Chapter 2

Collecting Network Data

2.1 Introduction

Compared to the collection of other types of data in the social sciences (e.g., classic survey data) the collection of social network data can be quite challenging. A major threat to validity in social network research stems from problems of missing data that are due to a number of different sources at a number of different stages in the research process. One major contributor to missing data is non-response in network surveys. Although non-response bias, and missing data more generally, is a concern in the collection of data of any type, it can be particularly vexing in the whole network case (Borgatti, Carley and Krackhardt 2006; Stork and Richards 1992). There are four primary ways in which missing data and subsequent error can enter into social network research (Kossinets 2005). First, missing data can enter into the picture if the network boundaries are not properly specified on theoretical or other grounds (Laumann, Marsden and Prensky 1983). Second, network surveys are extremely susceptible to non-response bias

in that missing actors and their links can affect structural and analytical outcomes at both the network and individual levels (Borgatti, Carley and Krackhardt 2006; Kossinets 2005). Respondents can refuse participation, can refuse to answer some or all network survey questions due to such things as lack of time, interviewee burden or question sensitivity and may drop out of a longitudinal study prematurely as a result. Third, the design of the study and subsequent sample or instrument design (e.g., types and forms of relational questions) for a given social network problem and context can also be important in limiting threats to validity (Bailey and Marsden 1999). Finally, issues of respondent reliability and accuracy have clearly been shown to produce errors of various kinds (Bernard, Killworth and Sailer 1984 ; Freeman, Romney and Freeman 1987).

Thus, we need to be aware of factors that minimize threats to validity in the collection of social network data, particularly in the case of complete networks. In this chapter we will look at data collection methods and related matters with regard to a number of both theoretical and practical issues concerning our ability to minimize such threats primarily in the whole or complete network context. The chapter is not a comprehensive treatment of all the possible ways to address validity concerns in the collection of network data but, rather, is meant to provide a general awareness of such problems and suggest some possible solutions. As they say, for warned is for armed!

2.2 Complete Versus Personal or Ego Networks

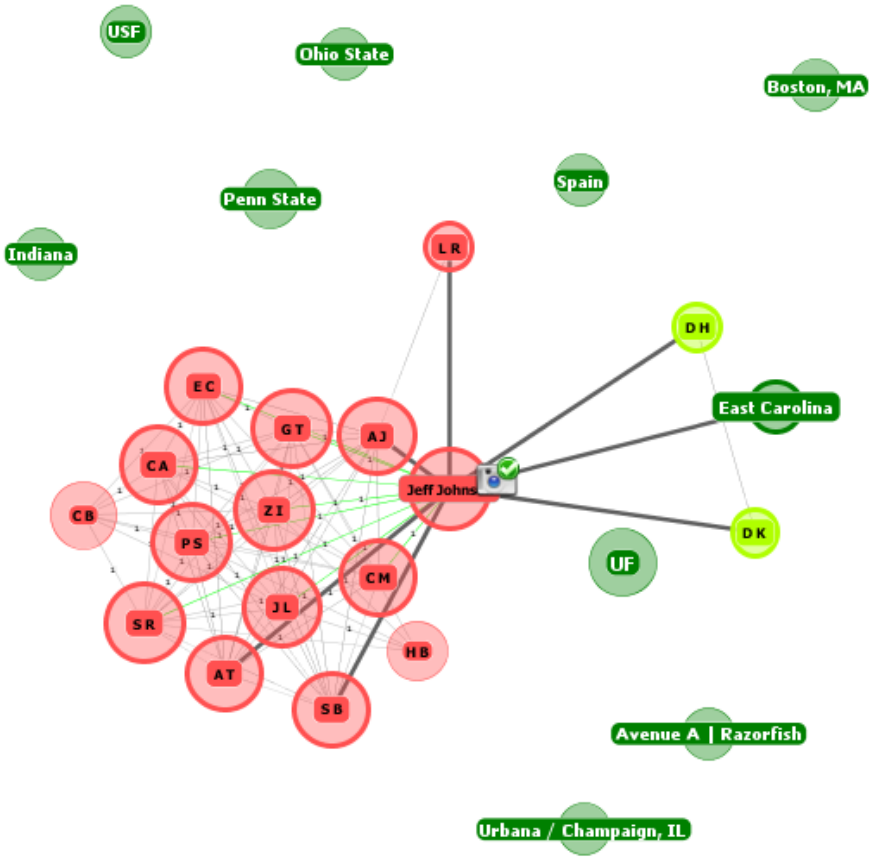
As discussed in the introduction, we can generally think of any network (e.g., food web, worldwide web, protein network, Facebook network) in two primary ways. The first is what we might call whole or complete networks in which we look at the relations among

all actors in the network however defined. Here all the dyadic relations among a bounded set of actors are documented and measured in some manner. The second is what we might call personal or ego networks in which we are interested only in sample of reported direct ties between an actor (referred to as ego) and some set of other actors (often called alters) however defined (e.g., friends, people who provided aid during a hurricane).

A simple way to think about this can be found among the many social networking sites on the internet. Let's take the networking internet site Facebook for example. If you (ego) have a page on Facebook somewhere on the page you list your personal interests (an attribute of ego) and the various groups to which you have some affiliation (which can also be thought of as a kind of network that will be discussed later). In addition, there is a list of "friends" (alters) with pictures and their online names. This page taken alone represents an ego or personal network, for you, in which we have information about you, ego (page owner), and all your Facebook relations here defined as "friends" (alters). The ties between ego and alters are either present or absent, what we call binary ties, but we can't really tell anything about the strength of any given relationship or tie other than it exists. If we were to take all the Facebook pages and look at all the dyadic binary relationships among all the "friends" we would be looking at a whole or complete network consisting of the entire Facebook "friends" network. Of course, we could look at only subsets of this whole network, say, by looking at only the network of people who consider themselves avid NASCAR fans or the network among those Facebook members who live in Boston or who are connected to East Carolina University. The manner in which one places boundaries on this whole network will depend on the purposes of the research. Figure 2.1 is an example of a Facebook 2-mode ego network for one of the

authors. This shows how an ego's alters are connected to one another and to other entities such as countries, universities, corporations, organizations, etc.

Figure2.1. 2-Mode ego network in Facebook for one of the authors using the Touch-Graph option.



One Facebook page or personal network probably wouldn't tell us much about the theoretical interesting aspects of Facebook or human behavior for that matter. For that we would probably want to take a sample (probably random) and look and the characteristics of egos and their alters. Let's say we have a hypothesis that women's personal networks

are more expansive, diverse and dense then men's personal networks. Although labor intensive, we could take a random sample of 200 pages and look at the attributes of ego's alters, the ties among ego's alters and the sheer size of ego's personal network. This would allow us, for example, to look at the homogeneity of ego nets (similarity of characteristics like political ideology or gender), the density of ego nets (i.e., density or the number of ties observed among the alters compared to the total number possible ties) and the number of alters (size) on average for both men and women.

If we really wanted to work hard we could look at the entire Facebook "friends" network and look at the exact same phenomena. We could, in fact, apply the same type of analyses (e.g., density, homogeneity) for the whole network study as we did with the ego network study to investigate differences between men and women. However, the collection of whole network data would allow for a much more expansive set of structural analyses at both the whole network level and the individual actor level as compared to the ego network approach. But then again, it would be a lot more work.

The type of approach one chooses (i.e., whole vs. ego) will depend on the purposes of the study. The personal network method has become particularly popular in the social sciences given that it fits well within a survey approach (Lin 2002; Wellman 2007). Personal network questions can be readily added to a standard survey instrument so that both traditional attribute data (e.g., education, income) can be collected along with individual based ego network data and then generally compared, for example, in some form of linear modeling. Recognizing the similarities among the types of relations and measures used in the two approaches, we will nevertheless treat them separately in our discussion on data collection methods and issues.

Despite the differences in levels of focus between these two primary approaches, the nature of questions reflecting dyadic relations and types of network measures used in either approach are strikingly similar. For example, degree centrality is the same measure of degree centrality whether one uses a complete or ego network approach. Differences may exist in how the data is managed and eventually analyzed (linear regression in the ego network case versus exponential random graph models in the complete network case), but there is a high degree of commonality in the nature of social relations investigated and the network measures of interest.

2.3 Sources of Social Network Data

Network data can come from a variety of different sources but it generally boils down to a distinction between primary versus secondary types of data. Secondary sources are those that already exist somewhere in print (e.g., fish exchange records, historical marriage records) or can be found electronically (e.g., Enron emails, Social Networking pages). It is not that these sources of network data don't require work to collect, collate and put into some format that allows for their analysis, they are secondary because they do not require any direct interviewing or observations with network members. Primary sources are primary because they involve the direct action either actively, passively or unobtrusively with the actors themselves. This, of course, has implications for the kinds of relations and their measurement. Secondary data by its historical and/or fixed nature dictates and limits the type of relations and levels of measurement that can be used in the course of the research. Primary data collection allows a greater deal of flexibility in the

type, measurement and number of relations to be studied. Some of the more computer-based automated forms of network data collection represent a transitional form between primary and secondary data. Although the data is collected directly, as in primary research, there are limitations on the types of relations available for study, as in secondary research.

2.4 Interviewee Response and Types of Relation

As noted in the introductory section of the book, networks can be made up of relations of various kinds involving a variety of different types of entities. We can have networks of people, proteins, predators and their prey, organizations, countries, terrorists, and so on. Such entities can have relations that involve sharing, the flows of material goods or energy, interactions, feelings, co-memberships and so on. Such relations can be simply present or absent or may take on some value representing the extent or degree to which any two entities have a relation. Some types of relations may be directed in that the relation goes in one direction but not necessarily the other, while yet for other types of relations direction has little or no meaning. Finally, relations for those interviewed can be emotionally neutral or emotionally charged. Thus, the type of network relation one measures will depend on the entities involved, the nature of the relation of interest to the researcher and the relation of network actors to the relations of interest. Table 2A, B and C , for example, show the types of relations relevant to people, organizations and countries. Some of these types of relations may evoke little emotional response on the part of actors as in kinship relations or may elicit much more emotion as in network questions eliciting affective relations such as who one likes or trusts or interaction

questions such as who one has had sex with. Thus, the type of relations studied and, as we shall see, the network context can impact the nature of responses on the part of actors.

Table 2A_. Types of Relations Among People.

Kinship • Mother of, Wife of, Son of	Affective • Likes, Trusts, Fond of
Non-kin Role-Based • Boss of, Professor of, Supervisor of • Friend of, Acquaintance of	Interactions • Hangs Out With, Talks to, Gives Advice to • Has Sex With, Shares Needle With, lends money to
Cognitive/Perceptual • Knows • Aware of What They Know	Affiliations • Belong to Same Clubs • Sits at the Same Table at Lunch

Table 2B_. Types of Relations Among Organizations.

Corporate Entities • Buy From/Sell to, Leases to, Outsources to • Owns Shares of, Subsidiary of • Joint Ventures With, Cooperate With, Sales Agreements With, Alliances With • Regulates, Controls	Via Corporate Members • Personnel Flows • Interlocking Directorates • Personal Friendships • Co-Memberships

Table 2C_. Types of Relations Among Countries.

Geopolitical Entities	Via Representatives or Citizens
• Trade with/Exports to, Imports from, Outsources to • Has Foreign Investment in • Joint Military Ventures With, Has Treaties with, Has Diplomatic Relations with, ,Has Economic Agreements With, Has Political and Military Alliances With • Regulates, Controls, Protects, Administrates, Occupies • Has Conflict With	• Migrant Flows or Diaspora • Interlocking Marriages • Visits Among Leaders • Co-Memberships in International Organizations (e.g., OPEC) • Student Exchange Programs

In actual research we must make a distinction between the theoretical and practical aspects of collecting social network data. The guiding theory underlying the network study may suggest the need to collect data of a specific kind among a specific set of actors. However, it is often the case that practically there may be problems in actor's abilities or willingness to meet these more theoretical needs. Or in the case of longitudinal network studies, the need to sustain actor participation throughout the research may involve many compromises in terms of the types of relations sought, the number of total questions asked and form of the network data collected. Depending on the context, some types of relational questions are more sensitive than others and this respondent sensitivity can impact interviewee's willingness to answer questions or participate in the research all together. Further, such sensitivity can vary by cultural context (e.g., economic relations may be more sensitive in some cultures than others), can vary over time (e.g., some relational questions may be of a more sensitive nature at the

beginning of a longitudinal study than towards the end) and may vary as a function of the data collection methods employed (e.g., face-to-face versus online interviews).

A good example of this issue comes from the work of Johnson, Boster and Palinkas (2003) in their study of the network dynamics at polar research stations. In the beginning of a four year study the researchers were initially interested in the formation of friendships and the ability of individuals to assess potential friendships within a short period of group formation in an attempt to understand the extent to which first impressions play out in the long run in terms of network evolution. One of the researchers attended the first training exercise of the first winter-over crew preparing to deploy to the geographic South Pole. During a break in training the crew members were given a questionnaire asking them to rank the other members of the crew from 1 to n-1 in terms of their guess about the likelihood of forming friendships with each person on the list over the coming winter. Immediately, several of the crew began to grumble and protest and one crew member threw down his pencil and walked out of the room. This resistance to the administered network question was related to two primary problems. It had been decided on theoretical grounds to use a full rank order to measure potential friendship, as opposed to other measures such as a likert scale, due mostly to the psychometric advantages of fully ranked data (Eudey, Johnson and Schade 1996). Further, the researchers were interested in affective relations in the station such that friendship seemed to be the best type of relation to study. However, two unanticipated factors almost sidetracked the entire research enterprise, an enterprise that involved a longitudinal design and required sustained cooperation on the part of crew members throughout the 9 months of the austral winter and in two subsequent winter-over crews.

First, it was discovered that the initial period of group formation was filled with great optimism (i.e., a utopian stage) where there was a general perception that everyone would get along and be friends over the course of the austral winter. The task of having people rank order one another in terms of potential friendship created quite a negative emotional response on the part of crew members since they believed at this point in the group formation process that "everyone" would be friends and ranking people meant that some people would be ranked near the bottom of people's list therefore implying a possible lack of friendship. Thus, both the type of relation and how it was measured (i.e., rank order), although theoretically sound and justified, was practically problematic. The mix of a rank order collection method and actor's judgments as to expectations of friendship fostered a "perfect storm" of sorts in terms of sensitivity and interviewee burden. Eventually the researcher met with the crew at an "all hands meeting" and there was a joint discussion and compromise to ask crew about "who one interacts with socially" and to measure it on an eleven point likert scale (0 to 10). The relational question and the method of measurement was ultimately determined in concert with those being studied. Ironically, this created a sense of investment in the design of the study on the part of the crew and helped foster an extremely high and sustained response rate over the winter.

In terms of the compromise in the type of relation elicited, further research found that asking people about who they interacted with socially achieved the same thing as asking directly about friendship (i.e., highly inter-correlated) without asking an emotionally charged question. The scale measurement was a much simpler and less daunting task and reduced respondent burden that was critical for sustaining member's

participation over the winter. In addition, the scales could be ranked and achieve the same result as a full rank ordering of the data thereby achieving theoretical needs without risking non-response on the part of network actors (see Eudey, Johnson and Schade 1996).

It needs to be pointed out that whereas people were hypersensitive about assessing friendship in the initial stages of group formation, they had fewer problems discussing and expressing their feelings about social relations of both the negative and positive kind towards the end of the study year. This reflects a temporal component in understanding question sensitivity and its ultimate impact on potential non-response bias. Thus, the maturity and other characteristics (e.g., cultural context) of the network itself may have an impact on the level of emotional reaction to one or a given set of network questions.

This is a good example of the potential impacts of network question sensitivity in that respondents often feel self conscious about reporting on relations with others, particularly those questions that get at affective relations such as who an actor likes or trusts or, even more so, who a person dislikes or distrusts. Some of these problems can be minimized by the method of questionnaire administration. Self administered questionnaires may, as opposed to face-to-face interviews, limit these feelings, but as seen in the example above, some questions evoke strong emotion that not even the form of survey administration can overcome. In addition, different data collection methods are better and worse at dealing with issues of study reliability and validity and in reducing potential non-response bias and in limiting sources of missing data.

2.5 Data Collection and Limiting Non-response Bias

As stated earlier whole network approaches are extremely sensitive to missing data (Borgatti, Carley and Krackhardt 2006). This is particularly true for smaller networks where the absence of actors or ties can have relatively large effects. The manner in which we collect network data can have a profound impact on actor participation and on reliability and validity of the social network data sought. There are no hard and fast rules about which form of data collection is best, but there are a number of trade-offs depending on the data collection method or methods employed.

There is a considerable literature on factors contributing to non-response bias in surveys of various forms (e.g., face-to-face, phone, mail out). These factors include survey fatigue, instrument burden, respondent skepticism as to study utility, issues of privacy and confidentiality, lack of respondent motivation, lack of time and lack of interest due to questionnaire salience. Some or all of these may play a role and different data collection methods and their implementation are better or worse at improving response rates.

Table 2.1 provides examples of some of the ways in which researchers have typically collected network data. The columns in the table represent a few of the trade-offs one should consider in the course of considering a data collection method in a network study. As we have seen from the polar research station network example above, some network questions may be more emotionally sensitive than others. Self-administered network surveys, including mail out and on line surveys, may limit the degree of self-consciousness on the part of respondents. In addition, such survey approaches suffer less from interviewer response effects (although contacting and gaining access to respondents may still be impacted by such effects). Finally, these data

collection methods are certainly the easiest to administer compared to the others. However, this is where any advantages end.

An important means for reducing non-response on the part of actors concerns the building of rapport (Johnson 1990). The problem with some of self-administered approaches, particularly mail out and on line surveys, is the limited ability to establish contact and create a relationship, no matter how minor, with respondents. Making a connection with respondents is critical in increasing response rates no matter the collection method used. Many of the solutions to solving non-response bias in surveys, particularly mail out and on line surveys, involves using multiple contact methods to enhance response rates. Dillman (1978) provides guidelines for overcoming some of the disadvantages of the mail out and phone surveys in terms of increasing response rates. However, face-to-face collection provides the greatest opportunity for establishing rapport with respondents. Additionally it facilitates the use of elicitation interviewing techniques for the collection of network data (Brewer 2000, Johnson and Weller 2002). Network elicitation is difficult to do in a less interactive context and limited in phone and group interview formats.

Some studies have advocated that low response rates in surveys are due less to potential respondent's resistance to participation and more a result of researcher's inability to gain access or actually find or track down respondents (Sosdian and Sharp 1980). Finding, tracking down, or contacting potential respondents can certainly be an issue. These issues have become ever more problematic for some methods, like phone surveys, where people may have been overwhelmed by telemarketers and the like (i.e., suffering from survey fatigue). With such technologies as caller id, people can now

monitor calls and choose not to answer the phone. Johnson (1990), for example, found that using respondents to call ahead to their listed alters in a snowball sample limited this problem, particularly in population of older adults who were suspicious of strangers.

Comparative research has shown that the different survey approaches vary in response or return rates. However, such studies have also found that differences in response rates can vary depending on the social, organizational or cultural context. For example, whereas in a comparison of different survey approaches mail out surveys win out in one context they may just as readily loose out to other methods, such as on line surveys, in yet other contexts. The point here is that the data collection method you choose should be sensitive to the given cultural and social context in which one plans to work (Church 2001). In addition, we advocate making as much contact with potential respondents as possible independent of the type of survey approach. In fact, the more one can engage in ethnographic on the ground efforts the better chances for higher response rates in network surveys.

Table 2.1 Forms of data collection and their features.

Form of Data Collection/Interview	Issues of Sensitivity	Interviewer Response Effects	Ability to Establish Rapport	Thoroughness (Ability for Elicitation)	Ease of Administration
Face-to-Face	Moderate	Moderate	Moderate-High	High	Low-Moderate
Self-Administered	Low	Low	Low	Low	Moderate
Mail Out	Low	Low	Low	Low	High
On Line	Low	Low	Low	Low	High
Phone	Moderate	Low-Moderate	Low-Moderate	Moderate	Moderate
Group Setting	Low-Moderate	Low-Moderate	Moderate	Low-Moderate	Moderate

2.6 Determining Network Boundaries in Whole Networks

Determining the boundaries of a social network is driven by both data collection constraints and the theoretical focus of the study. We will refer to these as boundaries determined on the basis of some theoretical a priori factors, or theory driven boundaries, and those based on exploratory, emergent properties of the network data itself (e.g., including perceptions on the part of actors), or data driven boundaries. Social network data may, for example, be only one of a variety of forms of data to be collected during the course of a study (e.g., demographics, attitudinal). In addition, there may be subsequent network data collection tasks that become increasingly burdensome to respondents as the size of the network increases, this is true for both ego network and whole network approaches. However, boundary specification problems are primarily a problem within whole network studies. Unless one is engaged in a fully exploratory enterprise it is usually theoretical factors that will determine the criteria for bounding a social network. If we are interested in "group" dynamics, for example, than the boundaries of the network must relate to the criteria for defining just what the group is. Often times such boundary specifications are easy and straight forward such as bounding the networks of members of a Monastery (Sampson), a karate club (Erickson), a prison (Bernard and Killworth), an environmental program (Johnson and Parks), a fast food restaurant (Krackhardt), an ocean-going research vessel (Bernard and Killworth), a polar research station (Johnson et al.), etc. It is not that people don't have ties outside the given group it is just that the researcher chooses to limit ties to those are members of that group for both practical and

theoretical reasons. There are plenty of examples of these well-specified networks in the literature and the UCINET datasets are replete with such bounded networks.

In yet other cases, however, network boundaries may be fuzzy, unknown or difficult to determine due to such things as potential network member's spatial and geographical dispersion, lack of a priori knowledge of who belongs, the covertness of potential members and, simply, sheer size. In such cases there may be no real clear boundaries as in the previous examples, but there may theoretically or methodologically informed cut points where it just makes sense to draw the line. Such cut points or boundaries will be based on some set of relational criteria in terms of such things as tie intensity or density, the number of ties for a given actor or some other relational factor, set of attributes or activities on the part of actors.

Laumann, Marsden and Pensky (1983) make a somewhat similar distinction in their discussion on network boundary specification in terms of nominalist and realist strategies for bounding social networks. In the realist approach it is the actors themselves that are the prime mover for determining network boundaries. In this case, subjective meaningfulness on the part of actors is what guides inclusion rules. In the nominalist approach, it is the researcher that imposes criteria for bounding the network on the basis of some conceptual or theoretical framework that he or she deems important. It is interesting that these two approaches are reminiscent of the emic and etic distinctions in anthropology reflecting culturally specified versus culturally neutral means for understanding culture and society.

In many ways there are analogies between boundary specification problems in network data collection and the presence of a sampling frame in survey research. The one

thing that all the group network examples above have in common is the existence of an a priori or an easily obtainable list of actors (i.e., people, countries). This is similar to having a sampling frame in survey research in that the universe of actors is known in advance. As such, the researcher can construct data collection instruments that allow actors to respond to every member in the group in terms of different forms of social relations (e.g., friendship, communication) using such means as checklists, ratings, rankings and matrices. Such lists can be obtained on the basis of theoretical (nominalist) or subjective (realist) criteria. In the former, an a priori list of actors in theoretically determined roles in an organization, for example, might be used to place boundaries on a network. In the more subjective case, the actors themselves might be used to determine members viewed as being a part of the network.

However, it is often the case in both survey research and network studies that a sampling frame or a list of network members is not available a priori. Thus, some strategy must be developed for the initial selection of respondents or actors who are generally unknown in advance.

Survey samples may employ frames at other levels of aggregation such as in multistage cluster samples (e.g., lists of churches to sample parishioners) to help make units of analysis known. Or they may use other means such as census tracks to randomly select households for the sample. Network studies can use the same strategies to begin to find the unknown but this is where the similarities end, at least in terms of whole network research. Whereas survey research is interested in a representative sample from a population (i.e., representative subset of the population), whole network studies are seeking to discover the entirety of a bounded network population for a given theoretical

problem. In other words, we can't just talk to a subset of actors but we must talk to all the actors in the network as defined by the study. Thus we must make a distinction between a whole network survey and a survey sample. In some sense we refer to this as a sample in the network case but we are still attempting to discover as many unknown network members possible, not just a representative subset as in ego network approaches used in combination with surveys. Whole networks in these cases are whole networks, albeit often with some fuzziness. Although this analogy in sampling approaches may help in getting started, it does little to help us in knowing when to call it quits in terms of bounding the network.

2.6.1 Strategies for Bounding Whole Networks

As stated earlier, determining network boundaries is a matter of both theoretical and methodological factors. Some networks have easily definable boundaries and their members can be readily identified and listed. However, in many networks no prior lists of members exist making boundaries difficult to determine. In such cases, network boundaries may be drawn based on quotas, some specified criterion (e.g., degree > 1) or based on levels of network saturation (i.e., redundancy, density) during the course of some respondent driven sampling scheme (e.g., snowball sampling). In yet other cases, network boundaries may be determined by methodological constraints such as those imposed by collecting cognitive networks that relate to the data response burdens placed on actors.

2.6.1.1 Quotas

In an example of the use of quotas, Johnson (1990) studied seafood consumption in a small Midwestern town in the United States. He was interested in how webs of interaction at different social class levels influenced individual actor's perceptions concerning various types of meat (e.g., beef, poultry, pork, seafood) and how they are processed (e.g., fresh, frozen, canned). The basic theoretical issue driving the research concerned the extent to which people who are connected by social ties at different social class levels share cognitive models of "kinds of meat". In interviews with city planners the general social class levels of various neighborhoods in the town were determined. A random seed in each of two neighborhoods, one upper middle class and one lower middle class was chosen. A snowball sample from each of the seeds was then conducted until 15 households from each of the seed snowball samples were achieved for a total of 30 households. The two class networks could then be compared in terms of differences in cognitive models concerning meats.

The number of households interviewed from each seed sample, of course, could have been greater than in the example above. The size of any particular quota will depend on the theoretical problem at hand. The example above only needed two relatively small socio-economically distinct sets of network related actors to compare in terms of their judged similarity of items in a cognitive domain. However, other research problems in this same community may have required a significantly larger quota sample from each of the original seeds. For example, an understanding of the nature of network ties that bridge social classes in this small community would have required many more waves of interviews in the course of the snowball sample.

2.6.1.2 Criterion for Inclusion

Using inclusion and exclusion rules can help bound a social network. Thus, as in the previous example the researcher could have chosen to interview and include only actors within the same geographically defined neighborhood. Thus, only actors within the seed neighborhoods described above could have been included. There could also have been criteria on the strength of ties for determining inclusion and exclusion criteria or some combination of both tie strength and geographical location. The point here is that these rules would be determined by the theoretical problem at hand and possibly by other constraints as well (e.g., financial, logistical).

Most small group studies are in fact good examples of social network boundaries determined on the basis of some criteria. Membership in a class room, participation in a karate club, and women in a Southern town who attended a set of social events are all example of studies in which some condition or set of conditions established network boundaries whether attributional, relational or event based. It is not the members of a class don't have ties to others outside the class or that members of the karate club don't have ties to people who don't belong or don't do karate. Rather, the research problem determines where the boundaries will be drawn given theory in combination with practical and logistical constraints. Sure, the ultimate validity of the study depends on the proper rules for inclusion and exclusion, but it is not necessarily true that the validity of a network study can be enhanced by increasing the number of network members included. Actor relevancy trumps network size in research on social networks!

2.6.1.3 Saturation and Redundancy

It is often the case that reasonably clear network boundaries emerge during the course of the research. This is particularly true when utilizing snowball samples or other respondent driven sampling methods. In a study of communication networks in a fishery in the Southeastern United States, Maiolo and Johnson (1988) used key informant freelists (Borgatti 1994) and commercial license lists to identify an initial set of seeds for a snowball sample. Although a commercial license list existed for commercial fishers and commercial dealers there was no such list for sportfishers who targeted king mackerel both an important commercial and sport species. In addition, just because someone was a commercial fisher did not mean they targeted king mackerel. Key informants known to target king mackerel in both the commercial and recreational sectors were asked to freelist fishers they knew that regularly targeted that species. This list provided a seed list from which to begin the snowball sample of actor's "who talked to each other about king mackerel fishing". However, the problem was where to draw the boundary around the communication network. During the course of the snowball sample, which was conducted by both phone and face-to-face interviews, there were periods of time when there was considerable sample saturation or name redundancy in the elicitation of alters. This saturation represented fuzzy boundaries around communication networks that were often related to geographical factors. Thus, boundaries were placed on the basis of tie intensity and redundancy in the course of the snowball sample. However, it should be noted that if the purpose of the study is to discover the nature of ties that connect

various areas of high redundancy or density in social networks then ties bridging these areas of high density need to be pursued and the redundancy criteria may need to be applied across several waves.

One might have noticed that many of the examples provided above involved the incorporation of a number of different strategies in the course of determining network boundaries, producing what we might think of as hybrid strategies for boundary specification. Theoretical criteria, for example, may be used initially in determining boundaries but may be adjusted or fine tuned based on subjective information gained from the actors themselves. An example of this comes from Johnson (1986) in his study of the diffusion of innovations through a network of commercial fishers. Initially Johnson used the commercial license list obtained from the North Carolina Division of Marine Fisheries to identify commercial license holders in a small fishing community in North Carolina. He could have used the list as the boundary for the network (nominalist), but he was interested in active fishers as perceived by the fishers themselves (realist) and the list included anyone who commercial fished no matter the extent (e.g., part-time fishers). Using the names from the list he wrote them on cards and initiated interviews with fishers in the community asking them to sort the names into piles according to how similar they perceived the fishers to be to one another. This unconstrained pile sort led was aggregated into a matrix of judged similarities and scaled as shown in Figure 2.2.

Based on the pile sort task it was clear that there were differences among the various license holders based on amount of income due to commercial fishing. Those to the left of the configuration were all perceived as fulltime fishers while those to the right were viewed as part time. The final set of actors used for the network survey included

only fulltime fishers identified in the analysis. Thus, the inherent dimensionality in individual actor's perceptions of one another was used to specify network boundaries. As other work has shown, dimensionality in social networks is an important element in understanding network structure (Freeman 1983) and dynamics (Peli and Bruggeman 2005).

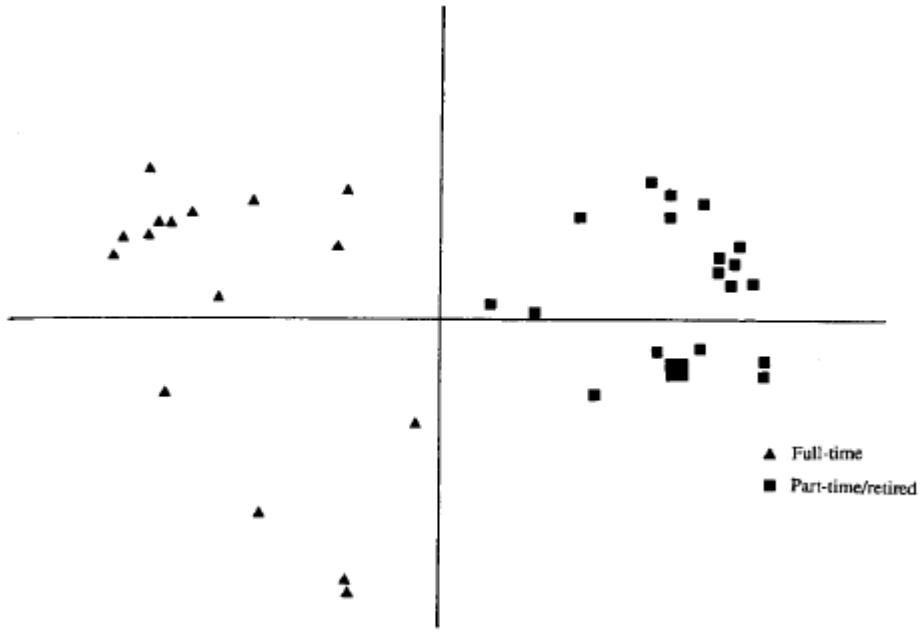


Figure 2.2. MDS of commercial fishers aggregated judged similarity of all commercial license holders in a small fishing community (from Johnson 1990).

2.7 Interviewee Burden

It may be that the size and particular boundaries for a network may be dictated by the methods employed. Some data collection methods are labor intensive and burdensome where such burden varies as a function of network size. A good example of this is

cognitive networks where people need to report on the network connections of all other actors in the network (Krackhardt 1987, Kumbasar et al.1994, Johnson and Orbach 2002). Thus, the bigger the network the more the respondent burden and the greater the need to bound the network to some reasonable size. There are a number of good examples of the use of cognitive networks in social network research and they generally involve the study of small well-bounded networks (see for example Kumbasar et al. and Krackhardt). However, how can the needs of studying a larger social collective be balanced with the methodological constraints of collecting cognitive network data?

In a study by Johnson and Orbach (2002) on political networks and the passing of a piece of environmental legislation there were potentially over 400 actors in the political network involving legislators, staff, resources managers, lobbyists and private citizens. The researchers were interested in the relationship between knowledge of the political landscape and political power. The problem was that political network had to limited to a reasonable size so that cognitive network data could be collected (i.e., each actor had to report on the ties of all the other actors in the network). This was potentially problematic given the status of the people to be interviewed and their time constraints (e.g., the President Pro Tem of the North Carolina Senate, cabinet level Secretaries, legislative committee chairs and co-chairs).

The network needed to be limited to no more than approximately 50 actors so as to limit the burden placed on interviewees (e.g., assuming reciprocated ties an unconstrained choice approach would involve 1200 dyadic choices on the part of the respondent). They used interviews with 10 knowledgeable key informants (Johnson 1990) of the legislative process to free list actors who were viewed to be “important” in

the development and passing of this piece of legislation. The top 45 names listed by the key informants were used to bound the network. For the cognitive network data collection the respondents were limited in the number of choices in that a fixed choice methodology of 3 was used. Actors were asked to list the three people on the list that each of the people talked to most about a given piece of legislation. This reduced the task to approximately 135 dyadic choices which was a much more reasonable task. As we shall see in the chapter on ego networks this is also a very important consideration in designing ego network surveys.

Interviewee burden, more generally, can lead to various kinds of non-response on the part of actors. There is plenty of literature discussing these issues in survey research and these apply here as well (Dillman 1977, Church 2001). However, unlike typical survey research where researchers are willing to accept at least some level of non-response bias as a matter of course, in whole network surveys such levels are totally unacceptable and provide real threats to the validity of any study. As we have seen from earlier examples, one of the objectives in maximizing response is not to anger or frustrate respondents. One potential source of respondent frustration and anger is the length of the interview itself, particularly if respondents feel time constraints. A major reason people state for their unwillingness to participate in surveys often concerns being "too busy" or lack of motivation (Sosdian and Sharp 1980). Network interviews and the complexity of certain social network methods can place both huge temporal and cognitive demands on respondents.

There are no hard and fast rules about how long is too long for an interview or how demanding is too demanding. However, the shorter the network survey instrument

the better, particularly if one is engaged in a longitudinal study where sustained participation is crucial. One rule of thumb for achieving an optimally sized network survey instrument is to include only those questions that are theoretically critical for the study at hand, no more or no less. If one is uncertain about the theoretical relevancy of a network question, conduct exploratory or ethnographic research to find out. Again, conducting ethnographic work prior to conducting a network survey can help in assuring the reliability and validity of network questions and in understanding the capacity of respondents to answer instruments of a given size (e.g., CEOs of companies and fishers in Cuba may face different time constraints).

2.8 Data Collection and Issues of Reliability

When we ask a network relational question we hope actor's responses are reliable and accurate. Even under the best of circumstances the nature and form of a response to a relational question can be highly variable in terms of reliability and accuracy. It is not that actors are intentionally lying or being evasive (although that can happen), but that certain forms of questioning are cognitively challenging to respondents even under the best of circumstances. Bernard, Killworth and Sailer (Killworth and Bernard 1976, 1979; Bernard and Killworth 1977; Bernard, Killworth and Sailer 1979,1982) in a series of papers showed that actors reports of temporally discreet behavioral interactions with others were suspect in terms of accuracy. This series of studies compared who people said they interacted with during a given period with who they actually interacted with and found high levels of errors in actor's reports of their behaviors in, for example, a fraternity, among ham radio operators and among deaf people communicating via

teletype. So if one asked “who were the people you talked to yesterday at the fraternity house?” there were generally a number of omission and commission errors in informant’s reports of who the interacted with. These findings are not all that surprising given a general problem with informant accuracy in retrospective data more generally (Bernard, Killworth, Sailer and Kronenfeld 1984).

Nevertheless, a number of other social network studies were conducted to better understand the nature of these problems with informant’s reports of their interactive behavior. Romney and Faust (1982) in a reanalysis of some of the BKS data (tech data) found that the accuracy of an informant depended on the extent of interaction with other network members. The more an actor interacted with other members of the group the greater the accuracy. Romney and Weller (1984) did a reanalysis of four of the BKS data sets and found that informant accuracy was even more a function of informant reliability or the an actor’s correlation to the aggregate response.

In a creative investigation of this problem Freeman, Romney and Freeman (1987) observed attendance at a symposium series in a university during one academic quarter. Following the final symposium of the quarter, attendees were asked to recall all the individuals that had attended. Not surprisingly there were inaccuracies in attendee’s reports of who attended. However, these inaccuracies were patterned in that omission errors tended to include those individuals who attended the final symposium but were mostly absent at the others, while commission errors tended to be individuals who usually attended the symposiums but happened to not be there for the final talk. Thus, individual recall reflected the long term or normative patterns of attendance and not the actual attendance at a single event. This is important in that if we are interested in

peoples reporting on normative or patterned network phenomena they are pretty good at it and we should adjust and construct are network questions accordingly.

The reliability of informant's or respondent's responses to network questions then can be influenced by a number of factors. As discussed above the nature of people's cognitive abilities to recall is an important consideration in constructing relational questions in surveys. Additionally, question sequencing or question embeddedness can imbue a range of interpretations as to the meaning of the relational questions that follow. Bailey and Marsden (1999) discuss this with regard to the name generator of who one "discusses important matters with" as used in the General Social Survey. They found that some individuals had difficulty articulating just what constituted "important matters" and that this was influenced by the sequencing of the questions. The nature of the preceding questions often influenced respondent's interpretations of just what constituted "important matters".

We must once again stress that many of these issues of questionnaire design and implementation important in minimizing non-response bias and increasing reliability and validity can best be resolved through the use of extensive ethnographic exploratory work. Only through a sound understanding of the network context can one hope to collect solid social network data.