3
Handling Relational Data

Social network analysis emerged as a set of methods for the analysis of social structures, methods that specifically allow an investigation of the *relational* aspects of these structures. The use of these methods, therefore, depends on the availability of relational rather than attribute data. In this Chapter I will show how these relational data can be collected, stored and prepared for social network analysis. Many of the general considerations that arise in handling relational data are not specific to this type of research. They are those that arise with all social science data: gaining access, designing questionnaires, drawing samples, dealing with non-response, storing data on computers and so on. These issues are adequately covered in the many general and specialist texts on research methods, and it is not necessary to cover the same ground here. However, a number of specific problems do arise when research concerns relational data. As these problems are not, in general, covered in the existing texts on research methods, it is important to review them here before going on to consider the techniques of social network analysis themselves.

The Organization of Relational Data

All social research data, once collected, must be held in some kind of **data matrix** (Galtung, 1967), a framework in which the raw or coded data can be organized in a more or less efficient way. At its simplest, a data matrix is a table of figures, a pattern of rows and columns drawn on paper. When the data set is large or complex, the data matrix may need to be stored on record cards or in a computer file. Whatever the physical form taken, the logical structure of the data matrix is always that of a table. In variable analysis, attribute data can be organized in a case-by-variable matrix. Each case studied (for example, each respondent) is represented by a row in the matrix, while the columns refer to the variables on which their attributes are measured. Figure 3.1 shows a simple form of such a data matrix, with illustrative variables. This is the way in which data are organized, on paper or in a computer, for most standard statistical procedures.

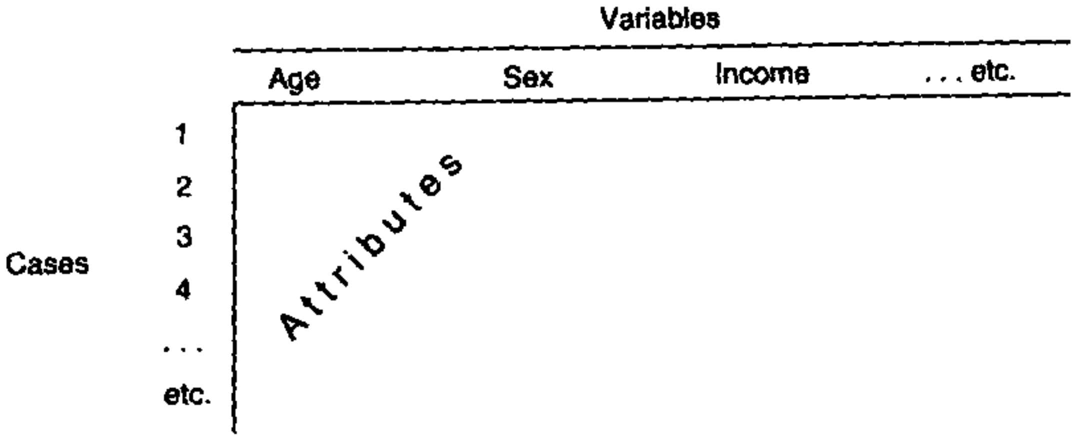


Figure 3.1 *A data matrix for variable analysis*

The case-by-variable data matrix cannot be used for relational data. These data must, instead, be seen in terms of a case-by-affiliation matrix. The cases are still the particular agents that form the units of analysis, but the affiliations are the organizations, events, or activities in which these agents are involved. The columns of the matrix, then, refer to the affiliations in terms of which the involvements, memberships, or participations of the agents can be identified. From this case-by-affiliation matrix can be derived information on the direct and indirect connections among the agents. In Figure 3.2, for example, a simple case-by-affiliation matrix is shown for the involvement of three people (labelled 1, 2 and 3) in three events (labelled A, B and C). Where a specific individual participates in a particular event, there is a '1' in the corresponding cell of the matrix; non-participation is shown by a '0' entry. It can be seen that all three people participate in event A, but none of them is involved in events B or C. Thus, the sociogram that can be drawn from this

		Affiliations		
		A	B	C
Cases	1	1	0	0
	2	1	0	0
	3	1	0	0

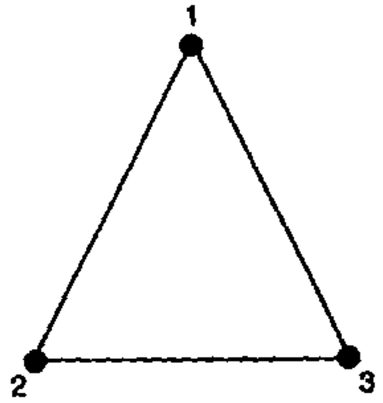


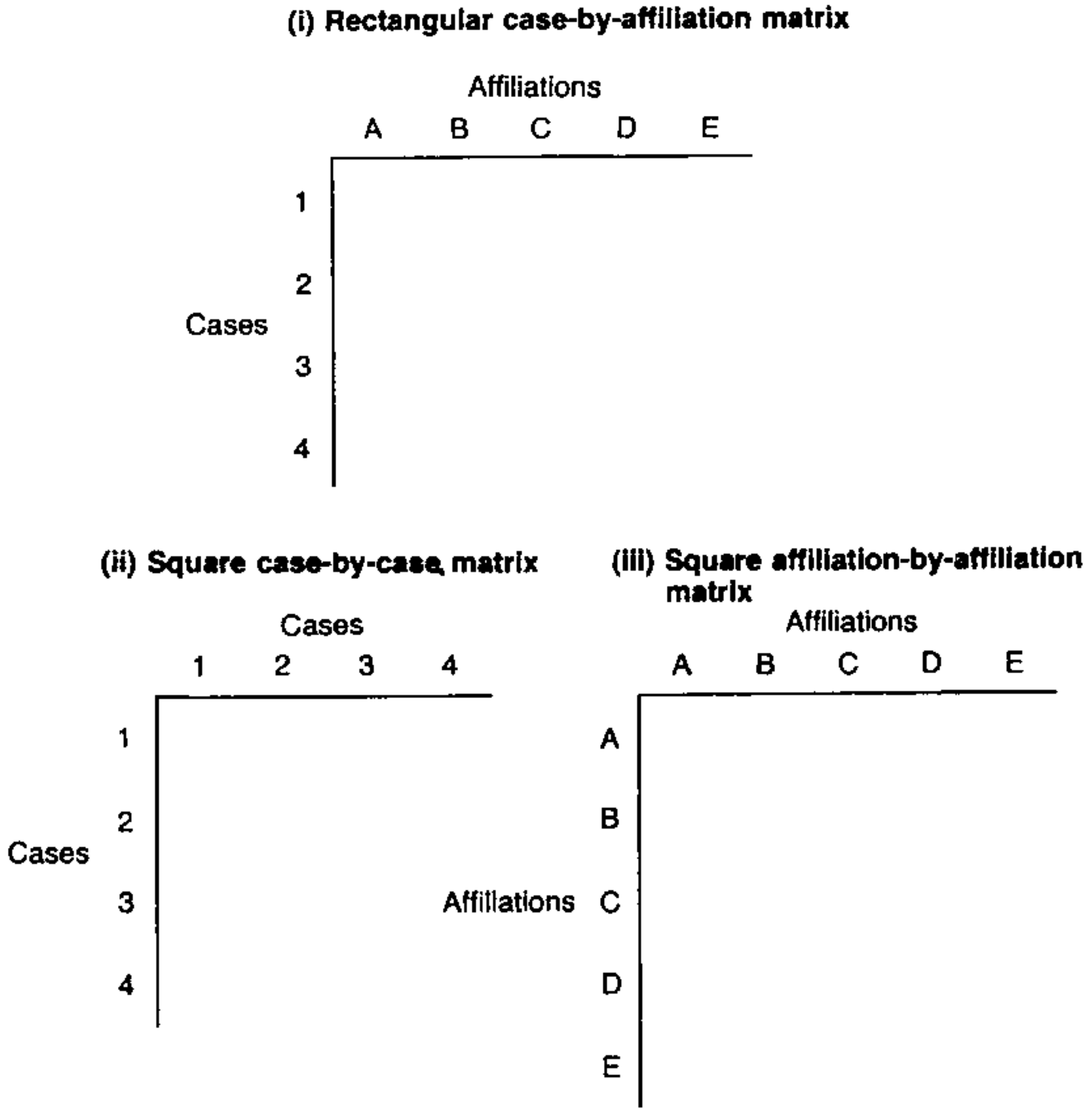
Figure 3.2 *A simple matrix and sociogram*

matrix shows a simple triad of mutual contacts among the individuals. The sociogram can be read as saying that each person meets the other two at a particular event.

It can be quite difficult to construct sociograms for even moderate-sized data sets. Lines will criss-cross one another at all sorts of angles to form a thicket of connections, and any visual appreciation of the structure is lost. Indeed, it may be quite impossible, using conventional manual methods of drawing, to construct a sociogram for a large network. For this reason, social network analysts have attempted to find alternative ways of recording the connections. Following the principle of the data matrix, the solution that has been most widely adopted has been to construct a case-by-case matrix in which each agent is listed twice – once in the rows and once in the columns. The presence or absence of connections between pairs of agents is represented by '1' or '0' entries in the appropriate cells of the matrix. This idea is not, perhaps, as immediately comprehensible as the sociogram, and so it is worthwhile spelling it out at greater length.

Figure 3.3 shows the general form of data matrices for social networks. The most general form for raw or coded data is what I have called the case-by-affiliation matrix, in which agents are shown in the rows and their affiliations in the columns. Such a matrix is described as being two-mode or 'rectangular', because the rows and columns refer to different sets of data. For this reason, the numbers of rows and of columns in the matrix are generally different.[1] From this basic rectangular data matrix can be derived *two* square, or one-mode matrices. In the case-by-case matrix both the rows and the columns will represent the cases, and the individual cells will show whether or not particular pairs of individuals are related through a common affiliation. This matrix, therefore, shows the actual relations or ties among the agents. It is exactly equivalent to the sociogram in the information that it contains. The second square matrix shows affiliations in both its rows and its columns, with the individual cells showing whether particular pairs of affiliations are linked through common agents. This matrix, the affiliation-by-affiliation matrix, is extremely important in network analysis and can often throw light on important aspects of the social structure that are not apparent from the case-by-case matrix.

Thus, a single rectangular matrix of two-mode data can be transformed into two square matrices of one-mode data.[2] One of the square matrices describes the rows of the original matrix and the other describes its columns. Nothing is added to the original data, the production of the two square matrices is a simple transformation of it. The rectangular matrix and the two square matrices are


(i) Rectangular case-by-affiliation matrix

(ii) Square case-by-case matrix

(iii) Square affiliation-by-affiliation matrix

Figure 3.3 *Matrices for social networks*

equivalent ways of representing the same relational data. In social network analysis the rectangular matrix is generally termed an 'incidence' matrix, while the square matrices are termed 'adjacency' matrices. These terms derive from graph theory, and they will be explained more fully in the following chapter. For the moment, it is sufficient merely to know the names, as they are the most generally used terms for relational data matrices. Most techniques of network analysis involve the direct manipulation of adjacency matrices, and so involve a prior conversion of the original incidence matrix into its two constituent adjacency matrices. It is critically important, therefore, that researchers understand the form of their data (whether it is incidence or adjacency data) and the assumptions that underpin particular procedures of network analysis.

In those situations where a researcher collects two-mode data on cases and their affiliations, then, it will generally be most appropriate to organize this information into an incidence matrix from which the adjacency matrices used in network analysis can later be derived. In some situations, however, it will be possible for a researcher to collect relational data in a direct case-by-case form. This would be the situation with, for example, friendship choices made within a small group. In this situation of what is called direct sociometric choice data, the information can be immediately organized in an adjacency matrix. Without entering into all the complications, there is, in this situation, no corresponding incidence matrix and no complementary adjacency matrix of affiliations. The reason for this, of course, is that all the agents have merely a single affiliation in common – the fact of having chosen one another as friends.[3]

For many social network purposes, the distinction between cases and affiliations may appear somewhat artificial. In a study of, say, the involvement of 18 women in 14 social events, it would seem only sensible to regard the women as the cases and the events as their affiliations. Indeed, this would be in line with the normal survey practice of treating the agents as the cases. But with such phenomena as overlapping group memberships, for example, the situation is far less clear-cut. This kind of research is interested in the extent to which a group of organizations overlap in their membership, with how similar they are in their patterns of recruitment. Both the groups and their members are agents in the sociological sense, and so both have an equal right to be considered as the 'cases'. The members may be treated as the cases, in which case the organizations of which they are members will be treated as their affiliations; or the organizations may be treated as cases and the members that they share will be seen as their affiliations. The choice of which set of agents to treat as the cases for the purpose of network analysis will depend simply on which is seen as being the most significant in terms of the research design.

This decision will normally have been reflected in prior sampling decisions. If the organizations are assumed to be of the greatest importance, then a sample of organizations will be selected for study and the only people who will figure in the subsequent analysis will be those who happen to be members of these organizations. In such a research design, the organizations have a theoretical priority and it would seem sensible to treat the members as indicating affiliations between organizations. As far as the techniques of network analysis are concerned, however, it makes no difference which of the two are regarded as the cases. The same procedures may be applied whichever choice is made, and it is the task of the researcher to decide which of them may have a meaningful sociological interpretation.[4]

The distinction between cases and affiliations, therefore, may generally be regarded as a purely conventional feature of research designs for network analysis. A further aspect of this convention is to place the cases on the rows of the incidence matrix and the affiliations on its columns. This is based on the conventions employed in attribute analysis, where the cases are treated as rows and the variables are treated as columns.

If the data matrix is to be used as a basic organizational framework for relational data, certain other conventions must also be understood. These other conventions can be recommended as the basis of best practice in network analysis, as they help to ensure maximum clarity in research discussions. Most readers will be familiar with the importance of conventions in basic mathematics. It is conventional when drawing ordinary graphs of variables, for example, to use the vertical axis for the dependent variable and to label this as the '*y*' axis. The horizontal axis is used for the independent variable and is labelled as the '*x*' axis. This convention prevents any confusion about how the graph is to be read and it ensures that any statements made about the graph will be unambiguous. The conventions surrounding the relational data matrix have the same purpose.

In the discussion of matrices, it is conventional to designate the number of rows in a matrix as '*m*' and the number of columns as '*n*'. It is also customary to list the rows first when describing its size. The overall size of a matrix can, therefore, be summarized by referring to it as an $m \times n$ matrix. The incidence matrix of Figure 3.3, for example, is a 4×5 matrix. It is also conventional to refer to the rows before the columns when describing the contents of any particular cell, and to use the letter '*a*' to refer to the actual value contained in the cell. Thus, the value contained in the cell corresponding to the intersection of row 3 with column 2 would be designated as $a(3,2)$. This can be generalized by using the convention of referring to the individual rows by '*i*' and the columns by '*j*'. Thus, the general form for the content of a cell is $a(i,j)$, where the researcher may then go on to specify the relevant values for i and j. These conventions are summarized in Figure 3.4.

The usefulness of the matrix approach to relational data can best be illustrated through a concrete example. Figure 3.5 contains some artificial data on interlocking directorships among companies. An interlocking directorship, or interlock, exists where a particular person sits as a director on the boards of two or more companies.

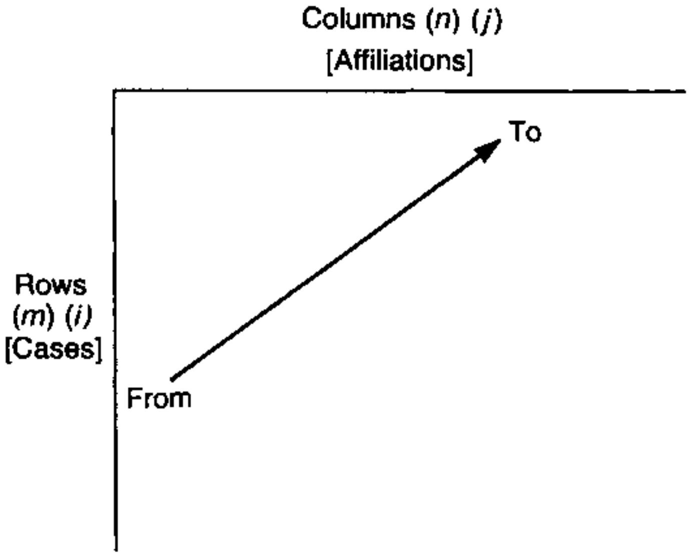


Figure 3.4 *Matrix conventions: best practice*

His or her presence on the two boards establishes a relation between the companies. In many investigations of interlocking directorships, it is the companies that are of central interest. For this reason, they are generally treated as the cases and so they are shown as the rows of the incidence matrix of Figure 3.5. The affiliations, shown in the columns of this matrix, are the directors that the companies have, or do not have, in common with one another. Each cell of the matrix contains a binary digit, '1' or '0', which indicates the presence or absence of each director on each company. Thus, company 1 has four directors (A, B, C and D), and director A sits on the board of company 2 as well as company 1. This means that there is an interlock between company A and company B. Adjacency matrix (ii) in the diagram shows the interlocks that exist among all companies. In this matrix, each cell shows more than the mere presence or absence of an interlock, it shows the number of directors in common between a pair of companies. The cells contain actual values, rather than simply binary digits, because companies may have more than one director in common. Thus, company 1 and company 4 have just one director in common (director C), while companies 2 and 3 have two directors in common (directors B and C). This can be confirmed by examining the columns of the original incidence matrix, which show that director C sits on companies 1 and 4, and that directors B and C each sit on companies 2 and 3.

The simplest kind of analysis of this adjacency matrix might suggest that the strength of a relation can be measured by the *number* of interlocks that it involves. The strongest relations, then, exist between companies 1 and 2 and between companies 1 and 3,

(i) Incidence matrix

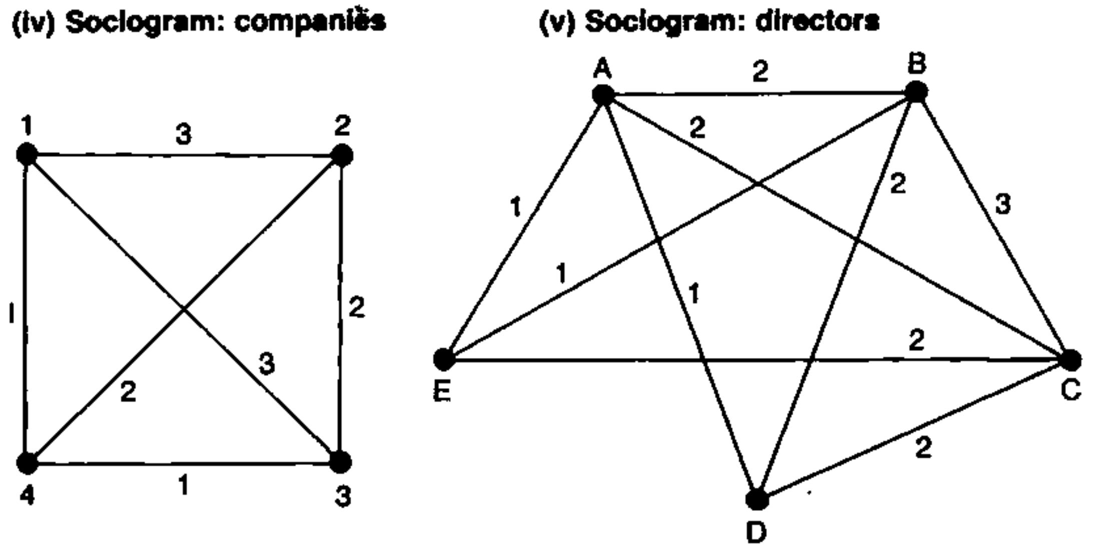
		Directors				
		A	B	C	D	E
Companies	1	1	1	1	1	0
	2	1	1	1	0	1
	3	0	1	1	1	0
	4	0	0	1	0	1

(ii) Adjacency matrix: companies-by-companies

	1	2	3	4
1	–	3	3	1
2	3	–	2	2
3	3	2	–	1
4	1	2	1	–

(iii) Adjacency matrix: directors-by-directors

	A	B	C	D	E
A	–	2	2	1	1
B	2	–	3	2	1
C	2	3	–	2	2
D	1	2	2	–	0
E	1	1	2	0	–



(iv) Sociogram: companiès

(v) Sociogram: directors

Figure 3.5 *Matrices for interlocking directorships*

each of these relations involving three directors. The weakest links would be those that involve just one director. The sociogram of companies indicates the structure of the matrix quite clearly, with the numbers attached to the lines indicating the strength or 'value' of the lines. This sociogram could equally well have been drawn in other ways: for example, with the thickness of the lines representing their value, or with the points connected by one, two, or three parallel lines. Each method would convey the same information about the structure of the matrix.

It will be recalled that it is possible to derive two adjacency matrices from a single incidence matrix. In this example, it is possible to derive not only the company-by-company adjacency matrix but also a director-by-director adjacency matrix. This matrix and its associated sociogram of directors, in Figure 3.5, show the relations among the directors that exist when they sit on the same company board. There is, for example, a strong relation between B and C, who meet one another on three separate corporate boards (the boards of companies 1, 2 and 3), and rather weaker, single board relations between A and D, between A and E, and between B and E. The sociogram of directors also illustrates such sociometric ideas as that D and E are relatively more 'peripheral' to the network than are the other directors: they have fewer connections, their connections are generally weaker, and they are not connected to one another.

The adjacency matrices shown in Figure 3.5 also illustrate some further general considerations in social network analysis. First, it is important to note something about the diagonal cells running from the top left to the bottom right. In matrix analysis this particular diagonal is referred to simply as 'the diagonal', because the cells are different from all others in the matrix. In a square matrix the diagonal cells show the relation between any particular case and itself. In some situations this is a trivial relation that exists simply by definition, while in others it may be an important feature of the network. The cells on the diagonal of matrix (ii) of Figure 3.5, for example, refer to the relation of each company to itself. In this example, these data would not have any particular meaning. The fact that a company is connected to itself through all its directors is true but trivial, as our concern is with *inter*-company relations. For this reason, the diagonal cells contain no values and should be ignored in the analysis. Many technical procedures in network analysis require the researcher to specify whether diagonal values are to be included or excluded, if this is at all ambiguous. For this reason, researchers must always be aware of the status of the diagonals in their matrices and will need to understand how particular procedures handle the diagonal values.

Figure 3.5 also shows that the adjacency matrices are *symmetrical* around their diagonals: the top half of each matrix is an identical, mirror image of its bottom half. The reason for this is that the data describe an 'undirected' network, a network in which the relation of company 1 to company 2, for example, is the same as the relation of company 2 to company 1. The existence of a relation between the two is considered independently of any question of whether the relation involves the exercise of power and influence in one direction but not in the other. For this reason, all the relational information in an adjacency matrix for an undirected network is contained in the bottom half of the matrix alone; the top half is, strictly speaking, redundant. Many analytical procedures in network analysis, therefore, require only the bottom half of the adjacency matrix and not the full matrix. For undirected networks, no information is lost in this method of analysis.

Undirected data are the simplest and easiest type of relational data to handle, and it is, perhaps, necessary to spend a little time in discussing some of the more complex types of data. One of the most important considerations in variable analysis is the level of measurement that is appropriate for a variable. This is the question of whether attribute data should be measured in nominal, ordinal, ratio, or interval terms. From this decision flow many other decisions about which particular analytical procedures can appropriately be used for the data. Similar measurement problems arise with relational data, according to whether the data are 'directed' and/or 'numbered'. Figure 3.6 uses these two dimensions to classify the four main levels of measurement in relational data.

		Directionality	
		Undirected	Directed
Numeration	Binary	1	3
	Valued	2	4

Figure 3.6 *Levels of measurement in relational data*

The simplest type of relational data (type 1) is that which is both undirected and binary. This is the form taken by the data in the incidence matrix of Figure 3.5. The adjacency matrices in that Figure contain relational data of type two: the relations are undirected but valued.[5] I have already shown that the 'valued' data (type 2) in the adjacency matrices of Figure 3.5 are derived from the binary data of the original incidence matrix. Values typically indicate the strength of a relation rather than its mere presence. The signed data that were discussed in the previous chapter in connection with theories of balance, are relational data where a '+' or '−' is attached to each line. These relations can be regarded as intermediate between the binary and valued types. Such data show more than simply the presence or absence of a relation, as the presence is qualified by the addition of a positive or negative sign; but the nature of the relation is indicated simply by the polarity and not by an actual value. It is, of course, possible to combine a sign with a value and to code relational data as varying from, say, −9 to +9. In such a procedure, the value could not represent simply the number of common affiliations between cases, as they cannot have a negative number of affiliations in common. The values must, therefore, be some other measure of the strength or closeness of the relation. Such a procedure would, of course, rest upon a *sociological* argument that produced solid theoretical or empirical reasons for treating the data in this way.

Valued data can always be converted into binary data, albeit with some loss of information, by using a cut-off value for 'slicing' or dichotomizing the matrix. In a slicing procedure, the researcher chooses to consider only those relations with a value above a particular level as being significant. Values above this level are sliced off and used to construct a new matrix in which values at or below this level are replaced by '0' entries and values above it are replaced by '1' entries. This procedure of slicing the data matrix is a very important technique in network analysis, and will be discussed more fully in Chapter 5. Directed data can also take binary or valued forms, and similar slicing procedures can be applied to reduce valued and directed data (type 4) to binary and directed data (type 3). It is also possible to reduce directed data to undirected data, by the simple expedient of ignoring the direction. Thus, a researcher may decide that the important thing to consider is the mere presence or absence of a relation, and not its direction. In this case, then, it makes sense to ignore the directionality of the data. A further matrix convention may appropriately be mentioned at this point. In adjacency matrices that contain directed data, the usual convention is to present the direction of a relation as running 'from' a row element 'to' a column element. Thus, the entry in cell (3,6) of a directed matrix would show the presence or strength of the relation directed *from* person 3 *to* person 6. The relation directed from person 6 to person 3 would be found in cell (6,3). This convention is shown in

Figure 3.4. It is for this reason that a directed matrix is asymmetrical around its diagonal, and that, therefore, the whole matrix must be considered, and not simply its bottom half.

Complex types of relational data can always be reduced to more simple types and, in the last instance, *any* type of relational data may be treated as if it were undirected and binary (type 1). Techniques appropriate to this type of data, therefore, have the widest application of all the techniques of social network analysis. It is not, of course, possible to undertake the reverse operation, converting simple to complex data, unless additional information is available over and above that contained in the original data matrix.[6]

Researchers must always take great care over the nature of their relational data. They must, in particular, be sure that the level of measurement used is sociologically appropriate. The attempt to use valued data in studies of interlocks, for example, rests upon assumptions about the significance of multiple directorships that may or may not be appropriate. It might be assumed, for example, that the *number* of directors in common between two companies is an indicator of the strength or closeness of a relation. Having four directors in common, on this basis, would mean that two companies are 'closer' than those that have only two directors in common. But is this a valid sociological assumption? If it is not, the mathematical procedure should not be used. Mathematics itself cannot provide an answer for the researcher. The relevance of particular mathematical concepts and models is always a matter for the informed sociological judgement of the researcher. Even if it is decided that it is reasonable to use valued data, the researcher must be alive to other assumptions that might be contained in the mathematical procedures. Does a procedure, for example, treat the values as ordinal or as ratio variables? In the former case, a value of four would be regarded simply as being stronger than a value of two; in the latter case the relationship would be regarded as being *twice* as strong. The choice of a level of measurement is, again, a sociological question and not a mathematical one.

The Storage of Relational Data

The analysis of very small data sets is often quite straightforward. An adjacency matrix and sociogram for a four- or five-person group, for example, can easily be constructed by hand. However, this becomes more difficult when the size of the network is any greater than this. When dealing with data sets that have more than about ten cases and five affiliations, it is all but essential to use a computer. Not only does computer processing save a considerable amount of time – the matrix re-arrangement undertaken by Homans in his investigation of the involvement of 18 women in 14 social events, for example, can be undertaken on a computer in a few seconds at most – it also allows analyses to be undertaken that are simply not possible by hand.

If relational data are properly stored, they can be managed and manipulated more efficiently. The need to use computers for network analysis, then, means that it is important to consider how the logical structure of the data matrix can be translated into a computer file. The first step is often to sort names of agents or events in order to generate listings that can be analysed for their connections. Research on interlocking directorships, for example, involves generating a list of directors in the target companies, sorting this into alphabetical order, and then identifying those names that appear two or more times. The most straightforward method for doing this is to use a text editor or word processor to create a data file, as the names can be typed in as text and then sorted and edited. Many word processors will allow data to be sorted into alphabetical or numerical order as an aid to its analysis and manipulation.[7]

The most usual result of this processing is data in 'linked list' format. In a linked list, each line of text in the file shows a case followed by its affiliations. It might show, for example, the name of a director followed by the names of all the companies of which he or she is a director. However, this cannot usually be transformed into an incidence matrix (as shown in Figure 3.3) simply with a word processor. Unless the user wishes to undertake some difficult – and error-prone – manual processing, it is useful to move the data directly into a social network analysis program such as UCINET. In this program, linked list data can be imported and can be invisibly converted to an incidence matrix. The program also allows the new data files to be directly edited in their original linked list format.

The linked list format of UCINET is presented on the screen as a spreadsheet, and a number of data processing tasks can, in fact, be carried out with a spreadsheet program such as EXCEL. Indeed, a spreadsheet can import linked lists directly from a word processor if no social network program is available. While the spreadsheet is, still, widely seen as a financial tool for accountants and stock market analysts, it is essentially an electronic matrix manipulator. Even the simplest of spreadsheets can be used to store and to organize relational data, and they can be used to prepare these data in files readable by many other specialist packages. Suitable spreadsheet programs are so widely available that it is worth considering them as a basic data storage and manipulation system for social network data. If the data have been converted from linked lists to matrices in binary or valued form, the spreadsheet can be used to calculate basic statistical measures, such as row and column sums, frequency distributions, correlations, and so on. Many of these measures can be converted into screen graphics and then printed out. Frequency distributions, for example, can be instantly plotted on a histogram or bar chart. While the major mathematical functions built into spreadsheets are the kind of financial and statistical procedures most appropriate for variable analysis, a number of spreadsheet programs include facilities for matrix mathematics that allow the calculation of various structural properties of networks.[8]

Data stored on a spreadsheet can be manipulated very easily, providing a solution to the practical problems of data preparation that have often plagued network analysis. Virtually all spreadsheets, for example, will allow rows and columns to be sorted into alphabetical or numerical order, automatically re-arranging the corresponding data. The spreadsheet's 'range' options can be used to specify particular parts of a matrix for copying to a new file. If, for example, a matrix of friendship relations among people is stored in a file, it is possible to select the male or the female data alone for separate analysis. It is even possible to transform an incidence matrix into its corresponding adjacency matrices. This kind of use of a spreadsheet, however, is probably best attempted only if other programs specifically designed for social network analysis are not available. The principal use of the spreadsheet should be to store the data and to carry out the straightforward data management functions of re-arrangement and manipulation.[9]

The two most widely used social network packages – UCINET and STRUCTURE – both store their data in simple matrix form, and it is easy to transfer an appropriate file directly from a spreadsheet to either of these packages.[10] For most purposes, it is best to import data into one of the specialist packages as early as possible, reading it back into a spreadsheet only when attribute data have to be added and used in statistical analyses. In these circumstances, in fact, it may be preferable to export the data files to a specialist statistical package such as SPSS.

One of the most powerful network analysis packages is GRADAP, but this program uses data files in a format that is different from the matrix structure discussed so far. GRADAP can exchange data files with SPSS, but it cannot handle direct input of the incidence and adjacency matrices. GRADAP data files can be produced in a spreadsheet or a word processor, or in the SPSS text editor, but the format is less intuitive than the data matrix. GRADAP aims at a complete translation of relational data into the terminology of graph theory, and so requires that there be an explicit identification of the points and lines that comprise the data. Instead of an incidence matrix, GRADAP requires that the adjacency matrices themselves be specified in two separate files: a 'point file' that lists the cases, and a 'line file' that lists each relation. A line is defined by the points at either end of it. Where the researcher has direct sociometric choice data, or where actual patterns of relations are observable in some way, this poses few problems. But where the data exist in linked lists or incidence matrices, it can be quite difficult to produce the necessary files. Even with the help of spreadsheet or database programs, a considerable amount of manual processing is required to produce the GRADAP files.

Line number	Tail	Head	Line info
1	1	2	A
2	1	2	B
3	1	2	C
4	2	3	B
5	2	3	C
6	1	3	B
7	1	3	C
8	1	3	D
9	1	4	C
10	3	4	C
11	2	4	C
12	2	4	E

Figure 3.7 *A GRADAP line file*

Figure 3.7 shows the form of a GRADAP line file for the data contained in the incidence matrix of Figure 3.5. There are 12 interlocking directorships in this network – the total can be confirmed by adding the values in the bottom half of the corresponding adjacency matrix (ii) of Figure 3.5. Each interlock is counted as a separate 'line' for data input to GRADAP, and so the line file contains 12 entries.[11] Each of the 12 lines is identified by the points that lie at its 'tail' and its 'head', and further information about the line (such as the name of the director responsible for it) can be added to the file.[12] A comparison of Figures 3.7 and 3.5 will confirm that the GRADAP line file contains all the information that is contained in the incidence matrix, and this is the reason why GRADAP can, from its line files, invisibly construct the two adjacency matrices.

Once a GRADAP file structure has been created, the program offers powerful data management facilities, acting almost like a specialist database management system for relational data. However, the program requires a thorough knowledge of graph theory if it is to be used for even the simplest of analyses. For the newcomer and the occasional network analyst, therefore, UCINET, together with a word processor, offers the best facilities. For the advanced user who is able to use other programs to generate the line file, GRADAP has many advantages. These programs are discussed more fully in the Appendix.

The Selection of Relational Data

In the first two sections of this chapter I have looked at the nature of relational data and at how it can be organized and managed for network analysis. Having clarified the ways in which the collected data can be organized and stored, it is possible to examine some remaining issues concerning the data collection process. I have argued that few distinct problems arise in this area, but the question of the *selection* of data is one that does pose considerable problems for social network analysis. These selection problems concern the boundedness of social relations and the possibility of drawing relational data from samples.

A common strategy in the study of small scale social networks has been to identify all the members of a particular group and to trace their various connections with one another. But this is a far from straightforward matter. Social relations are social constructs, produced on the basis of the definitions of the situation made by group members. A relation of 'close friendship', for example, may mean different things to different people, according to their conceptions of what it means to be 'close'. The researcher who simply asks respondents to identify their 'close friends' cannot be sure that all respondents will have the same understanding of 'closeness'. Respondents with a restrictive definition of closeness will draw narrow boundaries around themselves, while those with a more inclusive conception of friendship will recognize more extensive boundaries. The very boundaries of the group of close friends, therefore, will vary from one person to another. Any boundaries identified by the researcher through an aggregation of these individual perceptions may be wholly artificial – simple artefacts of question wording. If, on the other hand, the researcher explicitly defines 'close' – by, for example, frequency of interaction – he or she will be imposing a definition of closeness on the respondents and the boundaries of friendship may, again, be artificial.

This issue is important, as researchers often have unrealistic views about the boundaries of relational systems (Laumann et al., 1989). It is often assumed that the social relations of individuals will be confined to the particular group or locale under investigation. To the extent that connections outside this locale are ignored, the social network studied will be an imperfect representation of the full network. This is especially clear in the case of informal groups, such as street gangs, where the boundaries of the group are loosely drawn and where gang members' activities stretch well beyond its core membership (Yablonsky, 1962). But the same is also true for more formal groups. Kerr and Fisher (1957), for example, discussed the 'plant sociology' that focuses its attention on the physical boundaries of particular workshops and offices in isolation from the wider economy. Such investigations isolate their research locale from the larger regional, national and international systems in which they are embedded. Research that is confined to the local work situation may fail to identify those relations that extend beyond the plant.

In a similar vein, Stacey (1969) has criticized locality studies for their assumption that bonds of 'communal' solidarity are confined within the local social system. She holds that they must be seen as stretching out to entwine with the larger economic and political systems. Similarly, Laumann et al. (1983: 31) have argued that a locality study of the flow of money through a network ought not to limit its attention to that geographical locality. Many of the most important agencies in the circulation of money will lie outside the locality: federal government agencies, regional and national banks, multinational companies, and so on. If, as is likely, these are more important to the flow of money than are the local organizations and agencies, a locality-based research project faces the possibility of a totally inadequate view of the structure of the relevant network of transactions.

What these problems point to is the fact that the determination of network boundaries is not simply a matter of identifying the apparently natural or obvious boundaries of the situation under investigation. Although 'natural' boundaries may, indeed, exist, the determination of boundaries in a research project is the outcome of a theoretically informed decision about what is *significant* in the situation under investigation. A study of political relations, for example, must recognize that what counts as 'political', how this is to be distinguished from 'economic', 'religious' and other social relations, and the choice of boundaries for the relevant political unit, are all theoretically informed decisions. Researchers are involved in a process of conceptual elaboration and model building, not a simple process of collecting pre-formed data.

Assuming that relevant boundaries can be identified, the research may proceed to define the target population for study. Two general approaches to this task have been identified: the 'positional' and the 'reputational' approaches.[13] In the **positional approach**, the researcher samples from among the occupants of particular formally defined positions or group memberships. First, the positions or groups that are of interest are identified, and then their occupants or members are sampled. Unless the population under investigation is very small, this is likely to require some kind of enumerated list that covers the whole of the target population. Examples of this kind of strategy would be samples drawn from a school class, a village, a workgroup, or from institutions such as a political elite or corporate directorate. A familiar problem with positional studies is that of determining which positions to include. Studies of elites, for example, have often been criticized for their identification of top positions in institutional hierarchies, especially when the researcher offers no real justification for the cut-off threshold used to distinguish the 'top' from other positions in the institutional hierarchy. This problem is, of course, a reflection of the general boundary problem that has already been discussed, and it is important that researchers have theoretically and empirically justifiable reasons for the inclusion or exclusion of particular positions.

This sometimes involves an assumption that there are 'natural' sub-groups within the population. Research on business interlocks, for example, has often focused attention on the 'top 250' companies in an economy.[14] This research strategy involves the assumption that the division between the 250th and the 251st company forms a natural boundary between large-scale and medium-scale business. However, such boundaries can rarely be drawn with precision. There is a continuous gradation in size from large to small, and, while it may be possible to identify the points in the size distribution at which the gradient alters, it will not generally be possible to draw sharp boundaries. Indeed, most such research does not examine the overall size distribution for shifts of gradient, but simply uses an arbitrary and *a priori* cut-off threshold: while some researchers investigate the top 250, others investigate the top 50, top 100, or top 500 slices of the distribution.[15]

In the positional approach the selection of cases for investigation may sometimes follow from an earlier decision about the selection of affiliations. A directorship, for example, can be regarded as a person's affiliation with a company, and a researcher may already have decided to limit attention to a particular group of companies. In such a situation, the selection of directors for study is determined by the selection criteria used for the companies.[16]

The **reputational approach** can be used where there are no relevant positions, where there is no comprehensive listing available, or where the knowledge of the agents themselves is crucial in determining the boundaries of the population. In the reputational approach, the researcher studies all or some of those named on a list of nominees produced by knowledgeable informants. Those included on the list are those who are reputed to be the members of the target population. The informants are asked to nominate, for example, 'powerful members of the community', 'people of high standing in business', and so on, depending on the purposes of the research, and these nominations are combined into a target population. The choice of informants is, obviously, of crucial importance in the reputational approach. The researcher must have good reasons to believe that the informants will have a good knowledge of the target population and are able to report this accurately. Whether or not this is the case will often be known only when the research has been completed, and so there is an element of circularity in the strategy. For this reason, researchers ought to endeavour to come up with theoretical and empirical reasons for the choice of informant which are, so far as is possible, independent of the particular social relations under investigation.

This will not always be possible, and one particular variant of this reputational strategy, using the so-called 'snowballing' technique, follows exactly the opposite procedure. In this approach, a small number of informants are studied and each is asked to nominate others for study. These nominees are, in turn, interviewed and are asked for further nominations. As this procedure continues, the group of interviewees builds up like a snowball. Eventually, few additional nominees are identified in each round of interviews. In the snowballing method, the social relation itself is used as a chain of connection for building the group. By its very nature, however, a snowball sample is likely to be organized around the connections of the particular individuals who formed its starting point. For this reason, the method of selection tends to determine many of the relational features of the resulting social network. This network is built from the relations of a group of connected agents and, as Laumann et al. remark, 'it is scarcely informative to learn that a network constituted by a snowball sampling procedure is well-connected' (1983: 22).

A final strategy of selection, neither positional nor reputational, occurs when the columns of the incidence matrix are true affiliations and the researcher aims to select these separately from the cases. Such research would be concerned with choosing, say, the activities and events in which people are involved, independently of any positions or organizations that may have been used to identify the people themselves. In his study of New Haven, for example, Dahl (1961) used participation in the making of key decisions as the basis of selection. Involvement in decision-making, therefore, was seen as an 'affiliation' for which people could be given a binary or numeric value independently of whatever organizational positions they held. This allowed Dahl, or so he believed, to assess the relative power of different categories of agents, instead of assuming that power was an automatic correlate of position. A similar strategy was that of Davis (1941) and his colleagues in *Deep South*, where social events were studied, resulting in a matrix showing the participation of 18 women in 14 events. The problem in this kind of strategy, of course, is that of how to justify the choice of affiliations: have the most important events been chosen, and what is a 'key issue'? Selection of true affiliations, therefore, involves precisely the same problems as the direct selection of cases. Activities and events can be chosen because they are regarded as objectively significant (a variant of the positional approach) or because knowledgeable informants believe them to be important (a variant of the reputational approach).

I have written so far mainly of the selection of whole populations through complete or quasi-enumeration. But it may often be necessary to use sample data, and these matters become more complicated. Few sampling problems arise in small group studies, where it is generally possible to undertake a complete enumeration of all group members and of their relations with one another. When research on large-scale social systems is being undertaken, however, a complete enumeration may not be a viable aim, and there will be particularly intractable sampling problems. The sheer scale of the resources needed will often preclude the complete enumeration of large populations, but, even if such research proved possible – for example, in a census of population – the scale of the resulting data matrix would make any analysis impossible. As square adjacency matrices must be constructed before most network analyses can be undertaken, the data matrices can be quite enormous. Attribute data for, say, 1000 cases and 50 variables would involve 50,000 entries in a data matrix. Advances in computing have made such matrices relatively easy to handle for most statistical purposes. In the case of relational data, however, the case-by-case adjacency matrix for 1000 cases would contain 1,000,000 cells. In the case of a fairly small village with a population of 5,000 people, the adjacency matrix would contain 25,000,000 cells, which is beyond the capacity of most available computers and software.[17] For a national population running into the millions, the sheer quantity of data can hardly be imagined, and the computing power required to handle this simply does not exist outside the realms of science fiction.

It was, of course, similar problems that, in the pre-computer age, led to the development of sampling techniques that would allow, say, a sample of 1000 to be used instead of a complete enumeration of a population of many thousands. The statistical theory of sampling sets out the conditions under which attribute data collected from a sample of cases can be generalized into estimates for larger populations. It might be assumed, therefore, that sampling from large populations would provide a similar workable solution for social network analysis. Figure 3.8 gives a schematic account of the ideal sampling process in social network analysis. A particular population of agents will be involved in a complex system of social relations of all types that make up the total network. Within this relational system, sociologists may identify such 'partial' networks as those comprising economic relations, political relations, religious relations, and so on. When a strategy of complete enumeration is followed, the researcher can attempt to ensure that full information is obtained on all the relevant relations, and so can construct adequate models of the partial networks.

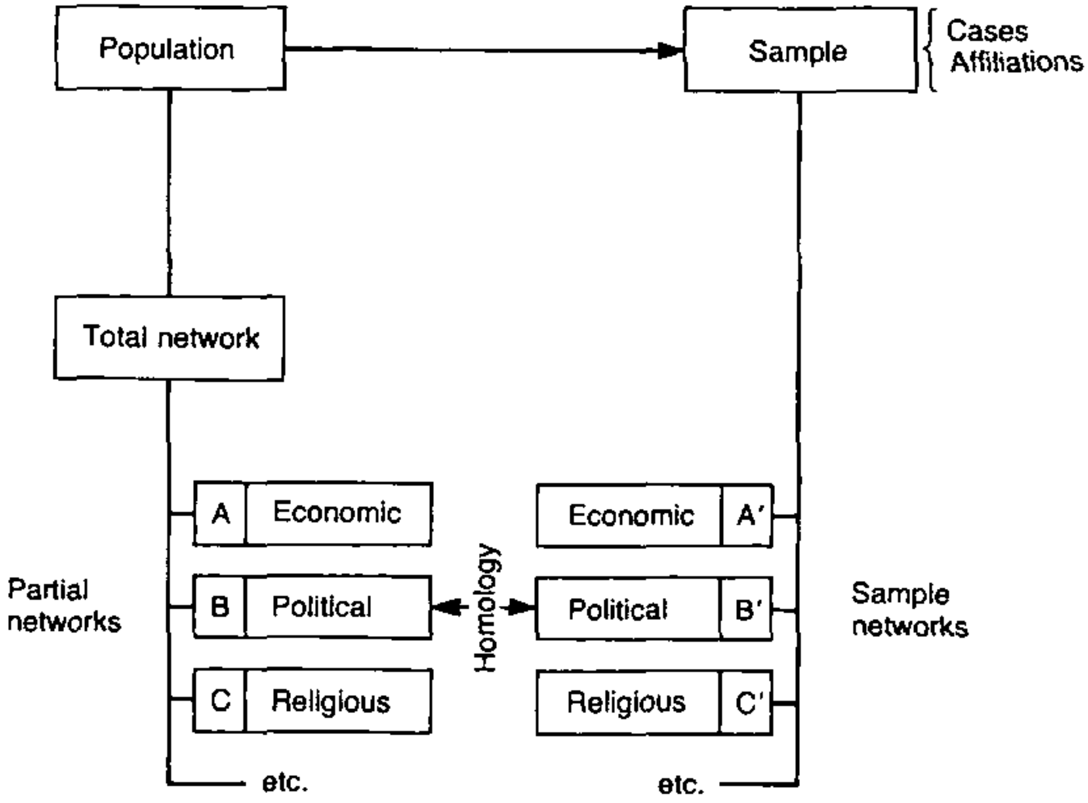


Figure 3.8 *Networks and sampling: the ideal*

The task of sampling would appear to be obvious and straightforward, involving nothing more than the general principles of sampling in survey research: a representative sample of cases is drawn from the population in question, their relations are investigated, and sample networks are constructed that will be homologous to the partial systems that occur in the population as a whole. But things are not, in fact, as simple as this. The general principles of sampling are based on the application of the theory of probability to large numbers of observations, and there are well-established mathematical rules for judging the reliability of sample data. There are no such rules for judging the quality of relational data derived from a sample; and there are good reasons for assuming that sampling may result in unreliable data. Although it is possible to draw a sample of 1000 cases for analysis and it might be possible to find a computer and program capable of handling a 1000 × 1000 adjacency matrix, there is no guarantee that the structure of this sample network would bear any relationship to the structure of the corresponding partial network. A representative sample of *agents*, does not, in itself, give a useful sample of *relations* (Alba, 1982: 44).

It might seem, at first sight, that this is not a real problem. The overall distribution of relations among agents and their 'density',[18] for example, might seem an easy matter to estimate from sample data: the sample provides data on the network attributes of the individual cases, and these can be used to calculate overall network parameters. The density of the friendship ties in a country, for example, could be assessed by asking a random sample of people how many friends they have. If the sample is large enough, these estimates ought to be reliable. But it is almost impossible to go beyond such basic parameters to measure the more qualitative aspects of network structure.

The reasons for this relate to the sparsity of the relational data that can be obtained from a sample survey of agents. Even if there was a perfect response rate and all respondents answered all the questions in full, many of the contacts named by respondents will not themselves be members of the sample. This means that the number of relations among members of the sample will be a very small subset of all their relations, and there is no reason to believe that the relations identified among the agents in the sample would themselves be a random sample of *all* the relations of these same agents. With a very large population, such as that of a national study, it is very unlikely that *any* member of a random sample will have any kind of social relation with others in the same sample. The probability of a connection existing between two individuals drawn at random from a population of many millions is so low as to be negligible. It is, therefore, unlikely that a researcher could say anything at all about the relational structure of the national population from a random sample. Burt (1983a) has made a rough estimate that the amount of relational data lost through sampling is equal to $(100-k)$ per cent, where k is the sample size as a percentage of the population. Thus, he argues that a 10 per cent sample involves the loss of 90 per cent of the relational data – even a massive 50 per cent sample would involve the loss of half of the data. Such a loss of data makes the identification of cliques, clusters and a whole range of other structural features virtually impossible in conventional sample research.

Sample data can also lead to difficulties in arriving at basic measures of the relational attributes of the particular individuals studied, especially if there is any amount of non-response in the survey. Imagine, for example, an attempt to estimate the sociometric popularity of agents in a network in which there is a very small number of very popular agents and a much larger number of less popular ones.[19] Because they exist in very small numbers, a sample is unlikely to include sufficient of the very popular agents to allow any generalizations to be made about the overall patterns of popularity in the network. This is akin to the problem of studying a small elite or dominant class through a national random sample survey. Unless the sample is very large indeed, they will not appear in adequate numbers, and a very large sample defeats much of the point of sampling. One way around this, of course, might be to use a stratified sample, in which popular agents have a higher probability of selection. The obvious difficulty with this, however, is that such a sampling strategy could be implemented only if the researcher already knew something about the distribution of popularity in the population.

There seem, at present, to be three different responses to these sampling problems. The first is to abandon any attempt to measure the global properties of social networks and to restrict attention to personal, ego-centric networks. This research strategy involves looking at the unrestricted choices that people make, including those to others not included in the sample, and calculating, for example, the density and certain other ego-centric features of their contacts. As no attempt is made to generalize about, for example, the density or 'close knit' texture of the overall network, sampling poses few difficulties other than those that arise in any kind of social research. This is the strategy used in studies of friendship and community undertaken by Wellman (1979), Fischer (1982) and Willmott (1986, 1987).

The second response is to use a form of snowballing. Frank (1978a, 1979) argues that researchers should draw an initial sample of cases and then collect information on all the contacts of the sample members, regardless of whether these are members of the original sample. These contacts are added to the sample and their contacts are discovered in the same way. By extending this process through a number of stages, more and more of the indirect contacts of the members of the initial sample will be discovered. The researcher must decide how far to continue this snowballing. This will generally be to the point at which the number of additional members added to the sample drops substantially, because names that have already been included are being mentioned for the second or third time. Frank has shown that such a snowballing method allows a reasonable estimate to be made of such things as the distribution of contacts and the numbers of dyads and triads. A snowball sample, of course, is not a random sample: the structure that is discovered is, in fact, 'built in' to a snowball sampling method itself. But this is precisely what is necessary in order to avoid the sparsity of connections found in a random sample. The assumption of the snowball sampling method is that the connected segment of the network that forms the sample network is representative of all other segments of the network. The researcher, then, must have some knowledge about the population and their relations in order to make this assessment of representativeness. But snowballing does, at the very least, make it possible to try to estimate which features of the structure may be an artefact of the sampling method itself and so to control for these in the analysis.[20]

The third response to the sampling problem is that of Burt (1983a), who has suggested a way of moving on to some of the more qualitative features of social networks. In particular, Burt is concerned with the identification of 'positions' or structural locations, such as roles. If it is assumed that agents in a similar structural location in a network will have various social attributes in common, then it is possible to use survey data on the typical relations between agents with particular attributes as a way of estimating what structural locations might exist in the network. From each respondent it is necessary to obtain information about their social attributes and the attributes of those to whom they are connected (including people outside the sample). Agents can then be grouped into sets of agents with commonly occurring combinations of attributes, and these sets can be arranged into a sets-by-sets square matrix that shows the frequency of relations between members of the various categories. It might be discovered, for example, that 70 per cent of white men have black male friends, while only 20 per cent of white women have black male friends. Such measures, argues Burt, provide estimates of the valued relations between social 'roles' that could be expected to occur if the researcher had undertaken a complete enumeration of all men and women in the population.

There are glimmers of what can be achieved in the study of large scale social systems using sampling methods. Though it might seem, at present, impossible to discover anything about such things as cliques and clusters from sample data, it is to be hoped that further advances in the techniques of network sampling will make this possible (Alba, 1982: 46; Frank, 1988).