Introduction to Social Network Methods

1. Social Network Data

This page is part of an on-line textbook by Robert A. Hanneman and Mark Riddle of the Department of Sociology at the University of California, Riverside. For more information, or to offer comments, you can send me e-mail.

Table of Contents

- Introduction: What's different about social network data?
- Nodes
  Populations, samples, and boundaries
  Modality and levels of analysis
- Relations
  Sampling ties

  Multiple relations
- Scales of measurement
- A note on statistics and social network data

Introduction: What's different about social network data?

On one hand, there really isn't anything about social network data that is all that unusual. Networkers do use a specialized language for describing the structure and contents of the sets of observations that they use. But, network data can also be described and understood using the ideas and concepts of more familiar methods, like cross-sectional survey research.

On the other hand, the data sets that networkers develop usually end up looking quite different from the conventional rectangular data array so familiar to survey researchers and statistical analysts. The differences are quite important because they lead us to look at our data in a different way -- and even lead us to think differently about how to apply statistics.

"Conventional" sociological data consists of a rectangular array of measurements. The rows of the array are the cases, or subjects, or observations. The columns consist of scores (quantitative or qualitative) on attributes, or variables, or measures. Each cell of the array then describes the score of some actor on some attribute. In some cases, there may be a third

dimension to these arrays, representing panels of observations or multiple groups.

Name	*Sex*	*Age*	*In-Degree*
Bob	Male	32	2
Carol	Female	27	1
Ted	Male	29	1
Alice	Female	28	3

The fundamental data structure is one that leads us to compare how actors are similar or dissimilar to each other across attributes (by comparing rows). Or, perhaps more commonly, we examine how variables are similar or dissimilar to each other in their distributions across actors (by comparing or correlating columns).

"Network" data (in their purest form) consist of a square array of measurements. The rows of the array are the cases, or subjects, or observations. The columns of the array are -- and note the key difference from conventional data -- the same set of cases, subjects, or observations. In each cell of the array describes a relationship between the actors.

Who reports liking whom?				
	Choice:			
Chooser:	Bob	Carol	Ted	Alice
Bob	---	0	1	1
Carol	1	---	0	1
Ted	0	1	---	1
Alice	1	0	0	---

We could look at this data structure the same way as with attribute data. By comparing rows of the array, we can see which actors are similar to which other actors in whom they choose. By looking at the columns, we can see who is similar to whom in terms of being chosen by others. These are useful ways to look at the data, because they help us to see which actors have similar positions in the network. This is the first major emphasis of network analysis: seeing how actors are located or "embedded" in the overall network.

But a network analyst is also likely to look at the data structure in a second way -- holistically. The analyst might note that there are about equal numbers of ones and zeros in the matrix. This suggests that there is a moderate "density" of liking overall. The analyst might also compare the cells above and below the diagonal to see if there is reciprocity in choices (e.g.

Bob chose Ted, did Ted choose Bob?). This is the second major emphasis of network analysis: seeing how the whole pattern of individual choices gives rise to more holistic patterns.

It is quite possible to think of the network data set in the same terms as "conventional data." One can think of the rows as simply a listing of cases, and the columns as attributes of each actor (i.e. the relations with other actors can be thought of as "attributes" of each actor). Indeed, many of the techniques used by network analysts (like calculating correlations and distances) are applied exactly the same way to network data as they would be to conventional data.

While it is possible to describe network data as just a special form of conventional data (and it is), network analysts look at the data in some rather fundamentally different ways. Rather than thinking about how an actor's ties with other actors describes the attributes of "ego," network analysts instead see a structure of connections, within which the actor is embedded. Actors are described by their relations, not by their attributes. And, the relations themselves are just as fundamental as the actors that they connect.

The major difference between conventional and network data is that conventional data focuses on actors and attributes; network data focus on actors and relations. The difference in emphasis is consequential for the choices that a researcher must make in deciding on research design, in conducting sampling, developing measurement, and handling the resulting data. It is not that the research tools used by network analysts are different from those of other social scientists (they mostly are not). But the special purposes and emphases of network research do call for some different considerations.

In this chapter, we will take a look at some of the issues that arise in design, sampling, and measurement for social network analysis. Our discussion will focus on the two parts of network data: nodes (or actors) and edges (or relations). We will try to show some of the ways in which network data are similar to, and different from more familar actor by attribute data. We will introduce some new terminology that makes it easier to describe the special features of network data. Lastly, we will briefly discuss how the differences between network and actor-attribute data are consequential for the application of statistical tools.

Return to the table of contents of this page

Nodes

Network data are defined by actors and by relations (or nodes and ties, etc.). The nodes or actors part of network data would seem to be pretty straight-forward. Other empirical approaches in the social sciences also think in terms of cases or subjects or sample elements and the like. There is one difference with most network data, however, that makes a big

difference in how such data are usually collected -- and the kinds of samples and populations that are studied.

Network analysis focuses on the relations among actors, and not individual actors and their attributes. This means that the actors are usually not sampled independently, as in many other kinds of studies (most typically, surveys). Suppose we are studying friendship ties, for example. John has been selected to be in our sample. When we ask him, John identifies seven friends. We need to track down each of those seven friends and ask them about their friendship ties, as well. The seven friends are in our sample because John is (and vice-versa), so the "sample elements" are no longer "independent."

The nodes or actors included in non-network studies tend to be the result of independent probability sampling. Network studies are much more likely to include all of the actors who occur within some (usually naturally occurring) boundary. Often network studies don't use "samples" at all, at least in the conventional sense. Rather, they tend to include all of the actors in some population or populations. Of course, the populations included in a network study may be a sample of some larger set of populations. For example, when we study patterns of interaction among students in a classrooms, we include all of the children in a classroom (that is, we study the whole population of the classroom). The classroom itself, though, might have been selected by probability methods from a population of classrooms (say all of those in a school).

The use of whole populations as a way of selecting observations in (many) network studies makes it important for the analyst to be clear about the boundaries of each population to be studied, and how individual units of observation are to be selected within that population. Network data sets also frequently involve several levels of analysis, with actors embedded at the lowest level (i.e. network designs can be described using the language of "nested" designs).

Return to the table of contents of this page

Populations, samples, and boundaries

Social network analysts rarely draw samples in their work. Most commonly, network analysts will identify some population and conduct a census (i.e. include all elements of the population as units of observation). A network analyst might examine all of the nouns and objects occurring in a text, all of the persons at a birthday party, all members of a kinship group, of an organization, neighborhood, or social class (e.g. landowners in a region, or royalty).

Survey research methods usually use a quite different approach to deciding which nodes to study. A list is made of all nodes (sometimes stratified or clustered), and individual elements

are selected by probability methods. The logic of the method treats each individual as a separate "replication" that is, in a sense, interchangeable with any other.

Because network methods focus on relations among actors, actors cannot be sampled independently to be included as observations. If one actor happens to be selected, then we must also include all other actors to whom our ego has (or could have) ties. As a result, network approaches tend to study whole populations by means of census, rather than by sample (we will discuss a number of exceptions to this shortly, under the topic of sampling ties).

The populations that network analysts study are remarkably diverse. At one extreme, they might consist of symbols in texts or sounds in verbalizations; at the other extreme, nations in the world system of states might constitute the population of nodes. Perhaps most common, of course, are populations of individual persons. In each case, however, the elements of the population to be studied are defined by falling within some boundary.

The boundaries of the populations studied by network analysts are of two main types. Probably most commonly, the boundaries are those imposed or created by the actors themselves. All the members of a classroom, organization, club, neighborhood, or community can constitute a population. These are naturally occuring clusters, or networks. So, in a sense, social network studies often draw the boundaries around a population that is known, *a priori*, to be a network. Alternatively, a network analyst might take a more "demographic" or "ecological" approach to defining population boundaries. We might draw observations by contacting all of the people who are found in a bounded spatial area, or who meet some criterion (having gross family incomes over $1,000,000 per year). Here, we might have reason to suspect that networks exist, but the entity being studied is an abstract aggregation imposed by the investigator -- rather than a pattern of institutionalized social action that has been identified and labeled by it's participants.

Network analysts can expand the boundaries of their studies by replicating populations. Rather than studying one neighborhood, we can study several. This type of design (which could use sampling methods to select populations) allows for replication and for testing of hypotheses by comparing populations. A second, and equally important way that network studies expand their scope is by the inclusion of multiple levels of analyis, or modalities.

Return to the table of contents of this page

Modality and levels of analysis

The network analyst tends to see individual people nested within networks of face-to-face relations with other persons. Often these networks of interpersonal relations become "social facts" and take on a life of their own. A family, for example, is a network of close relations

among a set of people. But this particular network has been institutionalized and given a name and reality beyond that of it's component nodes. Individuals in their work relations may be seen as nested within organizations; in their leisure relations they may be nested in voluntary associations. Neighborhoods, communities, and even societies are, to varying degrees, social entities in and of themselves. And, as social entities, they may form ties with the individuals nested within them, and with other social entities.

Often network data sets describe the nodes and relations among nodes for a single bounded population. If I study the friendship patterns among students in a classroom, I am doing a study of this type. But a classroom exists within a school - which might be thought of as a network relating classes and other actors (principals, administrators, librarians, etc.). And most schools exist within school districts, which can be thought of as networks of schools and other actors (school boards, research wings, purchasing and personnel departments, etc.). There may even be patterns of ties among school districts (say by the exchange of students, teachers, curricular materials, etc.).

Most networkers think of individual persons as being embedded in networks that are embedded in networks that are embedded in networks. Networkers describe such structures as "multi-modal." In our school example, individual students and teachers form one mode, classrooms a second, schools a third, and so on. A data set that contains information about two types of social entities (say persons and organizations) is a two mode network.

Of course, this kind of view of the nature of social structures is not unique to social networkers. Statistical analysts deal with the same issues as "hierarchical" or "nested" designs. Theorists speak of the macro-meso-micro levels of analysis, or develop schema for identifying levels of analysis (individual, group, organization, community, institution, society, global order being perhaps the most commonly used system in sociology). One advantage of network thinking and method is that it naturally predisposes the analyst to focus on multiple levels of analysis simultaneously. That is, the network analyst is always interested in how the individual is embedded within a structure and how the structure emerges from the micro-relations between individual parts. The ability of network methods to map such multi-modal relations is, at least potentially, a step forward in rigor.

Having claimed that social network methods are particularly well suited for dealing with multiple levels of analysis and multi-modal data structures, it must immediately be admitted that networkers rarely actually take much advantage. Most network analyses does move us beyond simple micro or macro reductionism -- and this is good. Few, if any, data sets and analyses, however, have attempted to work at more than two modes simultaneously. And, even when working with two modes, the most common strategy is to examine them more or less separately (one exception to this is the conjoint analysis of two mode networks).

Return to the table of contents of this page

Relations

The other half of the design of network data has to do with what ties or relations are to be measured for the selected nodes. There are two main issues to be discussed here. In many network studies, all of the ties of a given type among all of the selected nodes are studied -- that is, a census is conducted. But, sometimes different approaches are used (because they are less expensive, or because of a need to generalize) that sample ties. There is also a second kind of sampling of ties that always occurs in network data. Any set of actors might be connected by many different kinds of ties and relations (e.g. students in a classroom might like or dislike each other, they might play together or not, they might share food or not, etc.). When we collect network data, we are usually selecting, or sampling, from among a set of kinds of relations that we might have measured.

Return to the table of contents of this page

Sampling ties

Given a set of actors or nodes, there are several strategies for deciding how to go about collecting measurements on the relations among them. At one end of the spectrum of approaches are "full network" methods. This approach yields the maximum of information, but can also be costly and difficult to execute, and may be difficult to generalize. At the other end of the spectrum are methods that look quite like those used in conventional survey research. These approaches yield considerably less information about network structure, but are often less costly, and often allow easier generalization from the observations in the sample to some larger population. There is no one "right" method for all research questions and problems.

Full network methods require that we collect information about each actor's ties with all other actors. In essence, this approach is taking a census of ties in a population of actors -- rather than a sample. For example we could collect data on shipments of copper between all pairs of nation states in the world system from IMF records; we could examine the boards of directors of all public corporations for overlapping directors; we could count the number of vehicles moving between all pairs of cities; we could look at the flows of e-mail between all pairs of employees in a company; we could ask each child in a play group to identify their friends.

Because we collect information about ties between all pairs or dyads, full network data give a complete picture of relations in the population. Most of the special approaches and methods of network analysis that we will discuss in the remainder of this text were developed to be used with full network data. Full network data is necessary to properly define and measure many of the structural concepts of network analysis (e.g. between-ness).

Full network data allows for very powerful descriptions and analyses of social structures. Unfortunately, full network data can also be very expensive and difficult to collect. Obtaining data from every member of a population, and having every member rank or rate every other member can be very challenging tasks in any but the smallest groups. The task is made more manageable by asking respondents to identify a limited number of specific individuals with whom they have ties. These lists can then be compiled and cross-connected. But, for large groups (say all the people in a city), the task is practically impossible.

In many cases, the problems are not quite as severe as one might imagine. Most persons, groups, and organizations tend to have limited numbers of ties -- or at least limited numbers of strong ties. This is probably because social actors have limited resources, energy, time, and cognative capacity -- and cannot maintain large numbers of strong ties. It is also true that social structures can develop a considerable degree of order and solidarity with relatively few connections.

Snowball methods begin with a focal actor or set of actors. Each of these actors is asked to name some or all of their ties to other actors. Then, all the actors named (who were not part of the original list) are tracked down and asked for some or all of their ties. The process continues until no new actors are identified, or until we decide to stop (usually for reasons of time and resources, or because the new actors being named are very marginal to the group we are trying to study).

The snowball method can be particularly helpful for tracking down "special" populations (often numerically small sub-sets of people mixed in with large numbers of others). Business contact networks, community elites, deviant sub-cultures, avid stamp collectors, kinship networks, and many other structures can be pretty effectively located and described by snowball methods. It is sometimes not as difficult to achieve closure in snowball "samples" as one might think. The limitations on the numbers of strong ties that most actors have, and the tendency for ties to be reciprocated often make it fairly easy to find the boundaries.

There are two major potential limitations and weaknesses of snowball methods. First, actors who are not connected (i.e. "isolates") are not located by this method. The presence and numbers of isolates can be a very important feature of populations for some analytic purposes. The snowball method may tend to overstate the "connectedness" and "solidarity" of populations of actors. Second, there is no guaranteed way of finding all of the connected individuals in the population. Where does one start the snowball rolling? If we start in the wrong place or places, we may miss whole sub-sets of actors who are connected -- but not attached to our starting points.

Snowball approaches can be strengthened by giving some thought to how to select the initial nodes. In many studies, there may be a natural starting point. In community power studies, for example, it is common to begin snowball searches with the chief executives of large economic,

cultural, and political organizations. While such an approach will miss most of the community (those who are "isolated" from the elite network), the approach is very likely to capture the elite network quite effectively.

Ego-centric networks (with alter connections)

In many cases it will not be possible (or necessary) to track down the full networks beginning with focal nodes (as in the snowball method). An alternative approach is to begin with a selection of focal nodes (egos), and identify the nodes to which they are connected. Then, we determine which of the nodes identified in the first stage are connected to one another. This can be done by contacting each of the nodes; sometimes we can ask ego to report which of the nodes that it is tied to are tied to one another.

This kind of approach can be quite effective for collecting a form of relational data from very large populations, and can be combined with attribute-based approaches. For example, we might take a simple random sample of male college students and ask them to report who are their close friends, and which of these friends know one another. This kind of approach can give us a good and reliable picture of the kinds of networks (or at least the local neighborhoods) in which individuals are embedded. We can find out such things as how many connections nodes have, and the extent to which these nodes are close-knit groups. Such data can be very useful in helping to understand the opportunities and constraints that ego has as a result of the way they are embedded in their networks.

The ego-centered approach with alter connections can also give us some information about the network as a whole, though not as much as snowball or census approaches. Such data are, in fact, micro-network data sets -- samplings of local areas of larger networks. Many network properties -- distance, centrality, and various kinds of positional equivalence cannot be assessed with ego-centric data. Some properties, such as overall network density can be reasonably estimated with ego-centric data. Some properties -- such as the prevailence of reciprocal ties, cliques, and the like can be estimated rather directly.

Ego-centric networks (ego only)

Ego-centric methods really focus on the individual, rather than on the network as a whole. By collecting information on the connections among the actors connected to each focal ego, we can still get a pretty good picture of the "local" networks or "neighborhoods" of individuals. Such information is useful for understanding how networks affect individuals, and they also give a (incomplete) picture of the general texture of the network as a whole.

Suppose, however, that we only obtained information on ego's connections to alters -- but not information on the connections among those alters. Data like these are not really "network" data at all. That is, they cannot be represented as a square actor-by-actor array of ties. But

doesn't mean that ego-centric data without connections among the alters are of no value for analysts seeking to take a structural or network approach to understanding actors. We can know, for example, that some actors have many close friends and kin, and others have few. Knowing this, we are able to understand something about the differences in the actors places in social structure, and make some predictions about how these locations constrain their behavior. What we cannot know from ego-centric data with any certainty is the nature of the macro-structure or the whole network.

In ego-centric networks, the alters identified as connected to each ego are probably a set that is unconnected with those for each other ego. While we cannot assess the overall density or connectedness of the population, we can sometimes be a bit more general. If we have some good theoretical reason to think about alters in terms of their social roles, rather than as individual occupants of social roles, ego-centered networks can tell us a good bit about local social structures. For example, if we identify each of the alters connected to an ego by a friendship relation as "kin," "co-worker," "member of the same church," etc., we can build up a picture of the networks of social positions (rather than the networks of individuals) in which egos are embedded. Such an approach, of course, assumes that such categories as "kin" are real and meaningful determinants of patterns of interaction.

Return to the table of contents of this page

Multiple relations

In a conventional actor-by-trait data set, each actor is described by many variables (and each variable is realized in many actors). In the most common social network data set of actor-by-actor ties, only one kind of relation is described. Just as we often are interested in multiple attributes of actors, we are often interested in multiple kinds of ties that connect actors in a network.

In thinking about the network ties among faculty in an academic department, for example, we might be interested in which faculty have students in common, serve on the same committees, interact as friends outside of the workplace, have one or more areas of expertese in common, and co-author papers. The positions that actors hold in the web of group affiliations are multi-faceted. Positions in one set of relations may re-enforce or contradict positions in another (I might share friendship ties with one set of people with whom I do not work on committees, for example). Actors may be tied together closely in one relational network, but be quite distant from one another in a different relational network. The locations of actors in multi-relational networks and the structure of networks composed of multiple relations are some of the most interesting (and still relatively unexplored) areas of social network analysis.

When we collect social network data about certain kinds of relations among actors we are, in a

sense, sampling from a population of possible relations. Usually our research question and theory indicate which of the kinds of relations among actors are the most relevant to our study, and we do not sample -- but rather select -- relations. In a study concerned with economic dependency and growth, for example, I could collect data on the exchange of performances by musicians between nations -- but it is not really likely to be all that relevant.

If we do not know what relations to examine, how might we decide? There are a number of conceptual approaches that might be of assistance. Systems theory, for example, suggests two domains: material and informational. Material things are "conserved" in the sense that they can only be located at one node of the network at a time. Movements of people between organizations, money between people, automobiles between cities, and the like are all examples of material things which move between nodes -- and hence establish a network of material relations. Informational things, to the systems theorist, are "non-conserved" in the sense that they can be in more than one place at the same time. If I know something and share it with you, we both now know it. In a sense, the commonality that is shared by the exchange of information may also be said to establish a tie between two nodes. One needs to be cautious here, however, not to confuse the simple possession of a common attribute (e.g. gender) with the presence of a tie (e.g. the exchange of views between two persons on issues of gender).

Methodologies for working with multi-relational data are not as well developed as those for working with single relations. Many interesting areas of work such as network correlation, multi-dimensional scaling and clustering, and role algebras have been developed to work with multi-relational data. For the most part, these topics are beyond the scope of the current text, and are best approached after the basics of working with single relational networks are mastered.

Return to the table of contents of this page

Scales of measurement

Like other kinds of data, the information we collect about ties between actors can be measured (i.e. we can assign scores to our observations) at different "levels of measurement." The different levels of measurement are important because they limit the kinds of questions that can be examined by the researcher. Scales of measurement are also important because different kinds of scales have different mathematical properties, and call for different algorithms in describing patterns and testing inferences about them.

It is conventional to distinguish nominal, ordinal, and interval levels of measurement (the ratio level can, for all practical purposes, be grouped with interval). It is useful, however, to further divide nominal measurement into binary and multi-category variations; it is also useful to distinguish between full-rank ordinal measures and grouped ordinal measures. We will briefly

describe all of these variations, and provide examples of how they are commonly applied in social network studies.

***Binary measures of relations*:** By far the most common approach to scaling (assigning numbers to) relations is to simply distinguish between relations being absent (coded zero), and ties being present (coded one). If we ask respondents in a survey to tell us "which other people on this list do you like?" we are doing binary measurement. Each person from the list that is selected is coded one. Those who are not selected are coded zero.

Much of the development of graph theory in mathematics, and many of the algorithms for measuring properties of actors and networks have been developed for binary data. Binary data is so widely used in network analysis that it is not unusual to see data that are measured at a "higher" level transformed into binary scores before analysis proceeds. To do this, one simply selects some "cut point" and re-scores cases as below the cut-point (zero) or above it (one). Dichotomizing data in this way is throwing away information. The analyst needs to consider what is relevant (i.e. what is the theory about? is it about the presence and pattern of ties, or about the strengths of ties?), and what algorithms are to be applied in deciding whether it is reasonable to recode the data. Very often, the additional power and simplicity of analysis of binary data is "worth" the cost in information lost.

Multiple-category nominal measures of relations: In collecting data we might ask our respondents to look at a list of other people and tell us: "for each person on this list, select the category that describes your relationship with them the best: friend, lover, business relationship, kin, or no relationship." We might score each person on the list as having a relationship of type "1" type "2" etc. This kind of a scale is nominal or qualitative -- each person's relationship to the subject is coded by it's type, rather than it's strength. Unlike the binary nominal (true-false) data, the multiple category nominal measure is multiple choice.

The most common approach to analyzing multiple-category nominal measures is to use it to create a series of binary measures. That is, we might take the data arising from the question described above and create separate sets of scores for friendship ties, for lover ties, for kin ties, etc. This is very similar to "dummy coding" as a way of handling multiple choice types of measures in statistical analysis. In examining the resulting data, however, one must remember that each node was allowed to have a tie in at most one of the resulting networks. That is, a person can be a friendship tie or a lover tie -- but not both -- as a result of the way we asked the question. In examining the resulting networks, densities may be artificially low, and there will be an inherent negative correlation among the matrices.

This sort of multiple choice data can also be "binarized." That is, we can ignore what kind of tie is reported, and simply code whether a tie exists for a dyad, or not. This may be fine for some analyses -- but it does waste information. One might also wish to regard the types of ties as reflecting some underlying continuous dimension (for example, emotional intensity). The types

of ties can then be scaled into a single grouped ordinal measure of tie strength. The scaling, of course, reflects the predispositions of the analyst -- not the reports of the respondents.

Grouped ordinal measures of relations: One of the earliest traditions in the study of social networks asked respondents to rate each of a set of others as "liked" "disliked" or "neutral." The result is a grouped ordinal scale (i.e., there can be more than one "liked" person, and the categories reflect an underlying rank order of intensity). Usually, this kind of three point scale was coded -1, 0, and +1 to reflect negative liking, indifference, and positive liking. When scored this way, the pluses and minuses make it fairly easy to write algorithms that will count and describe various network properties (e.g. the structural balance of the graph).

Grouped ordinal measures can be used to reflect a number of different quantitative aspects of relations. Network analysts are often concerned with describing the "strength" of ties. But, "strength" may mean (some or all of) a variety of things. One dimension is the frequency of interaction -- do actors have contact daily, weekly, monthly, etc. Another dimension is "intensity," which usually reflects the degree of emotional arousal associated with the relationship (e.g. kin ties may be infrequent, but carry a high "emotional charge" because of the highly ritualized and institutionalized expectations). Ties may be said to be stronger if they involve many different contexts or types of ties. Summing nominal data about the presence or absence of multiple types of ties gives rise to an ordinal (actually, interval) scale of one dimension of tie strength. Ties are also said to be stronger to the extent that they are reciprocated. Normally we would assess reciprocity by asking each actor in a dyad to report their feelings about the other. However, one might also ask each actor for their perceptions of the degree of reciprocity in a relation: Would you say that neither of you like each other very much, that you like X more than X likes you, that X likes you more than you like X, or that you both like each other about equally?

Ordinal scales of measurement contain more information than nominal. That is, the scores reflect finer gradations of tie strength than the simple binary "presence or absence." This would seem to be a good thing, yet it is frequently difficult to take advantage of ordinal data. The most commonly used algorithms for the analysis of social networks have been designed for binary data. Many have been adapted to continuous data -- but for interval, rather than ordinal scales of measurement. Ordinal data, consequently, are often binarized by choosing some cut-point and re-scoring. Alternatively, ordinal data are sometimes treated as though they really were interval. The former strategy has some risks, in that choices of cut-points can be consequential; the latter strategy has some risks, in that the intervals separating points on an ordinal scale may be very heterogeneous.

Full-rank ordinal measures of relations: Sometimes it is possible to score the strength of all of the relations of an actor in a rank order from strongest to weakest. For example, I could ask each respondent to write a "1" next to the name of the person in the class that you like the most, a "2" next to the name of the person you like next most, etc. The kind of scale that would result from this would be a "full rank order scale." Such scales reflect differences in degree of

intensity, but not necessarily equal differences -- that is, the difference between my first and second choices is not necessarily the same as the difference between my second and third choices. Each relation, however, has a unique score (1st, 2nd, 3rd, etc.).

Full rank ordinal measures are somewhat uncommon in the social networks research literature, as they are in most other traditions. Consequently, there are relatively few methods, definitions, and algorithms that take specific and full advantage of the information in such scales. Most commonly, full rank ordinal measures are treated as if they were interval. There is probably somewhat less risk in treating fully rank ordered measures (compared to grouped ordinal measures) as though they were interval, though the assumption is still a risky one. Of course, it is also possible to group the rank order scores into groups (i.e. produce a grouped ordinal scale) or dichotomize the data (e.g. the top three choices might be treated as ties, the remainder as non-ties). In combining information on multiple types of ties, it is frequently necessary to simplify full rank order scales. But, if we have a number of full rank order scales that we may wish to combine to form a scale (i.e. rankings of people's likings of other in the group, frequency of interaction, etc.), the sum of such scales into an index is plausibly treated as a truly interval measure.

Interval measures of relations: The most "advanced" level of measurement allows us to discriminate among the relations reported in ways that allow us to validly state that, for example, "this tie is twice as strong as that tie." Ties are rated on scales in which the difference between a "1" and a "2" reflects the same amount of real difference as that between "23" and "24."

True interval level measures of the strength of many kinds of relationships are fairly easy to construct, with a little imagination and persistence. Asking respondents to report the details of the frequency or intensity of ties by survey or interview methods, however, can be rather unreliable -- particularly if the relationships being tracked are not highly salient and infrequent. Rather than asking whether two people communicate, one could count the number of email, phone, and inter-office mail deliveries between them. Rather than asking whether two nations trade with one another, look at statistics on balances of payments. In many cases, it is possible to construct interval level measures of relationship strength by using artifacts (e.g. statistics collected for other purposes) or observation.

Continuous measures of the strengths of relationships allow the application of a widerl range of mathematical and statistical tools to the exploration and analysis of the data. Many of the algorithms that have been developed by social network analysts, originally for binary data, have been extended to take advantage of the information available in full interval measures. Whenever possible, connections should be measured at the interval level -- as we can always move to a less refined approach later; if data are collected at the nominal level, it is much more difficult to move to a more refined level.

Even though it is a good idea to measure relationship intensity at the most refined level possible, most network analysis does not operate at this level. The most powerful insights of network analysis, and many of the mathematical and graphical tools used by network analysts were developed for simple graphs (i.e. binary, undirected). Many characterizations of the embeddedness of actors in their networks, and of the networks themselves are most commonly thought of in discrete terms in the research literature. As a result, it is often desirable to reduce even interval data to the binary level by choosing a cutting -point, and coding tie strength above that point as "1" and below that point as "0." Unfortunately, there is no single "correct" way to choose a cut-point. Theory and the purposes of the analysis provide the best guidance. Sometimes examining the data can help (maybe the distribution of tie strengths really is discretely bi-modal, and displays a clear cut point; maybe the distribution is highly skewed and the main feature is a distinction between no tie and any tie). When a cut-point is chosen, it is wise to also consider alternative values that are somewhat higher and lower, and repeat the analyses with different cut-points to see if the substance of the results is affected. This can be very tedious, but it is very necessary. Otherwise, one may be fooled into thinking that a real pattern has been found, when we have only observed the consequences of where we decided to put our cut-point.

Return to the table of contents of this page

A note on statistics and social network data

Social network analysis is more a branch of "mathematical" sociology than of "statistical or quantitative analysis," though networkers most certainly practice both approaches. The distinction between the two approaches is not clear cut. Mathematical approaches to network analysis tend to treat the data as "deterministic." That is, they tend to regard the measured relationships and relationship strengths as accurately reflecting the "real" or "final" or "equilibrium" status of the network. Mathematical types also tend to assume that the observations are not a "sample" of some larger population of possible observations; rather, the observations are usually regarded as the population of interest. Statistical analysts tend to regard the particular scores on relationship strengths as stochastic or probabilistic realizations of an underlying true tendency or probability distribution of relationship strengths. Statistical analysts also tend to think of a particular set of network data as a "sample" of a larger class or population of such networks or network elements -- and have a concern for the results of the current study would be reproduced in the "next" study of similar samples.

In the chapters that follow in this text, we will mostly be concerned with the "mathematical" rather than the "statistical" side of network analysis (again, it is important to remember that I am over-drawing the differences in this discussion). Before passing on to this, we should note a couple main points about the relationship between the material that you will be studying here, and the main statistical approaches in sociology.

In one way, there is little apparent difference between conventional statistical approaches and network approaches. Univariate, bi-variate, and even many multivariate descriptive statistical tools are commonly used in the describing, exploring, and modeling social network data. Social network data are, as we have pointed out, easily represented as arrays of numbers -- just like other types of sociological data. As a result, the same kinds of operations can be performed on network data as on other types of data. Algorithms from statistics are commonly used to describe characteristics of individual observations (e.g. the median tie strength of actor X with all other actors in the network) and the network as a whole (e.g. the mean of all tie strengths among all actors in the network). Statistical algorithms are very heavily used in assessing the degree of similarity among actors, and if finding patterns in network data (e.g. factor analysis, cluster analysis, multi-dimensional scaling). Even the tools of predictive modeling are commonly applied to network data (e.g. correlation and regression).

Descriptive statistical tools are really just algorithms for summarizing characteristics of the distributions of scores. That is, they are mathematical operations. Where statistics really become "statistical" is on the inferential side. That is, when our attention turns to assessing the reproducibility or likelihood of the pattern that we have described. Inferential statistics can be, and are, applied to the analysis of network data. But, there are some quite important differences between the flavors of inferential statistics used with network data, and those that are most commonly taught in basic courses in statistical analysis in sociology.

Probably the most common emphasis in the application of inferential statistics to social science data is to answer questions about the stability, reproducibility, or generalizability of results observed in a single sample. The main question is: if I repeated the study on a different sample (drawn by the same method), how likely is it that I would get the same answer about what is going on in the whole population from which I drew both samples? This is a really important question -- because it helps us to assess the confidence (or lack of it) that we ought to have in assessing our theories and giving advice.

To the extent the observations used in a network analysis are drawn by probability sampling methods from some identifyable population of actors and/or ties, the same kind of question about the generalizability of sample results applies. Often this type of inferential question is of little interest to social network researchers. In many cases, they are studying a particular network or set of networks, and have no interest in generalizing to a larger population of such networks (either because there isn't any such population, or we don't care about generalizing to it in any probabilistic way). In some other cases we may have an interest in generalizing, but our sample was not drawn by probability methods. Network analysis often relies on artifacts, direct observation, laboratory experiments, and documents as data sources -- and usually there are no plausible ways of identifying populations and drawing samples by probability methods.

The other major use of inferential statistics in the social sciences is for testing hypotheses. In

many cases, the same or closely related tools are used for questions of assessing generalizability and for hypothesis testing. The basic logic of hypothesis testing is to compare an observed result in a sample to some null hypothesis value, relative to the sampling variability of the result under the assumption that the null hypothesis is true. If the sample result differs greatly from what was likely to have been observed under the assumption that the null hypothesis is true -- then the null hypothesis is probably not true.

The key link in the inferential chain of hypothesis testing is the estimation of the standard errors of statistics. That is, estimating the expected amount that the value a a statistic would "jump around" from one sample to the next simply as a result of accidents of sampling. We rarely, of course, can directly observe or calculate such standard errors -- because we don't have replications. Instead, information from our sample is used to estimate the sampling variability.

With many common statistical procedures, it is possible to estimate standard errors by well validated approximations (e.g. the standard error of a mean is usually estimated by the sample standard deviation divided by the square root of the sample size). These approximations, however, hold when the observations are drawn by independent random sampling. Network observations are almost always non-independent, by definition. Consequently, conventional inferential formulas do not apply to network data (though formulas developed for other types of dependent sampling may apply). It is particularly dangerous to assume that such formulas do apply, because the non-independence of network observations will usually result in under-estimates of true sampling variability -- and hence, too much confidence in our results.

The approach of most network analysts interested in statistical inference for testing hypotheses about network properties is to work out the probability distributions for statistics directly. This approach is used because: 1) no one has developed approximations for the sampling distributions of most of the descriptive statistics used by network analysts and 2) interest often focuses on the probability of a parameter relative to some theoretical baseline (usually randomness) rather than on the probability that a given network is typical of the population of all networks.

Suppose, for example, that I was interested in the proportion of the actors in a network who were members of cliques (or any other network statistic or parameter). The notion of a clique implies structure -- non-random connections among actors. I have data on a network of ten nodes, in which there are 20 symmetric ties among actors, and I observe that there is one clique containing four actors. The inferential question might be posed as: how likely is it, if ties among actors were purely random events, that a network composed of ten nodes and 20 symmetric ties would display one or more cliques of size four or more? If it turns out that cliques of size four or more in random networks of this size and degree are quite common, I should be very cautious in concluding that I have discovered "structure" or non-randomness. If it turns out that such cliques (or more numerous or more inclusive ones) are very unlikely under the assumption that ties are purely random, then it is very plausible to reach the

conclusion that there is a social structure present.

But how can I determine this probability? The method used is one of simulation -- and, like most simulation, a lot of computer resources and some programming skills are often necessary. In the current case, I might use a table of random numbers to distribute 20 ties among 10 actors, and then search the resulting network for cliques of size four or more. If no clique is found, I record a zero for the trial; if a clique is found, I record a one. The rest is simple. Just repeat the experiment several thousand times and add up what proportion of the "trials" result in "successes." The probability of a success across these simulation experiments is a good estimator of the likelihood that I might find a network of this size and density to have a clique of this size "just by accident" when the non-random causal mechanisms that I think cause cliques are not, in fact, operating.

This may sound odd, and it is certainly a lot of work (most of which, thankfully, can be done by computers). But, in fact, it is not really different from the logic of testing hypotheses with non-network data. Social network data tend to differ from more "conventional" survey data in some key ways: network data are often not probability samples, and the observations of individual nodes are not independent. These differences are quite consequential for both the questions of generalization of findings, and for the mechanics of hypothesis testing. There is, however, nothing fundamentally different about the logic of the use of descriptive and inferential statistics with social network data.

The application of statistics to social network data is an interesting area, and one that is, at the time of this writing, at a "cutting edge" of research in the area. Since this text focuses on more basic and commonplace uses of network analysis, we won't have very much more to say about statistics beyond this point. You can think of much of what follows here as dealing with the "descriptive" side of statistics (developing index numbers to describe certain aspects of the distribution of relational ties among actors in networks). For those with an interest in the inferential side, a good place to start is with the second half of the excellent Wasserman and Faust textbook.

Return to the table of contents of this page
Return to the table of contents of the textbook
