

## Commentary: Sampling in Social Networks

Richard B. Rothenberg

*Emory University School of Medicine*

In classic statistical theory, if a random sample is drawn from a population whose underlying distribution is known, it may be assumed that the properties of the sample mirror those of the population (Snedecor and Cochran, 1972). On that cornerstone is built a statistical superstructure that permits estimation, hypothesis testing, assurance of internal validity, generalizability, and modeling. For a variety of actual sampling schemes — simple random, stratified, probability proportional to size, systematic, cluster, multistage — considerable mathematical work has established appropriate point estimate and variance formulas, and has defined the potential for bias and other threats to validity (Levy and Lemeshow, 1980). This body of work provides satisfying precision for the estimation of uncertainty in defining population characteristics.

### Random Graphs

In the field of network analysis, sampling theory has been associated with defining the mathematical properties of random graphs. Though others preceded them, Erdos and Renyi (1959, 1960) are credited with establishing the theoretical base for estimation of such properties. During the past several decades considerable effort has been invested in describing graphs, and many familiar properties of social network have been established for random graphs. Investigators have explored the mean and variance of degree in a graph (Frank, 1980; Rapoport, 1979a); the probability that a graph will be connected (Gilbert, 1959); the distribution of connected components in a graph (Frank, 1978a; Ling, 1975; Naus and Rabinowitz, 1975); and general types of estimation in large graphs under various sampling schemes (Frank, 1980, 1981, 1978b). One specific type of network investigation — snowball sampling (Goodman, 1961) — has also been the subject of theoretical mathematical investigation (Frank, 1979; Doreian and Woodward, 1992; Johnson *et al.*, 1989).

These investigations are cast in classic probabilistic terms. For example, the early study of Gilbert (1959) assumed a set of random points (1,2...N) through which  $N(N-1)/2$  lines could be drawn connecting them, creating  $2^{N(N-1)/2}$  possible graphs. A sample graph is obtained by making random, independent choices to draw a line between two points, the choice being made with common probability  $p$ . The author then derives expressions, for both small and large  $N$ , for the probability that the graph is connected (meaning that there is a path of some length between every point  $i$  and every point  $j$ ) and the probability that two points are connected. Similarly, in much of his work, Frank (1978a, 1978b, 1980, 1981) assumes an unknown, undirected graph of order  $N$  (i.e.,  $N$  nodes) and size  $R$  (i.e.  $R$  lines) which is sampled using simple random sampling (assuming a uniform distribution), or using Bernoulli sampling (assuming a selection probability of  $p$ ). He further assumes that the graph is simple, with no loops or multiple edges, and that contacts between different pairs of individuals are independent.

The tension between these assumptions and the pragmatic problems of obtaining network information are obvious, and thoroughly appreciated by those who invoke such assumptions (Frank, 1981; Rapoport, 1979b). In an overview of the approach to random networks, Rapoport says: "Mathematical modeling is a vehicle for absolutely rigorous reasoning and therein lies its advantage. A disadvantage of mathematical modeling is that it necessitates a drastic paring down of the details of the phenomena modeled...these simplifications...can impair or altogether destroy the model's pragmatic relevance."

## Network Sampling Schemes

Several sampling schemes have been proposed to bridge the gap between theory and practice. In 1976, Granovetter proposed multiple random samples (with replacement) of size  $n$  from a known population of size  $N$ . He attempted to show that the average density of the samples (density defined as the proportion of actual connections out of all possible connections) was a good estimate of the density of  $N$ . Such sampling thus provides a measure of the amount of interaction within a population. Multiple sampling has considerable intuitive appeal, and appears related to the concept of "triangulation," but its tie to classic probability theory may break down in the absence of knowledge about  $N$ .

As noted, snowball sampling has considerable theoretic appeal. The snowball procedure is defined as one that enlarges an original node sample by joining adjacent nodes (Frank, 1979): A final snowball sample with  $s$  stages places the outermost nodes at a distance  $s$  from the innermost node (assuming no recursion). Frank builds his mathematical theory by assuming that the sampling design of the snowball sample is simply an extension of the sampling design of the first stage. The fundamental assumptions are thus tied to underlying probability theory but are detached from some of the pragmatic realities. Some more pragmatic aspects of snowball sampling have been explored, however. Johnson *et al.* (1989) performed simulations to determine the relative impact of changes in the sample size, the number of stages, and the number of contacts named at each stage on the estimate of indegree centrality. They used two hypothetical 40x40 matrices, one of which had several small elite subgroups, and one of which had greater variation in the indegree of network members. The results were not straightforward, but did demonstrate an important trade-off between the number of stages (an expensive parameter) and the number of contacts (an inexpensive parameter). The model they use is certainly of heuristic value, and may serve investigators in preparing study designs, but is still dependant on preexisting knowledge about the underlying population. A more empirical study was conducted by Doreian and Woodard (1992) who demonstrated that use of a fixed list of agencies produced a network of very different characteristics from that produced by a snowball approach. They note the particular value of the snowball approach in responding to changing network circumstances, flexibility not available with preexisting lists. Though this latter study provides important information for the pragmatic aspects of network analysis, it is somewhat detached from the theoretical statistical basis for estimation.

In 1977, Klovdahl suggested a method for sampling that has the potential to serve pragmatic purposes yet retains a probabilistic base. He suggested a modification of snowball sampling wherein one contact is chosen at random from those named at each stage of sampling (thus the notion of a random walk). He demonstrated the ability of this technique to identify important network relationships in a large urban environment (Klovdahl, 1989). The formal

mathematical relationships engendered by this sampling technique have not been elucidated, however.

Thus, the connection between the mathematics of sampling and the exigencies of network research has been elusive. In many practical network situations--especially those involving rare or hidden populations —  $N$  is unknown (that is, the source population is difficult to define with regard to size, location, stability, and underlying distribution), and  $n$  is a nonrandom, nonprobabilistic sample that may or may not be representative, and whose statistical properties are unknown. In the absence of a well defined population (consider, for example, the multifocal nexi of drug users in a larger urban area, or the small invisible clusters of drug users in rural areas), it is self-evident that a probability sample is not possible. In the absence of a probability sample, the statistical superstructure collapses and, in principle, desirable statistical properties are not available to the investigator. The subsequent use of statistical tests that rest on assumptions of random sampling from a known underlying distribution is problematic. The absence of a statistical cornerstone has been a concern of investigators in the field and a source of skepticism for those in other disciplines.

### **Ascending Methods**

The obvious alternative to probability sampling is to go directly to persons of interest and ask them about their networks. Such egocentric data can be combined for sociometric purposes, and the structure of a group of people with characteristics of interest can be developed. A number of terms have been used to describe this procedure: multiplicity sampling (Sirken and Levy, 1974); site sampling (TenHouten *et al.*, 1971); targeted sampling (Watters and Biernacki, 1989); key informant sampling (Deaux and Callaghan, 1984; 1985); purposive sampling (Warwick and Lininger, 1975); strategic sampling (Hunt, 1970); judgment sampling (Honigmann, 1970; Bernard, 1988; Peltó and Peltó, 1979); and dimensional sampling (Arnold, 1970). (Note that most authors are careful to distinguish these procedures from convenience sampling, a process by which one recruits only those easily available, e.g. drug users in treatment clinics as opposed to drug users on the street.) Several reviews contrast and compare these methods (Watters and Biernacki, 1989; Johnson, 1990; Spreen, 1992), and make it clear that these mechanisms differ in detail, but not in their conceptual framework. (The exception is multiplicity sampling, a probability-based procedure that can be used as part of large sample surveys to study uncommon events.)

The basic approach embodied in these methods is typified by targeted sampling (Watters and Biernacki, 1989), a procedure that attempts to combine several mechanisms. The authors describe a systematic set of steps used to identify a population of interest and construct the social network from informants. They first use epidemiologic methods to describe geographic areas based on readily available aggregate data. They then use ethnographic methods to describe the population with regard to approximate size, location, and characteristics. Participants are then recruited through the active efforts of street outreach workers, using "chain-referral sampling." The authors stress the need to use interim findings to shape research questions, and the need for a flexible approach that can respond to changing information. Note that this flexibility is analogous to the techniques of sequential analysis used for probability based trials.

The concept that unites these methods is that of ascending data gathering (Erdos and Renyi, 1960), best described through comparison with its obverse, descending data gathering.

The latter begins with the total population and subdivides it by a classification system based on available data (e.g., age, sex, race). Successively smaller groups are increasingly homogeneous with regard to the variables used, and increasingly distant from other groups. In contrast, ascending methods start with the universe of single individuals and attempts to combine them based on variables that are not readily available (for example, drugs users in a small rural population center). The resultant groups may well be heterogeneous with regard to the readily available descending variables, and could not have been constructed by the descending route. The fundamental difference between the two approaches is that the descending method chooses a sample; the ascending method constructs a population.

## Relating Probability and Ascending Methods

Investigators who use ascending techniques face three important questions that link their method to probability sampling. These questions deal with issues of estimation of population totals from samples, of validity, and of generalizability.

*Estimation: What is the true size of the group that has been constructed?* The population reached in a study is an unknown subset of some larger group, but not a sample in the usual statistical sense. It may have intrinsic coherence, but it cannot be used to estimate the size of a larger group without invoking other techniques (e.g. capture-recapture methods, or demonstration by multiple means that the potential for further members is exhausted).

*Validity: How representative is this subset of the total group from which it is drawn?* The representativeness of the subset cannot be calculated directly. In a sense, the subset defines itself, and the next addition to it (i.e., the next observation) can be judged for probability that it is a sample observation from the known subset or from some other group. (For example, in a simple case, an observation  $x_i$  has a vector of characteristics  $q_{mnpqr}$ . The subset of interest has already been determined to have certain proportions of these characteristics:  $p_m, p_n, p_q,$  and  $p_r$ . The corresponding proportions in the overall population (used for testing the alternate hypothesis) would be  $p_M, p_N, p_O,$  and  $p_R$ . The observation  $q_{mnpqr}$  can be tested to see if it is a random sample from a population with a joint probability  $p_m p_n p_q p_r$  (null hypothesis) or from some alternate population  $p_M p_N p_O p_R$ . After the subset has reached a certain size, sequential testing of the new  $x_i$ 's can provide the investigator with increasing assurance that new recruits confirm the characteristics of the subset. Contrarily, the arrival of  $x_i$ 's that differ from the subset signals incompleteness or heretofore unsuspected diversity. In either case, this approach provides the investigator with a basis for saying that the subset is representative of some larger group (of unknown size) and provides the basis for a stopping rule. Note, however, that this procedure simply formalizes information that would be immediately evident to a field worker: the arrival of a new person who is different from the group already recruited. Again, as noted, there is a fundamental analogy between ethnographic and recruitment methods and sequential analysis.

*Generalizability: How much is the observed group like other groups?* The question cannot be answered only with information on the subset studied (for example: it would difficult to assume that drug users in rural population centers are like drug users in inner city urban settings). If, however, the general characteristics of a new area are similar to those of one already studied, and if (less expensive) ethnographic investigation is confirmatory, network characteristics may be assumed, obviating the need for more expensive investigation. The establishment of network

typologies, and their application in appropriate sociodemographic settings, is analogous to the creation of synthetic estimates in probability sampling procedures (Levy and Lemeshow 1980).

Thus, the ascending methodologies represent an alternate paradigm for investigating populations, but have analogies with the methods of probability sampling. A major difference is, as noted, that the latter provide a high level of precision in estimating uncertainty. The level of precision is lower for ascending methods. The approaches would appear to occupy different portions of the measurement continuum, and the difference between them is more quantitative than qualitative.

But analogies are not identities, and the use of a statistical superstructure for data collected under alternate assumptions may be problematic. Many investigators appear to feel that, despite the violation of assumptions, the use of statistical methods provides some interesting results, and connects social network analysis to a common scientific language. Once having chosen a purposive (or judgment or targeted) sample, investigators use it as if it were representative, and manipulate it accordingly.

If, however, one can accept the notion that the subset of persons studied as part of an ascending method represents a population, the application of statistical methods may not be unacceptable. If, in addition, one can verify that this "population" bears a definable relationship to some larger group, the techniques are even more justified. In some network studies the assumption that the studied group is a population is literally true (the members of a corporate hierarchy, or the inmates in a detention center). In other studies, it is a *de facto* premise: by ending data collection, the investigators place a boundary around the network they have described, and look within. Network measures (such as connected components or centrality measures) are calculated and a variety of statistical and epidemiologic techniques are used within the group.

### **An Alternative Paradigm**

But as noted, in the ascending approaches, estimation about some larger population (to which this smaller population is related) and generalizability to other populations cannot be based on sampling theory. Estimation and generalizability rest, rather, on the use of other methods to describe the larger population and to provide comparisons with the smaller population observed. These other methods are based on the substitution of epidemiologic and ethnographic observations for a numerical enumeration of the source population. Information about the source population is obtained through empirical observation; information about the subset is compared to this empirical "standard." A fundamental strength of this approach is that ethnographic assessment is ongoing, and presumes that the source population will change. The validity of a subset is not bound to a snapshot of the source.

At first blush, this appears to be an *a posteriori* justification for an approach with which we are stuck. It certainly does not meet the strict probabilistic criteria that are traditionally used in sampling theory. But, in actual practice, sampling is often far closer to the new paradigm than the traditional statistical one. It is well recognized, for example, that a random or probabilistic sample has a certain, definable probability of not being representative of the underlying population. In addition, we often over- or under-sample specifically to meet special analytic needs or because of an *a priori* understanding of the characteristics of the underlying population

and the difficulty of sampling some portion of it. Finally, we are often unsure of the size, location, or distribution of the source population. Pretending is not confined to network sampling.

It is perhaps paradoxical that we find more certainty in chance than in empiricism, and that a mechanized statistical procedure is preferred to continuing observation. The preference may arise from two observations: (1) using ethnographic approaches to validate the representativeness of a network may not furnish a familiar and satisfying statistic; (2) ethnographic methods are unfamiliar to many and can leave a sense of incompleteness or uncertainty. The challenge for investigators is to stop pretending that network sampling uses standard statistical methodology, and to improve the credibility of the alternate paradigm.

## References

- Arnold, D.O. 1970. "Dimensional sampling: An approach for studying a small number of cases". *American Sociologist* 5:147-50.
- Bernard, H.R. 1988. *Research Methods in Cultural Anthropology*. Newbury Park, CA: Sage.
- Deaux, E. and Callaghan J.W. 1984. "Estimating statewide health risk behavior: a comparison of telephone and key informant survey approaches". *Evaluation Review* 8:467-92.
- Deaux, E. and Callaghan, J.W. 1985. "Key informant versus self-report estimates of health risk behavior". *Evaluation Review* 9:365-8.
- Doreian, P., and Woodard, K.L. 1992. "Fixed list versus snowball selection of social networks". *Social Science Research* 21:216-33.
- Erdos, P. and Renyi, A. 1959. "On random graphs". *Publicationes Mathematicae Debrecen* 6:290-7.
- Erdos, P. and Renyi, A. 1960. "On the evolution of random graphs". *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5:17-61.
- Frank, O. 1978a. "Estimation of the number of connected components in a graph by using a sampled subgraph". *Scandinavian Journal of Statistics* 5:177-88.
- Frank, O. 1978b. "Sampling and estimation in large social networks". *Social Networks* 1:91-101.
- Frank, O. 1979. "Estimation of population totals by use of snowball samples". *Perspectives on Social Network Research*. 319-47.
- Frank, O. 1980. "Estimation of the number of vertices of different degrees in a graph". *Journal of Statistical Planning and Inference* 4:45-50.
- Frank, O. 1981. "A survey of statistical methods for graph analysis". In Leinhardt SL (ed) *Sociological Methodology* [Abstract].
- Gilbert, E.N. 1959. "Random graphs". *Annals of Mathematical Statistics* 30:1141-14.
- Goodman, L.A. 1961. "Snowball sampling". *Annals of Mathematical Statistics* 32:148-70.
- Granovetter, M. 1976. "Network sampling: Some first steps". *American Journal of Sociology* 81:1267-303.
- Honigmann, J.J. 1970. "Sampling in ethnographic fieldwork". *Handbook of Method in Cultural Anthropology*. New York: Columbia University Press.

- Hunt, R.G. 1970. *Strategic Selection: A purposive sampling design for small numbers research, program evaluation, and management*. Buffalo: State University of New York, Survey Research Center.
- Johnson, J.C. 1990. "Selecting ethnographic informants". A Sage University Paper. In: *Anonymous Qualitative Research Methods Series No. 22*. Newbury Park, California: Sage Publishing Inc.
- Johnson, J.C., Boster, J.S., and Holbert, D. 1989. "Estimating relational attributes from snowball samples through simulation". *Social Networks* 11:135-58.
- Klovdahl, A.S. 1977. "Social networks in an urban area: First Canberra study". *Australian and New Zealand Journal of Sociology* 13:169-75.
- Klovdahl, A.S. 1989. *Urban social network: Some methodological problems and possibilities*. Ablex Publishing Corp. 176-210.
- Levy, P.S. and Lemeshow, S. 1980. *Sampling for Health Professionals*. Belmont, CA: Lifetime Learning Publications.
- Ling, R.F. 1975. "An exact probability distribution on the connectivity of random graphs". *Journal of Mathematical Psychology* 12:90-8.
- Naus, J.I. and Rabinowitz, L. 1975. "The expectation and variance of the number of components in random linear graphs". *The Annals of Probability* 3(1):159-61.
- Pelto, P.J. and Pelto G.H. 1979. *Anthropological Research: The Structure of Inquiry*. Cambridge: Cambridge University Press.
- Rapoport, A. 1979a. "A probabilistic approach to networks". *Social Networks* 2:1-18.
- Rapoport, A. 1979b. "Some problems relating to randomly constructed biased networks". *Perspectives on Social Network Research* 119-37.
- Sirken, M.G. and Levy, P.S. 1974. "Multiplicity estimation of proportions based on ratios of random variables". *Journal of the American Statistical Association* 69:68-73.
- Snedecor, G.W. and Cochran, W.G. 1972. *Statistical Methods*. Ames, Iowa: The Iowa State University Press.
- Spreen, M. 1992. "Rare populations, hidden populations, and link-tracing designs: what and why?" *Bulletin de Méthodologie Sociologique* 36:34-58.
- TenHouten, W.D., Stern, J and TenHouten, D. 1971. "Political leadership in poor communities: applications of two sampling methodologies". *Urban Affairs Annual Review* 5:215-54.
- Warwick, D.P. and Lininger, C.A. 1975. *The Sample Survey: Theory and Practice*. New York: McGraw-Hill.
- Watters, J.K., and Biernacki, P. 1989. "Targeted sampling: Options for the study of hidden populations". *Social Problems* 36(4):416-30.
- van Meter, K.M. 1980. "Methodological and design issues: Techniques for assessing the representativeness of snowball samples". in Lambert EY (ed) *The Collection and Interpretation of Data from Hidden Populations*. NIDA Research Monographs. 31-43.