their eight best friends in school. They then traced the acquaintance chains created by the student's choices. Note that, even though a roster of eligible actors was given implicitly by the school membership requirement, it was large enough for the design to be regarded as free recall. Recall, on the other hand, was then fixed to a choice of eight who, finally, have to be ranked. Since friends ranked seventh and eighth turned out to be failures significantly more often then those ranked first or second, fixing the number of nominations may have obliged participants to name more friends than they actually had.

An example of a rating/ranking design on the full roster is provided in Box 13.

It is part of the principle of ego-centered network studies that alteri are not known prior to the survey. Since free recall is characterized by a potential bias and recall errors (Borgatti and Molina 2003), a number of approaches to assist and direct free recall have been developed. They are described in detail in Section 3.2.3.

Ego-centered network studies are akin to population studies in that the data sought serves to characterize the focal actors and not a larger structure. When they are conducted in large, open populations, surveying alteri is generally not an option so that the egos must provide information about all relationships and the alteri themselves. Interviews rather than questionnaires can help to increase the willingness of participants, and enable the researcher to react to responses, for example by asking for clarification. On the other hand, this may also result in an interviewer bias.

Observation

The canonical means of passive data collection is observation without intrusion, although participatory observation is common as well.

Observation is particularly helpful when actors are difficult to survey, for example small children, individuals from a different culture, and some types of aggregate actors. Field observation is, therefore, the standard approach in anthropology, in which the interest is in observing regular social situations as they unfold, ideally, without interference and over long periods of time (Bernard 2011; McCurdy, Spradley, and Shandy 2005).

Direct observation and active recording of actions and interactions is increasingly complemented by analysis of digital traces. The exploding use of electronic means of communication is producing a wealth of observational data. Some of this is done purposefully with an analytical objective in mind, some is generated for unknown later use.

More generally, online transaction data such as phone calls, purchase orders, blog postings, wiki edits, or even just search engine use and website browsing are recorded routinely for business, quality control, customization, and marketing. The same is true for credit card payments, bonus card usage, reservations, joint purchase of multiple products, key card access, mobile phone game playing, TV and broadband usage, movement with location-aware devices, and many other everyday activities. At least partial observation of regular social activity has thus become a standard activity, albeit by companies and governments rather than ethnogra-

Box 14: Whyte's Street Corner Society

William F. Whyte, a pioneer in participatory observation, spent three and a half years living in a slum district, almost entirely inhabited by Italian immigrants. In his study "Street Corner Society" (Whyte 1998), he gives an account of his experiences and analyzes the organizational structures of slums and corner gangs. Asked why White chose this particular slum, he answered, "Cornerville best fitted my pictures of what a slum district should look like" (Whyte 1998: 283). "It had more people per acre living in it than any other section of the city. If a slum meant overcrowding, this was certainly it." (Whyte 1998: 283)

The study was time-consuming. That was partly because Whyte had very little background in community study and partly because the topics of interest "...depended upon an intimate familiarity with people and situations" (Whyte 1998: 357). Whyte joined in the local ways of life. For a time he lived with an Italian family, he participated in activities such as bowling, baseball, softball, and cards. He earned the confidence and friendship of the people and became a part of their community.

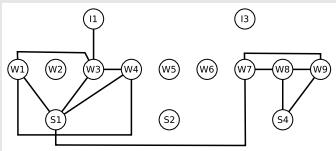
phers. For businesses and service providers, the statistical analysis of such data is a competitive necessity, and network analysis is part of the toolbox used. Some recorded data is released to the public or, often subject to disclosure conditions, researchers.

Some common differences to survey data and traditional observational data are worth pointing out. Recorded transaction data are very often vast but form an uncontrolled sample. Systematic and technology-induced biases, such as automatic transactions or spam, may require major data cleaning. More generally, the significance of a datum is typically a technical one. An SMS text message is a communication act but its precise content and significance is generally unknown. Similarly, the definition of a Facebook friend is immanent to the system. On the other hand, recorded data are often more convenient and cheaper to collect and there are fewer issues regarding recall, accuracy, or subjectivity.

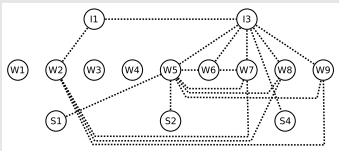
As with other forms of data collection there are many issues associated with electronically recorded data. We mention two that are of particular relevance for empirical studies. The first major issue is whether the available data is useful for the intended analysis. Because recording is of

Box 15: The Hawthorne Experiments

Roethlisberger and Dickson (1939) analyze the results of a series of inquiries into the effects of work and working conditions at Western Electric Company (Hawthorne Works). The experiences were divided into room studies, interviewing studies, and observational studies. The inspectors selected a vertical section of a department and "... placed the group to be studied in a separate room" (Roethlisberger and Dickson 1939: 387). A room study examined factors in the work place environment that affected worker fatigue. The interviewing studies identified worker attitudes. And through the observation of small work groups, the inspectors collected data about the social relationships and social structure within the group.



friendships



antagonisms

These two figures show friendships and antagonisms within the bank wiring observation room. The rooms consist of inspectors, wiremen, and soldermen. The first figure shows that friendships tend to cluster in two groups. The second figure shows that antagonism is mainly aimed at people without strong friendships.

limited flexibility, the kinds of observations that are feasible or available as secondary data may not yield data that is useful for the formulation and testing of a hypothesis of interest. The second major issue is ethical. In the majority of cases digital traces are produced by persons who are not aware that they are being observed. Obtaining consent may be challenging to organize and users who are aware of the recording may use a system differently.

A recent study that is highly interesting because it involves many of the aforementioned issues is Bakshy, Rosenn, Marlow, and Adamic (2012).

Box 16: Following and Being Followed

A relation called "following" on Twitter, a microblogging platform, is studied in Huberman, Romero, and Wu (2009). For each user of Twitter the researchers obtained for their data set the number of followers and followees (people followed by a user) the user has declared, along with the content and datestamp of all his/her posts. The data set consisted of a total of 309,740 users, who on average posted 255 posts, had 85 followers, and followed 80 other users. Among the 309,740 users only 211,024 posted at least twice. They call them the active users. They also define the active time of an active user by the time that has elapsed between his first and last post. On average, active users were active for 206 days. The researchers were interested to find out how many people each user communicates directly with through Twitter. Therefore they defined a user's friend as a person at whom the user has directed at least two posts. Using this definition they were able to find out how many friends each user has and compare this number with the number of followers and followees they declared.

Archives

Archival sources can be very useful in passive data collection. They are often the only means of obtaining data about units that are observable in principle, but inaccessible to the researcher. Examples include very large-scale and past networks. Archives are generally inexpensive alternatives, and can also be instrumental in plausibility checking and data completion (see for example Huberman et al. 2009).

Among the many types of archives are libraries, email and calendar entries, church registers, archaeological records, company reports, membership rosters, and all kinds of public, governmental, or commercial databases.

The best-known example of a historic network study is situated in Renaissance Florence (Padgett and Ansell 1993). With data from a multitude of historical records, it is argued that Cosimo de' Medici attained political control by being able to fill structural holes between a number of different political players.

Another example is a study of organizational affiliations of activists in the nineteenth-century women's movement in the USA (Rosenthal, Fin-

grudt, Ethier, Karant, and McDonald 1985). Similarly, more contemporary studies of interlocking directorates (e.g., Mintz and Schwartz 1985; Burt 1983) derive relations among banks and corporations from board membership records.

Bibliographic databases are an important source for studies in the sociology of science and, despite the many pitfalls, for research performance rankings.

3.2.2 Boundary Specification

Complete networks have a clearly defined boundary separating actors participating in dyads from everyone who does not. This boundary is specified before or during data collection and has been considered the problem of "where to set limits in the analysis of social networks that in reality do not have any obvious limits at all." (Barnes 1979: 414)

Depending on the scenario, there may well be self-evident boundaries. In studies interested in the social structure of bounded groups, such as school classes, business units, or service recipients with a unanimous membership criterion, the boundary is given. This does not imply, however, that members are easy to reach or that surveying or observing them is realistic. It may still be necessary to apply sampling.

In other scenarios it may be difficult to establish a criterion for membership in advance. By overestimating the scope of a network, it remains feasible to adjust an initial boundary by excluding members after data collection.

In surveys that use a choice design, at least, it is difficult to include actors once data collection is underway. Adjustments based on "the relative frequency of interaction, or intensity of ties among members as contrasted with non-members" are common (Wasserman and Faust 1994: 31).

Laumann, Marsden, and Prensky (1989) give a detailed account of the boundary specification problem. They distinguish between the criteria used (actor attributes, relation type, affiliation, or combinations thereof) and the perspective taken (realist or nominalist). Boundary specification from a realist perspective aligns with the perception of the actors. While they may not know each other, they do share the idea of belonging to a definite group such as the alumni of a particular university. Researchers take a nominalist perspective when membership criteria are not those normally applied by the actors themselves as in early adopters of a certain technology.

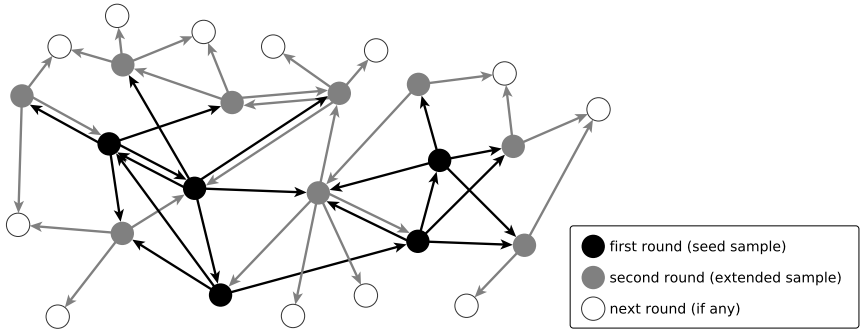


Figure 12: Two rounds of snowball sampling.

The above examples for initial boundary specification comprise *positional approaches*, in which membership is decided via attributes such as personal characteristics or affiliations with a group. An alternative are *relational approaches*, such as the boundary shrinking described above. The dominant method in this category is *snowball sampling*, during which additional actors are included in a survey if they are nominated in recall items. Criteria for both, that is the inclusion of new actors and stopping the snowball, must be established and clearly communicated. Figure 12 illustrates two waves of snowball sampling on a fictitious network.

Nominations of new actors can be solicited not only via tie-based recall but also by tapping the knowledge that respondents may have about the subject matter. For example, a questionnaire may include a name generating item such as "Which organizations participated in the Iran election protests?" Note that respondents need not be related to these organizations but can help with the implementation of a positional boundary specification.

Clearly, all strategies have advantages and disadvantages. Overly inclusive approaches increase the burden on data collection. However, overly exclusive approaches risk missing actors with an impact on the findings (Mizruchi and Marquis 2006).

Moreover, a researcher might decide to look at participation of specific actor in an incident to answer a research questions such as "Which organizations participated in the Iran election protests?" This event-based strategy is different from formal affiliations and memberships and can help you to identify relevant groups and subjects for your research project. The setting of network boundary is a special problem for complete-network

studies. Ego-network studies define the boundary during data collection for example with name generators. We discuss the different forms of data collection for complete and ego networks in the next part of this chapter.

3.2.3 Alter Recall

In a recall design, survey participants are asked to list ties with other actors that may be unknown to the researcher. This is almost always the case in ego-centered network data collection when egos are asked to list alteri but may also be relevant in relational approaches to boundary specification in other network studies.

The corresponding survey items are generally divided into *name generators* and *name interpreters*. With the help of name generators, the respondent is confronted with a specific relation and asked "with whom" he or she is related in this particular way (Diaz-Bone 1997: 52). A name generator defines the boundary of an egocentric network and is, therefore, of central importance for the entire analysis. Generators can be differentiated into *interpersonal* and *global generators* (Pfenning 1995: 46).

Interpersonal name generators relate to a context or a stimulus. Social contexts comprise the interaction spaces of ego and his or her network partners. These include immediate and extended family, the circle of friends, neighbors, colleagues, members of associations, and the broader circle of acquaintances. Individual persons from these social contexts are surveyed. "Crucial for the quality of such applications is the number of the surveyed social contexts" (Pfenning 1995: 47).

The advantage of this surveying method lies in the fact that it takes several spaces of social contact into account and captures most accurately the social environment of the surveyed subjects in terms of completeness. "The interpretation of the individual contexts [...] is problematic (Who is a friend? Who is an acquaintance?) as is the necessity to specify a generalized stimulus (e.g., With whom do you frequently have contact at work? Who are your best three friends?, etc.). This leads to the differentiation between functionally (i.e., also spatially) defined contacts (colleagues, neighbors, and family) and freely selected social relations (friends, acquaintances)." (Pfenning 1995: 47)

Instead of social contexts, stimulus-related interpersonal name generators use specific stimuli, such as shared leisure activities, discussion of important issues, giving of assistance, etc. This helps to avoid deficits and conceptual confusion in the use of social contexts. The biggest method-

ological problem here lies in the definition of the stimulus specifications. The question as to the centrality of social interactions must be answered in relation to this along with the questions "concerning the completeness of the surveying of the network persons, the fulfillment of the stimulus specifications in accordance with the interaction typology of social interconnectedness, the social web and social network." (Pfenning 1995: 47) It is particularly difficult to capture this information in the case of weak ties as they mostly represent merely situation-related, punctual, and, in many cases, one-off contacts. A long list of very specific interactions is needed to survey such contacts.

Global generators avoid the identification of alteri altogether. In contrast to the context-related name generators, questions about individual persons are not asked here within the specified contexts but "globally or generalized on the basis of the structure of the totality of the relevant social relations." (Pfenning 1995: 47) A survey of this kind could take the form of the ego being asked whether the majority of his or her friends are usually of the same opinion as he or she is, whether most of his or her friends know each other, etc. It is up to the ego him or herself "to carry out and specify a balancing of the individual dyads." (Pfenning 1995: 48) The advantage of this method lies in the almost complete capture of the spaces of social contact. The disadvantage lies in the fact that no dyad-related structural parameters can be calculated and the "personal union" is not controllable (cf. Pfenning 1995: 48). "This lack of selectivity and the balancing of the social relations by the surveyed subject renders this process a particularly subjective form of surveying." (Pfenning 1995: 48)

Once the egocentric network has been delineated, further information is obtained about the persons named by the ego, i.e., the alteri. These additional questions are called *name interpreters*. Generators and interpreters combined are referred to as a *network instrument*.

Name Generators

Burt (1984) developed a network instrument for the General Social Survey (GSS), a national random sample ($n = 1534$) in the USA. It starts with the following name generating question.

> "Q1. From time to time, most people discuss important personal matters with other people. Looking back the last six month – that would be back to last August – who are the people *with whom you discussed an important personal matter*?" (Burt 1984: 331, emphasis in source)

No limit on the number of alteri listed by ego is set. However, if fewer than five alteri are named, interviewers are supposed to probe "anyone else?"

Fourteen subsequent questions are then used as name interpreters. These are applied only to the five persons named first. This restriction is justified with reference to the time taken in the interview to survey the alter-alter relationships as their number grows quadratically with the number of alteri (Burt 1984: 314f).

Question Q2 records whether the ego feels very well acquainted with the alteri, and the persons, with whom the ego feels particularly well acquainted. Question Q3 asks whether the alteri see each other as strangers if they meet on the street, and Q4 asks whether the alteri consider themselves as very well acquainted with each other. Using the following name interpreters, additional information about the alteri is surveyed in relation to gender (Q5), age (Q12), religion (Q13), party preference (Q14), ethnicity (Q6), education (Q7), and income (Q15). Other interpreters specify the surveyed relationship on the basis of the duration of the relationship (Q9), the frequency of contact (Q8), the role assumed by the alter for the ego (Q10), and the topics discussed (Q11).

The use of an instrument for collecting egocentric network data with only one name generator tends to elicit small networks of core ties (Marsden 1987) but "many conceptual understandings of networks extend beyond core ties to include more mundane forms of social support" (Marsden 2005: 12).

An instrument with multiple name generators was designed for the Northern California Community Study (NCCS) (Fischer 1982). It was used in 1977/1978 in a sample stratified by community size in California ($n = 1050$). The Fischer instrument uses ten situation-related or stimulus-related questions on communicative interaction, practical support, and social activities.

The instrument's name generator (Fischer 1982: 315; McCallister and Fischer 1978: 137) includes persons,

1. whom the respondent would ask to look after their house when the respondent is away;
2. with whom the respondent speaks about his or her work;
3. who have helped the respondent with tasks in or around the house in the past three months;
4. with whom the respondent has eaten in the past three months or whom he or she has visited (or by whom he or she has been visited);

5. with whom the respondent sometimes speaks about shared leisure activities and hobbies;
6. with whom the unmarried respondent is in a relationship;
7. with whom the respondent speaks about personal matters;
8. whose advice is significant for the respondent in taking important decisions;
9. from whom the respondent would borrow money if he or she needed it (if this does not mean that he or she takes out a loan or uses savings);
10. who live in the respondent's household as adults.

Alter-alter relationships are surveyed for up to five alteri. Generators (1), (4), (5), and (7)–(9) are used here for the compilation of a list of these alteri. The first persons surveyed by these generators are listed in each case. Persons from the respondent's household and persons already listed are passed over (Fischer 1982: 332). The ego is then asked whether the listed persons know each other well. The age-age relationships are recorded by the Fisher instrument using this interpreter. At the end of the interview, a second and more comprehensive list is compiled of all persons named by the informant in the course of the interview for all ten generators. The respondent is then asked whether the list is complete or whether a person who is important to him or her is missing. Using name interpreters, the role relationship of the alter for the ego (the specified roles are: relation, colleague, neighbor, friend, acquaintance, other) and the gender of the alter are surveyed for all of the persons on the list. The intimacy of the relationship, the distance between homes (up to five minutes by car, more than one hour away), the availability of a meeting place (café, park or other) that can be reached within five minutes, the similarities between the ego and the alteri in terms of profession, ethnicity or nationality, religion, and the (leisure) activities pursued are also recorded. As is the case with the Burt instrument, the relationships between the alteri are also surveyed using the Fischer instrument for five people named by the ego. Because the structure of the surveyed egocentric network is only known for a partial network, this partial network is also referred to as a "small Fischer network."

Position Generators

Rather than identifying particular alteri using name generators, a position generator measures linkages to specific locations. This instrument has proven particularly useful for investigations of the productivity of general individual social capital, i.e., social capital research about general populations that does not focus on a particular life domain.

Lin (2001) assumes that access resources arise through embeddedness in a social structure that can be mobilized in a targeted way. He describes these resources as social capital. There are two methods for measuring such resources: first, the use of interpersonal name generators and, second, the position generator.

Lin states that name generators are insufficient as, by definition, they are linked to content unless there is information available about the population or content of the universe (roles, familiarity, geography, etc.). Hence, Lin proposes a different measurement instrument which he developed: the position generator. According to Lin, the position generator measures the access to network members via their professional position, which is understood as an existing social resource based on the prestige of the job in question within a hierarchically organized social structure. This instrument is not exactly easy to use but it is efficient. Using this instrument, different basic entities can be studied in a differentiated way. It is based on a clear theoretical foundation (distribution of prestige which can be referred back to, highest access prestige, and the number of accesses to different positions). In addition to positions, respondents identify relationship types (family, friend, acquaintance) for each accessed position.

Lin, Fu, and Hsung (2001) applied the position generator shown in Figure 13 to studies carried out in Taiwan. They also asked (Lin et al. 2001: 69) about the number of daily contacts

> In an ordinary day, how many people are you roughly in contact with?
> ☐ 0–4 ☐ 5–9 ☐ 10–19 ☐ 20–49 ☐ 100 or more

and how well the respondents knew these persons ("1. Know almost all of them; 2. Know most of them; 3. About half and half; 4. Don't know most of them; 5. Know almost none of them"). A value was created from these data which indicated that the higher the value, the less familiar the respondent is with his daily contacts. This corresponds to Granovetter's theory of the strength of weak ties, i.e., that extensive, less familiar contacts enlarge the networks and provide better access to social capital.

Resource Generators

Another instrument for the measurement of individual social capital involves tapping resources rather than context or positions (van der Gaag and Snijders 2005). This instrument also differs from other measurement instruments for social capital; it does not include the mapping of an egocentered network (as with the use of name generators), instead it bears

Q1. Among your relatives, friends, or acquaintances, are there people who have the following jobs?
Q2. If so, what his/her relationship to you?
Q3. If you don't know anyone with these jobs, and if you need to find such a person for private help or to ask about same problems, who among those you know would you go through to fin such a person? Who would he/she be to you?
Q4. What jobs does he/she do?

(see Q1–Q4 above)	Q1 1. Yes 2. No	Q2 see list below	Q3 see list below	Q4 see list below
a. High school lecturer				
b. Electrician				
c. Owner of small factory/firm				
d. Nurse				
e. Assemblymen/women at provincial or city/county level				
f. Truck driver				
g. Physician				
h. Manager of large factory/firm				
i. Police (regular policeman)				
j. Head of division, county/city government				
k. Housemaid or cleaning worker				
l. Reporter				
m. Owner of big factory/firm				
n. Lawyer				
o. Office workman or guard				

Figure 13: Generator based on positions (Lin 2001).

more similarity to the position generator (Lin and Dumin 1986). It differs from the position generator, in particular, by referring directly to accessed social resources rather than occupational prestige. Whereas the position generator can be used to measure access to social resources useful in instrumental actions, the information retrieved by the resource generator can refer more clearly to social resources useful in expressive actions.

The structure of the resource generator is the same as that of the position generator and shown in Figure 14. However, a defined list of resources is used here which encompass the concrete sub-resources of social capital and hence render the respondents' access to resources visible. As with the position generator, in addition to the resources, the strength of a relationship, relationship type (family member, friend, or acquaintance),

Do you know anyone who…	Acquaintance	Friend	Family member
1. can repair a car, bike etc.	☐	☐	☐
2. owns a car	☐	☐	☐
3. is handy repairing household equipment	☐	☐	☐
4. can speak an write a foreign language	☐	☐	☐
5. can work with a personal computer	☐	☐	☐
6. can play an instrument	☐	☐	☐
7. has knowledge of literature	☐	☐	☐
8. has senior high school (VWO) education	☐	☐	☐
9. has higher vocational (HBO) education	☐	☐	☐
10. reads a professional journal	☐	☐	☐
11. is active in a political party	☐	☐	☐
12. owns shares for at least Dfl. 10,000	☐	☐	☐
13. works at the town hall	☐	☐	☐
14. earns more than Dfl: 5,000 monthly	☐	☐	☐
15. owns a holiday home abroad	☐	☐	☐
16. sometimes has the opportunity to hire people	☐	☐	☐
17. knows a lot about governmental regulations	☐	☐	☐
18. has good contacts with a newspaper, radio or TV station	☐	☐	☐
19. knows about soccer	☐	☐	☐
20. has knowledge about financial matters (taxes, subsidies)	☐	☐	☐
21. can find a holiday job for a family member	☐	☐	☐
22. can give advice concerning a conflict at work	☐	☐	☐
23. can help when moving house (packing, lifting)	☐	☐	☐
24. can help with small jobs around the house	☐	☐	☐
25. can do you shopping when you are ill	☐	☐	☐
26. can give medical advice, when you are dissatisfied with your doctor	☐	☐	☐
27. can lend you a large sum of money (Dfl. 10,000)	☐	☐	☐
28. can provide a place to stay if you have to leave your house temporarily	☐	☐	☐
29. can give advice concerning a conflict with family members	☐	☐	☐
30. can discuss what political party you are going to vote for	☐	☐	☐
31. can give advice on matters law	☐	☐	☐
32. can give a good reference when you are applying for a job	☐	☐	☐
33. can babysit for your children	☐	☐	☐

Figure 14: Generator based on resources (van der Gaag and Snijders 2005).

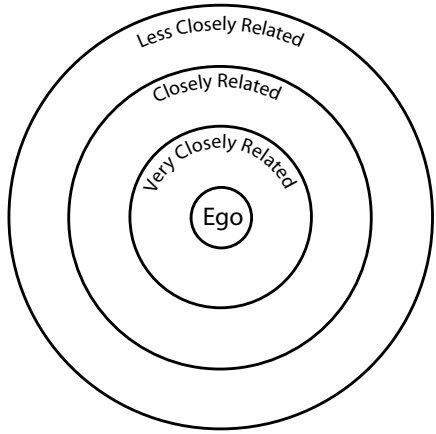


Figure 15: Network card (Kahn and Antonucci 1980).

etc. are also recorded. In this case also, it is not possible to eliminate incompatibility problems arising from its use in different contexts as the list of specific resources makes a series of variation possibilities in different contexts possible. Hence the composition of the instrument depends on the systematic, theoretical consideration which represent social resources, i.e., the "general social capital" of individuals.

Network Maps

Kahn and Antonucci (1980) describe procedures for identifying providers of social support, using both affective and role relation name generators and concentric circle diagrams for listing network members in relation to a respondent.

This instrument is particularly suited to the recording of the emotional proximity and importance of persons in a network. Like the other generators discussed in this section, it is also particularly suited for surveying egocentric networks, in which little is known about the field in which the respondents operate, e.g., immigrants. The respondent is presented with the diagram shown in Figure 15, which consists of three concentric circles. Ego, the respondent, is made the focal actor in the innermost circle. The respondents are asked to enter names or initials of those persons, to whom they feel emotionally connected or who are important to them personally in the three circles. The persons to whom they feel closest and who are most important to them are entered in the innermost circle. De-

	incomplete	inaccurate	inconsistent
causes	non-response data loss ⋮	poor design error ⋮	perception intention ⋮
effect	missing information	misleading information	contradictory or complementary information

Figure 16: Data quality issues.

pending on the information required, attribute data, such as age, gender, nature of relationship (family, friend, colleague, etc.), duration of the relationship, frequency of contact, etc. can then be surveyed for all of the persons entered in the circles.

3.3 Quality Issues

Data will never be perfect. In most cases it will not even describe exactly what is needed to answer a research question but only related phenomena that are observed or represented more easily. Even if it did provide such answers, lack of access, measurement error, transcription, and numerous other factors will inevitably cause quality impairments.

Hence, it is important to be aware of the types of quality issues that you may encounter. Knowing about them and some of their possible sources is more helpful than a series of dos and don'ts as it will help you to maintain a critical attitude towards the significance of patterns in the data and computational results.

Instead of going through a list of possible sources of error, or how to avoid them, we have organized this section according to the type of problem and its consequences. An overview is provided in Figure 16.

Incomplete Data

Missing data (Little and Rubin 1989; Allison 2001; Little and Schenker 1995) is always a nuisance because, by definition, an attempt was made to collect the data but it was not available to the researcher. Possible causes of missing network data (Kossinets 2006) include the misspecification of

network boundaries, inaccessibility, non-response, drop-out (in longitudinal studies), data loss, and many more. Of course, you should always try to obtain data as completely as possible; if this were not necessary, you would be looking for the wrong data in the first place.

Whether missing data affect your conclusions depends on

- how much data are missing,
- which data are missing, and
- how they became missing.

For instance, a study hypothesizing that the degree of social support received is negatively related to income may fail to find an association if high-income respondents are less likely to report it. This is an example of a systematic bias and called *missing at random* because it is conceivable that other attributes, such as level of education, are also related to income, and may, therefore, be helpful in *imputing* plausible values for the missing ones without distorting the results. Since priors were drafted in a lottery, for instance, the three values missing in Figure 7 could be imputed quite reasonably by distributing the number of seats not covered by the other families according to the number of family members eligible at the times.

Imputation of network data (Robins, Pattison, and Woolcock 2004) has not yet been applied extensively although there is a growing body of literature on predicting unobserved or future relationships (Liben-Nowell and Kleinberg 2007).

One extreme case in which there is no underlying systematic bias is called *missing completely at random*. In population samples this can often be dealt with by ignoring missing data because the assumption implies that the only consequence is a reduced sample size. While this approach may also apply to ego network studies, it is almost certain to fail in complete-network studies because of the inherent dependencies of overlapping dyads and the frequent use of indirect relations in analyses. Even if some studies suggest concrete numbers, such as an 80 percent response rate, above which, for example, the identification of the most central actors becomes reliable, we recommend that you do not apply them. At best, they are rules of thumb that are useful across a large number of studies with incomplete data sets of similar type. Your study is likely to be dealing with only a few networks, and most network-analytic techniques are very sensitive to even small changes in the data. Simply compare a ring-like structure and a chain of links: A single missing relationship makes all the difference for the actors at the ends of the chain.

At the other extreme, the bias is again systematic, however, dependencies are not with respect to any of the observed data. This is called *missing not at random* and the toughest problem to deal with. The potential for this to arise can at least be reduced sometimes by introducing additional attributes into the design of a study.

Inaccurate Data

Data are *inaccurate* if they are not missing but comprised of values that deviate from the intended ones. Sources of inaccuracy include measurement error, informant lack of recall, intentional lying, and many others.

Active data collection methods are highly dependent on the willingness and capacity of respondents to provide accurate information. Inaccuracy is particularly common when respondents have a different understanding of the definition of a relationship. This once again highlights the need for the precise and comprehensible formulation of questionnaire items. Are you sure that the respondents understand what they are supposed to answer? Do they have appropriate options? Can they reasonably be expected to know this information? Even in the given situation?

While in most cases we know when values are missing, it is difficult to assess when they are accurate (Butts 2003). Inaccuracy may be systematic or random. Data generation processes that consistently produce the same information (possibly from different angles) are called *reliable*, irrespective of whether the information in question is accurate or subject to systematic error. If the information is generally accurate but perturbed by random error, the process is referred to as *valid*.

Hence, both reliability and validity are necessary to obtain accurate data. The implementation of data collection methods according to these criteria may depend on the analytic interest, however. For example, Pfenning (1995) compares the reliability of Burt and Fisher's instruments, which are described above, with respect to the properties of the resulting ego networks. Because of the larger number of stimuli, Fischer's instrument generates more consistent alter lists. In a test-retest approach, 63 percent of alteri were named in both tests using Fischer's instrument, compared to 45 percent for Burt's. While this resulted in a more reliable heterogeneity indicator, a size indicator varied more strongly than with Burt's instrument.

Inconsistent Data

We say that data are *inconsistent* if variables supposed to represent the same information have different values. This includes the case in which one of them is missing. Sources of inconsistent data include repeated or alternative measurements of the same property, the mismatch of a set of particular values and their supposed aggregate, impossible values, and many more.

Bernard, Killworth, and Sailer (1980; 1982); Bernard, Johnson, Killworth, Kronenfeld, and Sailer (1985) observed social interactions within several different groups (students, managers, radio amateurs, a group of blind people, and employees of a social research institute) and then asked members of these groups about their relationships. As it turns out, just about half of the ties named by the respondents matched the observed ones. These studies have inspired further research with similar questions (Hammer 1985; Freeman and Romney 1987; Freeman, Romney, and Freeman 1987). An important conclusion for network research is that the recall of social interactions appears to be governed more by an underlying cognitive structure than by the actual short-term interactions. As a consequence, observed interactions may not be indicative of the more stable, longer-term relations in a group.

Another common case of inconsistent data is encountered when actors are asked to self-report relationships that exist between them. Since every relationship is assessed twice, once by each actor in the dyad, this is bound to lead to disagreement, even for relations that may appear indisputable to a researcher. If it is not possible to repeat the questioning, the data is frequently recoded by aggregating the two responses into one. For non-valued relations, the two main options are to treat an unconfirmed relationship as either

- absent (*minimum symmetrization*) or
- present (*maximum symmetrization*).

The term *symmetrization* derives from undirected relations in which responses can be seen as an asymmetric generalization of the relation. Neither is particularly satisfying because we either lose or impute data based on the understanding that there is a true situation that at least one of the respondents describes inaccurately. It may be more appropriate, however, to embrace the inconsistency and analyze whether the disagreement is actually informative in terms of the different perceptions of a relationship that actors may have.

An extreme case of this view underlies the concept of cognitive social structures. Actors assess every relationship not necessarily to arrive at a better consensus value but to investigate different cognitive models and how they relate to network behavior.

3.4 Ethical Considerations

Ethical considerations play an important role in all stages of a network study.

For example, during the early research design phase, you should consider the degree of intrusion imposed on the subjects while collecting network data from them: Indirect observation of subjects' behavior is less intrusive than interviews and surveys directly collected from the subjects, however, it may also be perceived as voyeuristic.

In the later stages of a network study, data storage and the protection of privacy should be considered carefully. Decisions regarding the way, in which findings are reported and possible implications discussed with the subjects need to be made towards the end of the project.

Many research institutions require researchers to comply with their internal ethics guidelines for research involving human subjects. While we recognize that there are differing international practices, we would like to highlight the importance of the elements that constitute the institutional review process before researchers start to collect and analyze network data.

First, researchers need to inform their subjects that the questions are not designed to cause harm to their position when they participate in the research study. This form of human subject protection is called informed consent, and most internal review boards require subjects to sign a form confirming that they understand the potential risks of revealing their network relationships. The process is necessary to ensure that individual rights are protected and that the study is ethically appropriate and may proceed (Klovdahl 2005). Borgatti and Molina (2003) highlights the importance of subject protection so that researchers do not compromise the successful approval of future network studies.

One of the main tensions in network research is that network data cannot be anonymous: Instead, both the respondents and their contact names must be known to the researcher to enable the collection of data

and conduct of the analysis. In addition, the respondents and their nominated contacts must be identifiable to match attributes, such as gender, political affiliations, etc. to them and their relationships. This means that network data cannot be independent and this creates significant responsibilities for researchers (Breiger 2005).

Another ethical consideration already addressed earlier in this chapter concerns decisions about the inclusion or exclusion of network actors. As soon as the researcher imposes an artificial boundary around the perceived network, he or she includes or excludes actors that may or may not have to be included. Connected to this is the decision on how to handle non-respondents: Are they still considered to be part of the network because they were mentioned by other actors in the network? Or, do they need to be removed from the data analysis because they themselves opted out of participating in the data collection? Kadushin (2005) suggests that every researcher must make the best possible assessment as to how the network data may be used, particularly in the case of military or public health networks.

When all of the data have been collected, you will be confronted with decisions on how to handle the data and protect the confidentiality of the responses and anonymity of subjects (Borgatti and Molina 2005).

One way to ensure the confidentiality of the responses is the assignment of id numbers to each subject (which we recommend anyway, see above) and removal of all other identifiers from the data. This way individuals are less likely to be identified, although it must be considered that their specific role or position in the social structure may expose them any way. This stresses once again that data on overlapping dyads may be heavily interdependent.

Another way to ensure subject anonymity is to display and publish only aggregated network data so that roles and positions cannot be identified even by those who know the context or are represented in it. We also recommend that this practice be adopted when researchers discuss their findings with potential clients. We consider this an important practice to protect participants from facing potentially negative ramifications.

A more comprehensive overview of the question of ethics may be obtained from the special issue of the journal *Social Networks* entitled "Ethical dilemmas in social network research."[1]

1 *Social Networks 27* (2), 2005.

3.5 Summary

A prerequisite for empirical hypothesis testing is the availability of suitable data. Network studies generally involve data for two types of observational units, actors, and dyads. The latter are the main characteristic of network studies. Because data defined on dyads is indexed by two actors rather than one, data tables need to be adapted accordingly.

We have distinguished primary data collected by yourself from secondary data that are already available. Secondary data are mostly collected by previous studies or extracted from databases and archives. Typical strategies for primary data collection are surveys (questionnaires, interviews, etc.) and observation (field observation, electronic transaction recording, etc.). In general, surveying is a form of active data collection in which a response is elicited by providing stimuli such as questionnaire items, whil observing is a form of passive data collection in which regular behavior is recorded.

Ideally, the decision for any of these strategies depends exclusively on the data that would concur with your research question. In many cases, a compromise must be made because the desired data may be impossible or too costly to obtain.

For primary data collection, the crucial decision concerns which actors to include. In ego-centered network studies this amounts to deciding on the population of focal actors and a sampling strategy. It is thus akin to other population sampling approaches. For complete networks, however, the population coincides with the actors of the network. It must, therefore, be bounded more restrictively, either by defining the boundary explicitly (and possibly narrowing it later on) or by expanding the network during data collection as in snowball sampling.

We also discussed several instruments for network surveys in detail. These use various techniques for assisting respondents to recall actors, to whom they are related in a given way. While they are primarily used in ego-centered network studies, similar techniques are useful in snowball sampling.

As usual, actor and tie attributes may require cleaning, filtering, normalization, and other kinds of preparatory transformation. While this may be inherent in the design of the data collection process, it may also be triggered by data quality issues. In this context, we differentiated between incomplete (missing), inaccurate (unreliable or invalid), and inconsistent (inaccurate or more fine-grained) data. A special transformation

is anonymization which is merely one technique for dealing with ethical issues such as privacy.

3.6 Exercises

1. For a network study on status and reputation of developers in an open-source software project, assume that you can obtain data on the assignment of developers to reported bugs.
   - What are the units of observation?
   - Is this active or passive data collection?
   - Are these primary or secondary data?
   - How would you organize the data? Why?
2. Explain the difference between a monadic and a dyadic variable to a friend or colleague.
3. Given a network of 1024 actors with 4211 ties, how many cells does the actor-by-actor table have? Compare this to the number of entries in a tie-by-attribute table and an actor-by-ties list.
4. What defines the boundary of a network?
5. Mobile phone service providers are interested in understanding which conditions prompt customers to terminate their calling plans (i.e., in predicting churn). It is assumed that calling patterns and churn of important calling partners are relevant indicators.
   - Outline a data collection strategy for a personal-network survey.
   - What if you can gain access to a provider's contract and call data?
6. Suppose you are planning to survey candidates for committee positions in a political party regarding their allies and competitors among all party members.
   - Formulate survey items for these relations.
   - How are you going to determine network boundaries?
   - Are there ethical considerations to address?
7. Describe and compare the data that is obtained from interviews using network maps or resource generators.

8. Discuss the differences between face-to-face interviews and online surveys. What are implications of these differences for the data obtained via one of the instruments described in this chapter?
9. Invent a scenario in which inconsistent data is more informative than data corrected for consistency.