Network Sampling: Some First Steps
Author(s): Mark Granovetter
Source: *The American Journal of Sociology*, Vol. 81, No. 6 (May, 1976), pp. 1287-1303
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/2777005
Accessed: 16/01/2011 09:58

Network Sampling: Some First Steps[1]

Mark Granovetter
Harvard University

Social network research has been confined to small groups because large networks are intractable, and no systematic theory of network sampling exists. This paper describes a practical method for sampling average acquaintance volume (the average number of people known by each person) from large populations and derives confidence limits on the resulting estimates. It is shown that this average figure also yields an estimate of what has been called "network density." Applications of the procedure to community studies, hierarchical structures, and interorganizational networks are proposed. Problems in developing a general theory of network sampling are discussed.

Sociologists and anthropologists have discussed and studied communities since their disciplines began. As the communities studied have increased in size, the fact that not all community members have social relations with one another has become a matter of prominent theoretical focus. The metaphor most consistently chosen to represent this situation is that of the "social network"—a device for representing social structure which depicts persons as points and relations as connecting lines. (Good general discussions are found in Barnes 1969; Bott 1957; Mitchell 1969; White, Boorman, and Breiger 1976).

Most discussions of network ideas, however, have had practical application only to small groups. Inability to apply the ideas effectively to larger structures has stemmed in part from the lack of a theoretical framework in which to place the network metaphor and in part from the absence—and perceived difficulty—of methods applicable to and statistical understanding of large networks. In an earlier paper (1973) I suggested some theoretical leads for the application of network ideas to macrosociology; here I explore some statistical and methodological avenues which, when more fully developed, should help to bring the network perspective more squarely into the mainstream of sociological research.

It is clear why network methods have been confined to small groups: existing methods are extremely sensitive, in their practicality, to group

[1] An early version of this paper was delivered at the Mont Chateau conference on the anthropological study of social networks, sponsored by the Mathematical Social Science Board, Morgantown, West Virginia, May 16–19, 1974. Discussions begun at that conference led to crucial improvements. In particular, the progress reported here would not have been possible without the collaboration and stimulation of Paul Holland; remarks by Samuel Leinhardt triggered important parts of the work. I am also indebted to Harrison White, Stanley Wasserman, and Ove Frank for valuable criticism.

size because they are population rather than sampling methods. In a group of size N, the number of potential (symmetric) ties is $[N \cdot (N - 1)/2]$ (i.e., proportional to N^2), so that any method meant to deal with the total population faces insuperable obstacles for groups larger than a few hundred. A group of 5,000, for instance—which we might think of as a small town—contains over 12 million potential lines in its network. Yet most Americans live in much larger aggregates, which analysts nevertheless, in community studies, persist in thinking relevant as social units. Implicitly, these studies often make arguments about the community's total network, but they rarely do so explicitly, because no methods exist for investigating such an object. Implicitly, again, what all such studies do, and must do, is to sample from that network. But because the procedure is not explicit and no statistical theory guides it, we are left guessing about the representativeness of the patterns of social relations found. This uncertainty is particularly noticeable when the "sampling" procedure is one of participant observation, but representativeness is problematic even if the procedure consists of asking a random sample of the community some sociometric questions. Just as an enormous advance in sociological work ensued when the general theory of random sampling was developed and applied to sociological problems, so the full development of network ideas in macrosociological perspective must await a comparable theory of network sampling. At present, only a few analysts have attacked this problem (Goodman 1961; Bloemena 1964; Capobianco 1970; Frank 1971), and only Frank attempts a comprehensive statistical treatment.

In this paper, I show that for one simple but important property of social networks, "density," a straightforward and practical method can provide acceptable sampling estimates even for very large populations. I then suggest applications of the method and discuss the more general problems of sampling from networks.[2]

Network density is the ratio of the number of ties actually observed to the number theoretically possible. In small groups, density is usually treated as a measure of group "cohesion" (Festinger, Schacter, and Back 1950, chap. 5) and as a partial indication of the extent to which a group

[2] I should stress that the basic results here are made possible by the pathbreaking work of Ove Frank (1971), professor of statistics, University of Lund, Sweden. Even more important than his actual results is Frank's demonstration that network sampling problems can be attacked with relatively standard methods of statistical inference (e.g., the use of indexing variables), although their application requires a good deal of imagination. Another important breakthrough which deserves to be followed up is Goodman's paper on "snowball sampling" (1961). Snowball sampling is not appropriate in the present paper, however, because its practicality is limited to cases where respondents make a fairly small number of sociometric choices. I mean to develop methods relevant to respondents' *entire* friendship networks, including the many people they know whom it would not occur to them to choose in a limited-choice situation. On the general significance of "weak ties," see my 1973 paper in this *Journal*.

is "primary" or "closed" (see Homans 1974; Bott 1957). In communities or larger settings, density has been used to indicate levels of "modernization" (see Mayer 1961; Tilly 1969). A good general discussion of density can be found in Barnes (1969).

DENSITY AND THE "HOW-MANY-PEOPLE" PROBLEM

Before discussing the sampling method, I want to make a detour to show that finding the density in networks is actually, given certain limitations, equivalent to answering the question "How many people do people know?" That question, though one might suppose its answer to be a fundamental social fact, has actually been studied very little, and no systematic information exists for representative populations.

Initially, the problem may seem straightforward: If we want to know how many people someone knows, why not ask him? To be quite frank, no *direct* evidence shows that this procedure would give poor results. Indirect evidence and everyday experience, however, suggest that most individuals could give only a very rudimentary estimate. The obvious method would be to ask a respondent to write out a list of all the people he knows and count the names. But such a substantial proportion of one's contacts are seen infrequently that they would come to mind only with some difficulty, particularly during the limited time one could allow or expect for the schedule to be filled out.

Gurevitch (1961) used an ingenious extension of this method, which insured far greater accuracy. He asked each respondent to keep a daily diary over a period of 100 days, listing, each day, all the persons he came into contact with who also knew *him* by this criterion (p. 1, n). This time period was chosen because "the net increment during the tail end of this period was small enough to justify termination of the procedure" (p 43).

The method, a variant of time-budget techniques, gives excellent results but has serious drawbacks. The most serious and obvious is that such sustained commitment of respondents can probably only be secured on a paid basis. Gurevitch's sample consisted of 15 individuals who responded to a notice offering pay for this activity and three unpaid volunteers (who presumably knew him). The size and character of his sample make generalization from it impossible, and the cost of the method makes reasonable samples impractical to draw. Another difficulty is that the method misses contacts seen less frequently than every 100 days. In a place where someone has lived for some years, there may be a substantial number of these.

A very different method becomes available if we shift the focus away from the individuals to the communities in which they live. In this more macroscopic perspective, we may view acquaintance volume as a char-

acteristic not simply of individuals but of the entire community. In fact, if we are willing to sacrifice individual detail and investigate the *average* number of people known to people within a bounded community (in which case we also miss, of course, contacts outside the boundary we set), the question does reduce to that of network density, as follows: In a group of size N, where N_t ties are observed, the density measure, D, is $N_t/[N(N-1)/2]$, where all ties are assumed symmetric. But since each of the N_t observed ties represents *two* cases of someone knowing someone else, the total number of contacts in the group must be $2N_t$, and the average number per person $2N_t/N$. Call this quantity V, for average acquaintance volume. Simple algebra now shows that $V = (N-1)D$. Hence, any method which finds density also finds average acquaintance volume. In what follows, I will often describe networks in terms of their average acquaintance volume instead of their density, given the greater intuitive appeal of the former. When I discuss applications, it will be clear that the equivalence also has substantive importance.

THE SAMPLING METHOD

Given a population of size N, the method proposed is to take a number of random samples from that population, each of size n (with replacement), and *within* each such sample ask each respondent some sociometric question about *each other respondent*. Which sociometric question is asked depends on the purpose of the particular investigation. If the main focus, for example, were the "how-many-people" question, it would be sufficient to ask whether the respondent knew each of the other $n - 1$ respondents by name. In this method, frequency of contact is irrelevant—people seen only every few years, or less often, have the same chance of being named as those seen every day. One could assure himself that results would be accurate for a given sample by providing as a stimulus not only the names of the $n - 1$ others, but also other relevant information such as address or occupation; even photographs could be used to be sure that acquaintances whose faces were better remembered than their names would be recognized.

In the language of graph theory, each sample, once lines are drawn among respondents corresponding to their sociometric responses, is a "random subgraph" from the population.[3] By averaging the densities found in the various samples taken, one arrives at an estimate of the density in the population network. (In the Appendix I give a proof that this estimate is unbiased.)

Two sampling parameters need to be set: the number of samples taken

[3] The use of random subgraphs was first suggested to me by Samuel Leinhardt.

and the size of each sample. One can imagine taking a large number of small samples (e.g., thousands of random pairs) or a small number of large samples (e.g., a few samples of several hundred or more). Some previous work on network sampling has focused on the idea that the relevant sampling unit ought to be the "tie"—that one should sample not from the N individuals but from the $N(N-1)/2$ possible lines in the network, to see in which cases lines are actually observed (Capobianco 1970; Niemeijer 1973; Tapiero, Capobianco, and Lewin 1975). This has a certain intuitive appeal but needs to be seen as a special case of sampling random subgraphs, with each subgraph containing two points. In effect, it is one example of the "large number of small samples" strategy referred to above. In this perspective, the calculations below make it clear that the "small number of large samples" is almost invariably a more efficient strategy.

The crux of the statistical problem then, is to determine what combinations of the two sampling parameters will insure a good estimate of acquaintance volume.

The first step must be a formula for the variance of our density estimate. This can be derived easily from a crucial result obtained by Frank (1971, p. 92), who shows that when exactly *one* subgraph of size n is sampled from a population of size N, and T denotes a random variable, the number of ties observed in this subgraph,

$$\begin{aligned}\text{Var}\,(T) = {} & (N-n)n(n-1)(n-2)s^2(a)/(N-1)(N-2)(N-3) \\ & + (N-n)(N-n-1)n(n-1)s^2(C)/2(N-2)(N-3), \quad (1)\end{aligned}$$

where $s^2(a)$ is the variance of the true vector of "outdegrees," that is, individual acquaintance volumes, and $s^2(C)$ is the variance of the true (population) sociomatrix.

By definition, our density estimate, to be called $\hat{D}$, is equal to $T/\binom{n}{2}$. Thus, $\text{Var}\,(\hat{D}) = \text{Var}\,(T)(4/[n^2(n-1)^2])$. Now, suppose more than one subgraph of size n is sampled—namely, w such samples—and we average all their density estimates to get an overall estimate $\hat{D}_{av}$. Neglecting the covariance among the pooled estimates,[4] we then have

$$\begin{aligned}\text{Var}\,(\hat{D}_{av}) = {} & (1/w)\left[\frac{2(N-n)}{(N-2)(N-3)n(n-1)}\right] \\ & \cdot \left[\frac{2(n-2)}{(N-1)}s^2(a) + (N-n-1)s^2(C)\right]. \quad (2)\end{aligned}$$

Equation (2) shows, as one might suspect, that the variance of the

[4] It is safe to neglect the covariances so long as $n \ll N$, or where n is moderate, so long as w is small. As will be shown below, these conditions apply in nearly all conceivable cases.

density estimate depends on the detailed structure of the actual population network, as given by $s^2(C)$ and $s^2(a)$. It is easily shown (see Frank 1971, pp. 70–72) that $s^2(C) = D(1 - D)$ and that

$$s^2(a) = (N - 1)s^2(C) - \frac{N - 1}{N} \cdot \sum\sum_{i \neq j} \{C_{ij} - [a_i/(N - 1)]\}^2,$$

where a_i is the number of ties involving person i, that is, the sum of row i in the sociomatrix.

Since the parameter $s^2(C)$ is completely determined by the density, the first step in arriving at a density estimate is to guess at the *true* density for a given population; to fix $s^2(a)$ requires more complex assumptions about how that density is distributed. First of all, since it is a variance, it must $= 0$ if $a_1 = a_2 = \ldots a_n$—that is, if every individual knows the same number of other people. The number, a_i, would then be exactly the average acquaintance volume and would hence minimize Var $(\hat{D}_{av})$ and the required sample size, for fixed density and confidence limits. (This is also clear from inspection of eq. [2].)

Correspondingly, $s^2(a)$ is maximized, for a given density, when all the acquaintanceship is concentrated in the smallest possible number of people, and everyone else knows no one. In order to say what this "smallest possible number" is, we must first specify the maximum number of people anyone might know. The larger this number, the higher the variance of $s^2(a)$, since the number who know any people can then be quite small compared to population size. A conservative procedure would be to set this number high and imagine that, in our maximum baseline population, if someone knows *any* people, he knows the maximum number. I will set this figure at 2,000. (In the Gurevitch 100-day diary study [1961] the largest number of acquaintances reported by any respondent was 658.) In a population of 100,000, then, with average acquaintance volume of 100, the maximal case would exist if 95,000 people knew no one, and each of the 5,000 others knew 2,000 from among one another. If $(N \cdot V)/2{,}000 < 2{,}000$, some modification is needed in the definition of "maximal." For instance, where $N = 10{,}000$ and $V = 100$, the present definition suggests a maximal graph as one with 9,500 people who know no one and 500 who each know 2,000 others. But this is internally inconsistent, since "knowing" is symmetric. In such cases, the logical procedure is to take $(N \cdot V)^{1/2}$ as the number of people who know anyone at all, and assume that each of them knows each other person in the group, and no one else knows anyone. Utilization of this procedure yields, if $N = 10{,}000$ and $V = 100$, a graph in which a set of 1,000 people know each other and 9,000 know no one.

Given fixed D (or V) this condition maximizes the variance of $\hat{D}$. The

overall result is plausible: Variance (and hence required sample size) is minimized in the fully homogeneous network and maximized in the maximally cliqued one.

It seems clear to me that the latter situation is much further from reality than the former. Even people who moved into a community yesterday are not isolates, and the great majority are likely to know some moderate number of people. Although I have no direct evidence, I would be surprised if more than, say, 20% of a population knew *many* more than the average number of people. Let me, then, arbitrarily define a "typical" population as one in which this is so. Specifically, let f_k be the proportion of the population whose acquaintance volume is equal to k times the average. For volume of 100 or 500, suppose $f_{0.5} = 0.4$, $f_1 = 0.4$, $f_{1.5} = 0.1$, $f_2 = 0.075$, $f_4 = 0.025$. For volume of 1,000, f_4 should be zero, in keeping with our stricture that no one knows more than 2,000 others; for that case, let $f_{0.5} = 0.3$, $f_1 = 0.5$, $f_{1.5} = 0.1$, $f_2 = 0.1$. Using this somewhat arbitrary notion of a typical population distribution of acquaintance volume, and the previously defined minima and maxima, we can arrive at some idea of what sample size will be needed to get a decent density estimate.

Let our idea of "decent" correspond to permitting a 20% error. While statisticians will blanch at this definition of decent, I argue that such a range will serve most purposes well enough. When two communities are compared whose true acquaintance volumes fall within the range of one another's 20% error limits, it seems doubtful that the true difference would be of much substantive significance anyway. In any case, formula (3) below can be modified to permit more narrow error limits.

Suppose the density estimates from a subgraph of size n are approximately normally distributed about the true density. Then 95% of the estimates will fall within 1.96 SD of the true density. That is, our 95% confidence interval of 20% tolerable error requires that

$$0.2D \geqslant 1.96\left\{\frac{1}{w}\left[\frac{2(N-n)}{(N-2)(N-3)n(n-1)}\right]\right.$$
$$\left.\left[\frac{2(n-2)}{(N-1)}s^2(a) + (N-n-1)D(1-D)\right]\right\}^{1/2} \tag{3}$$

For specified N, n, D, and $s^2(a)$, we can solve the inequality for w, giving us the minimum number of samples of size n needed to come within the specified limits. (Changes in degree of error tolerated or in level of confidence required can be introduced by substituting the desired values for 0.2 or 1.96.)

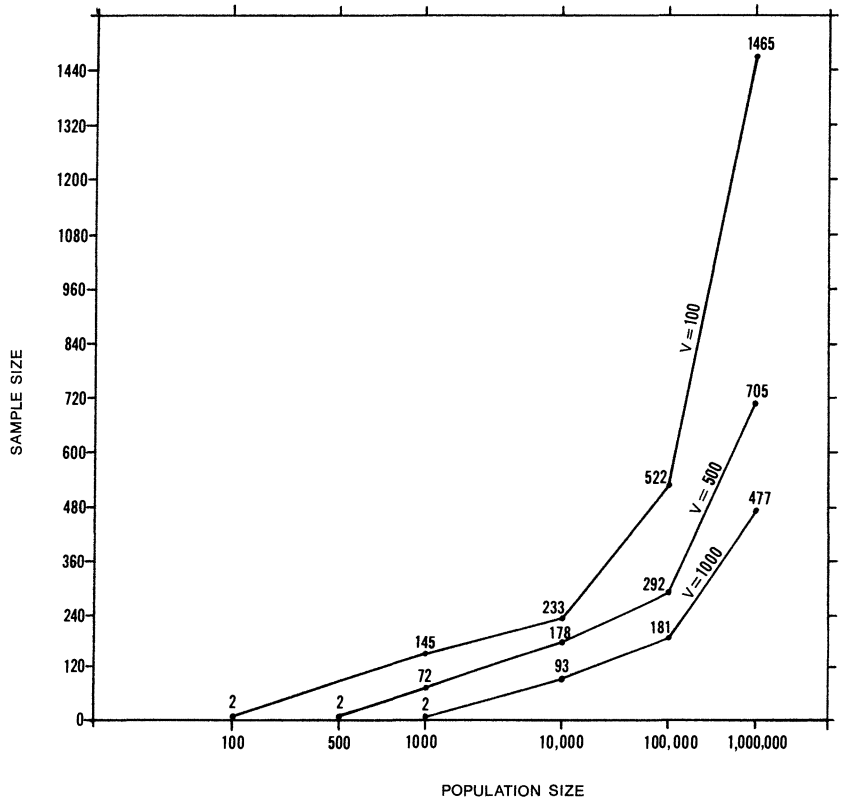

FIG. 1.—Sample size required for 95% confidence limits on network acquaintance volume, 20% error, for population sizes from 100 to 1,000,000, and true volume (V) of 100, 500, and 1,000.

Figure 1 gives the values of n for which $w = 1$ (i.e., only *one* sample of size n is needed) under various assumptions about acquaintance volume (density) and population size (N). In this graph, the population is assumed to show the "typical" distributions of acquaintance volumes outlined above, except that where volume is 500 and $N = 1{,}000$, the 0.3, 0.5, 0.1, 0.1 pattern is used. With the exception of new towns in very early stages, it seems unlikely that any communities would fall outside the acquaintance volume range of 100–1,000 used in figure 1.

The numbers here suggest that even for rather large populations the sample size required for the estimate is modest. Some empirical experience will be needed before we can say what the largest sample is in which we can, as a practical matter, expect each individual to answer sociometric questions about each other. I would think that this could be done easily

for samples of 500.[5] Figure 1 shows that only for the community of 1 million would a larger sample be needed. In such cases, if indeed 500 were the practical cutoff for this method, the difficulty could be met by taking *multiple* samples of size 500. With a population of 1 million and true average acquaintance volume of 500, two such samples would suffice; where true volume drops to 100, eight would be needed.

A nice irony of the method outlined is that it is of only marginal value for *small* populations, unlike all other network techniques. For a population of size 50, for example, one would need a random sample of at least 25 for a decent estimate. While this might occasionally be of some value, the real power of the method appears only for larger networks, ones at least in the hundreds.

Table 1 shows these computations not only for the "typical" case, but

TABLE 1

SAMPLE SIZES NEEDED TO MEET 20% ERROR, 95% CONFIDENCE LIMITS, FOR $w \doteq 1$

Population Size (N)	Average Acquaintance Volume	Value of n for Which $w \doteq 1$		
		Minimum	"Typical"	Maximum
1,000,000	100	1,382 (7.69)	1,465 (8.01)	7,460 (22.24)
1,000,000	500	619 (1.54)	705 (1.86)	1,415 (3.83)
1,000,000	1,000	438	477	668 (1.53)
100,000	100	436	522 (1.08)	6,800 (15.25)
100,000	500	195	292	1,165 (2.44)
100,000	1,000	138	181	425
10,000	100	136	233	2,570 (6.62)
10,000	500	60	178	1,030 (2.20)
10,000	1,000	41	93	370
1,000	100	40	145	454
1,000	500	14	72	137
1,000	1,000	...	...	...

NOTE.—Where a single sample larger than 500 would be needed, the number of samples required of size 500 is given in parentheses.

also for the minimum and maximum variance conditions specified above. Where $n > 500$, the number of samples of size 500 needed to meet the criterion is indicated in parentheses. As one might expect, required sample size goes down as population size goes down and as acquaintance volume goes up. The figures for "typical" populations are much closer to the minimum than the maximum. An analyst with a population suspected to be highly cliqued, however, can take comfort from the fact that even the sample size required for the maximum degree of cliquing is within practical bounds, for most cases.

[5] In some situations this may be doubled (see the section below on one-way questioning).

One final issue, mentioned above, is the trade-off between sample size and number of samples. At the beginning of my work, I was intrigued by the idea that a large number of samples of pairs or triads would provide the key to network sampling. Table 2 shows why this is not so. For pair

TABLE 2

TWO ILLUSTRATIONS OF TRADE-OFF BETWEEN SAMPLE SIZE (n) AND NUMBER OF SAMPLES TAKEN (w)

n	w	Maximum Number in Samples	Maximum Number of Questions Asked
	$N = 100{,}000$; $V = 500$; 20% Error, "Typical" Variance of V		
2	19,112	38,224	38,224
3	6,398	19,194	38,388
10	439	4,390	39,510
50	19	950	46,550
100	5	500	49,500
200	2	400	79,600
292	1	292	84,972
	$N = 1{,}000{,}000$; $V = 500$; 20% Error, "Typical" Variance of V		
2	191,984	383,968	383,968
3	64,022	192,066	384,132
10	4,281	42,810	385,290
50	160	8,000	392,000
100	40	4,000	396,000
500	2	1,000	499,000
705	1	705	496,320

sampling, a prohibitive number of people would have to be contacted. The total number of people one would have in all samples goes down steadily as the size of the samples increases. (Since sampling is with replacement, 19,112 samples of two from a population of 100,000 would involve fewer than 38,224 persons, but the expected number is not far from the maximum, so that this offers little help.)[6] The maximum number of questions asked of all respondents is arrived at by multiplying the maximum number of respondents by $(n - 1)$. This number increases with n, but the decline in the total number of respondents sampled is far more rapid. Since so much of an interviewer's time consists of finding respondents, the strategy of taking many small samples would therefore be much less practical than that of taking a lesser number of larger ones.[7]

[6] This calculation results from formulas developed by Paul Holland and Stanley Wasserman.

[7] I am indebted to James Beniger for pointing out errors in an earlier draft of this paragraph and of table 2.

More important, at a theoretical level, the larger the sample, the more properties of the population network can be estimated. In pair sampling, for instance, nothing *but* density could be estimated, since the subgraphs sampled have no structural properties other than the 1–0, yes-no question about the one potential tie. In a single, large random subgraph, by contrast, nearly all relevant properties appear. On the question of estimating the *full distribution* of acquaintance volumes, for example, Frank shows that the problem is simplified (although it is still not solved) if the sample size is "so large that all the frequencies that it is intended to estimate will have a chance of being represented in the sample graph" (1971, p. 114).

For both theoretical and practical reasons, then, I believe that network sampling must go in the direction of a few large samples rather than many small ones.

ONE-WAY QUESTIONING[8]

The above discussion is conservative, in that it assumes a method in which each possible tie is asked about from both ends—that is, each of the n sampled individuals is asked about each *other* one, so that i is asked about j as well as j about i. This means that we are asking $[n(n-1)]$ questions to find out about $[n(n-1)]/2$ ties. In most cases we will want to do this, either because we are interested in the symmetry or asymmetry of the sociometric choices or because we have more confidence in the validity of the information when both ends of a tie affirm its existence. If we are not interested in symmetry, however, and are willing to assume that a tie exists whenever one participant says it does, the sampling work can be cut in half. It is easy to arrange for each of the n sampled individuals to be asked, not about *each* of the $n-1$ others, but about roughly $(n-1)/2$ others; information is still obtained about each possible tie. Where, for example, 500 people are sampled, table 3 shows one simple scheme for using this method.

Such a procedure might be especially reasonable in obtaining acquaintance volume estimates for large cities. Whereas in a sample of 1,000 it might not be practical to ask each member questions about 999 others, questions about 500 others could probably be managed.

SOME APPLICATIONS

The applications that follow are illustrative only; they are far from exhaustive.

[8] This section was suggested by comments of Harry Collins.

TABLE 3

SOCIOMETRIC QUESTIONS FOR 500 RESPONDENTS UNDER ONE-WAY QUESTIONING

Respondent No.	Is Asked about Respondents No.
1	2–251
2	3–252
3	4–253
.	.
.	.
.	.
250	251–500
251	252–500
252	253–500, 1
253	254–500, 1–2
.	.
.	.
.	.
498	499–500, 1–247
499	500, 1–248
500	1–249

NOTE.—Since $n - 1$ is an odd number, each respondent cannot be asked *exactly* $(n - 1)/2$ questions, half are asked $(n/2)$ and half $(n - 2)/2$ since $n/2[(n/2) + (n - 2)/2] = [n(n - 1)]/2$.

1. *Sense of community.*—A central focus of community studies has long been the question of what determines whether residents feel a "sense of community" where they live (e.g., Nisbet 1953; Stein 1960). While this concept is ambiguous and involves many intangibles, a central part of all analysts' notion of "sense of community" is the existence of a relatively dense network of social ties over the specified area. Ethnographic studies make it clear that a crucial part of the sense of "belonging" in a place is the constant encounter with familiar, friendly faces in the course of everyday life. Young and Willmott (1962) found, after analyzing the close-knit Bethnal Green area of East London, that residents who had moved to "Greenleigh," a government-sponsored suburban new town, were extremely unhappy. Their number of social contacts fell precipitously, and they consequently experienced the general atmosphere as cold and "unfriendly." Young and Willmott explained the contrast by the differing ecology of the two areas—the city being more densely packed—and intermixing residential and commercial functions. This arrangement facilitates sociability, whereas the suburban one makes it difficult and artificial (cf. Jacobs 1961, chaps. 1–12).

But Willmott later studied Dagenham, a community very similar to Greenleigh except that it had been settled in the 1920s. He was surprised, after the study of Greenleigh, to see the extent to which Dagenham's residents had reestablished a sense of community more typical of urban

East London; most people felt the atmosphere to be friendly, and considerable visiting was reported (Willmott 1963, pp. 58–64). His conclusion is that earlier studies, such as that of Greenleigh, "have put altogether too little emphasis on sheer length of residence" (p. 111). The implication is that the six or seven years of Greenleigh's existence, although it seemed a substantial time to the authors, was actually brief in relation to the time required to build up dense social networks from scratch. But we are left guessing about how "sheer length of residence" has its effect. Other studies of new towns, or suburbs where a substantial influx arrives at one time, report far more cheerful social results than those found in Greenleigh (see Berger 1960; Gans 1967). What explains these differences? What is the process by which a sense of community develops, and what determines the requisite time?

The sampling method outlined in this paper offers a useful research tool for these issues. A time series of average acquaintance volume in a new town, beginning early and continuing for a number of years, would, in conjunction with ethnographic work, yield important insights. The statistical comparisons made possible by the sampling method would allow interesting questions to be posed which could not be answered by a single case study. For example: how important is *initial* acquaintance volume in determining the rate of community development? It may be that new towns in which there exist, at the outset, a substantial number of interpersonal ties are successful far more quickly than others. The initial ties may serve a pump-priming function—satisfying people's interim need for sociability and smoothing the way for new ties—since existing friends are a crucial source of new ones. Initial ties may come about in a variety of ways; for example, people may be more likely to move to a new town if they already know someone who lives there (see MacDonald and MacDonald [1964] on "chain migration"), or economic factors, such as plant relocations, may prompt a considerable migration from one place to another (see Berger 1960). Substantive questions such as that of the source of initial ties can easily be incorporated into a survey using the sampling method proposed here. When a respondent is given the stimulus of another name from the sample, and the response indicates some relationship, a variety of questions about the relationship—its origin, intensity, duration—can be asked, depending on the nature of the inquiry.

In this case, a comparative study could relate initial network density to the subsequent rate of increase. Different patterns of increase or decrease might be found to correlate with changes in a community's political or economic structure. In this connection, measures of average acquaintance volume for any communities, old or new, might be of great potential value as social indicators. Use of them would extend the recent wave of

interest in such indicators to interactional measures. Most indicators currently in use consist, by contrast, of individual characteristics (happiness, income, illness) aggregated over large numbers.

2. *Hierarchy.*—Recent theoretical work on acquaintance networks makes a good argument for the idea that unreciprocated sociometric choices indicate a status differential, the chooser occupying a lower status (Davis and Leinhardt 1972; Holland and Leinhardt 1971; Bernard 1974). Investigations thus far have been limited to groups of a couple of hundred, given data-processing difficulties. The sampling device discussed here offers an entrée into this question for larger populations. My statistical analysis is unchanged if the tie in question is unreciprocated instead of symmetric. Estimates for the density of both types are yielded by a single sample as follows: From the sample sociomatrix, construct two submatrices, one containing only the symmetric, the other only the asymmetric ties. Use each submatrix to arrive at a density estimate for its type of tie.[9] The ratio of the asymmetric density to the overall density might be an interesting measure of the degree of dyadic hierarchy in the group and a useful parameter for comparison of groups. For a single group, a time series of the measure would offer some insight into the evolution, stability, or disintegration of hierarchy. (A fuller treatment of hierarchy would require methods for sampling the average properties not only of dyads, as in this paper, but also of triads, which I do not deal with here.)

3. *Interorganizational networks.*—Networks whose nodes are organizations have recently generated increasing interest. Often, however, a network comprises too many organizations for any study to be feasible. In a given industry, say electronics, it might be of interest to know the extent to which firms interchange personnel. (In Granovetter 1974 I discuss the substantive significance of this question.) The degree of such interchange might be a nice parameter in comparison of different industries. All that would be needed to apply the findings of the present paper is adoption of some minimum level of personnel flow between a pair of companies as constituting a "tie" between them. A square sociometric matrix for the set of firms could then be filled in with the usual 1–0 entries. Insofar as flow were asymmetric, it might be possible to infer hierarchy, making this application a special case of the one suggested above.

DISCUSSION

In this paper I have argued that sociologists interested in the idea of social networks must attend to the development of a related theory of sam-

[9] But notice that dividing ties found into *any* set of categories requires a larger sample size, the increase to be determined by the category of tie with lowest (assumed) true density. Otherwise, estimates for the various categories will carry errors larger than would be tolerable when all ties are considered identical in quality.

pling, if they are to incorporate macrolevel concerns into the framework. Building on results in Frank (1971), I have shown that a relatively simple sampling procedure will yield good estimates of network density, or average acquaintance volume.

This is only a small step in network-sampling theory, however, since density is a crude, global measure of interaction structures. It is, in fact, the global aspect of density which makes it comparatively simple to estimate by a method which is both intuitively appealing and practicable. More detailed measures of network structure, especially measures of local variations and inhomogeneities in the network, are necessarily more difficult to estimate from small samples, since such samples, tapping only average properties of the entire graph, give poor representation to rare events. Moreover, when an estimating procedure *is* found which yields an unbiased estimate of such parameters, great computational and conceptual difficulty ensues in finding the variance of such estimates, without which no confidence limits can be established. Frank, for example (1971, pp. 109–15), develops formulas for estimating from the sample graph the *exact* distribution of acquaintances in the population. The unbiased estimator, however, no longer has the intuitive flavor that our density estimate has but depends instead on complex sets of equations. Moreover, the variance of the estimate "is of no practical use, because it involves unknown population parameters that seem hard to estimate" (Frank 1971, p. 112).

Nevertheless, the realization that some good results can be achieved without recourse to wholly new methods should be an incentive to those with good mathematical and statistical skills to push ahead in an area which promises considerable rewards for the time invested.

APPENDIX

In this Appendix I prove that density estimates from random subgraphs provide an unbiased estimate of true population network density.

To prove that the estimate is unbiased, consider w sets of n individuals, each set a random sample from a population of size N, with replacement.

Let T_i be a random variable, the number of ties actually observed in sample i, where $i = (1, 2 \ldots w)$. Let $P_k =$ the probability (based on empirical relative frequencies) of observing k ties in such a set of size n, where $k = 0, 1, 2, \ldots [n(n-1)]/2$, $\Sigma P_k = 1$.
Then:

$$E(T_i) = \sum_{k=1}^{n(n-1)/2} kP_k. \qquad \text{(A1)}$$

In the population graph, there are $P_k \binom{N}{n}$ sets of n people with k ties in

the set. Since any given tie appears in $\binom{N-2}{n-2}$ *different* sets of size n, the total number of ties in the population is:

$$\sum_{k=1}^{n(n-1)/2} kP_k \binom{N}{n} \Big/ \binom{N-2}{n-2} \tag{A2}$$

This quantity divided by $\binom{N}{2}$ gives an expression for network density (D). Then, substituting from (A1), we have:

$$D = E(T_i) \Big/ \binom{n}{2} = E\left[T_i \Big/ \binom{n}{2} \right] \tag{A3}$$

Thus, $T_i / \binom{n}{2}$, which is the density observed in sample i, gives an unbiased estimate of the population density. Call this estimate $\hat{D}_i$. Now if w separate samples are taken, we have $E(\hat{D}_1) = E(\hat{D}_2) = \ldots E(\hat{D}_w) = D$. Since $E(\hat{D}_1 + \hat{D}_2 + \ldots \hat{D}_w) = E(\hat{D}_1) + E(\hat{D}_2) + \ldots E(\hat{D}_w)$, it follows that

$$E[(\hat{D}_1 + \hat{D}_2 + \ldots \hat{D}_w)/w] = wD/w = D, \tag{A4}$$

that is, the average density estimate from the w samples also unbiasedly estimates population density.

REFERENCES

Barnes, J. A. 1969. "Networks and Political Process." Pp. 51–76 in *Social Networks in Urban Situations,* edited by J. C. Mitchell. Manchester: Manchester University Press.

Berger, Bennett. 1960. *Working-Class Suburb.* Berkeley: University of California Press.

Bernard, Paul. 1974. *Association and Hierarchy.* Ph.D. dissertation, Harvard University.

Bloemena, A. R. 1964. *Sampling from a Graph.* Amsterdam: Mathematics Centrum.

Bott, Elizabeth. 1957. *Family and Social Network.* London: Tavistock.

Capobianco, M. F. 1970. "Statistical Inference in Finite Populations Having Structure." *Transactions of the New York Academy of Science,* 11th ser. 32:401–13.

Davis, James A., and S. Leinhardt. 1972. "The Structure of Positive Interpersonal Relations in Small Groups." Pp. 218–51 in *Sociological Theories in Progress.* Vol. 2, edited by J. Berger, M. Zelditch, and B. Anderson. Boston: Houghton-Mifflin.

Festinger, L., S. Schachter, and K. Back. 1950. *Social Pressures in Informal Groups.* New York: Harper.

Frank, Ove. 1971. *Statistical Inference in Graphs.* Stockholm: Försvarets Forskningan-stalt.

Gans, Herbert. 1967. *The Levittowners.* New York: Knopf.

Goodman, Leo. 1961. "Snowball Sampling." *Annals of Mathematical Statistics* 32 (1):148–70.

Granovetter, Mark. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78 (May): 1360–80.

———. 1974. *Getting a Job: A Study of Contacts and Careers.* Cambridge, Mass.: Harvard University Press.

Gurevitch, Michael. 1961. "The Social Structure of Acquaintanceship Networks." Ph.D. dissertation, Massachusetts Institute of Technology.
Holland, Paul, and S. Leinhardt. 1971. "Transitivity in Structural Models of Small Groups." *Comparative Group Studies* 2 (May): 107–24.
Homans, George. 1974. *Social Behavior*. New York: Harcourt, Brace, Jovanovich.
Jacobs, Jane. 1961. *The Death and Life of Great American Cities*. New York: Random House.
MacDonald, John S., and Leatrice MacDonald. 1964. "Chain Migration, Ethnic Neighborhood Formation and Social Networks." *Milbank Memorial Fund Quarterly* 42 (January): 82–97.
Mayer, Phillip. 1961. *Townsmen or Tribesmen?* Capetown: Oxford.
Mitchell, J. C. 1969. "The Concept and Use of Social Networks." Pp. 1–50 in *Social Networks in Urban Situations,* edited by J. C. Mitchell. Manchester: Manchester University Press.
Niemeijer, Rudo. 1973. "Some Applications of the Notion of Density." Pp. 45–64 in *Network Analysis: Studies in Human Interaction,* edited by J. Boissevain and J. C. Mitchell. The Hague: Mouton.
Nisbet, R. A. 1953. *The Quest for Community*. New York: Oxford University Press.
Stein, M. 1960. *The Eclipse of Community*. Princeton, N.J.: Princeton University Press.
Tapiero, C., M. Capobianco, and A. Lewin. 1975. "Structural Inference in Organizations." *Journal of Mathematical Sociology* 4(1): 121–30.
Tilly, Charles. 1969. "Community: City: Urbanization." Mimeographed. Ann Arbor: University of Michigan.
White, Harrison, S. A. Boorman, and Ronald Breiger. 1976. "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions." *American Journal of Sociology* 81 (January): 730–80.
Willmott, P. 1963. *The Evolution of a Community*. London: Routledge.
Young, M., and P. Willmott. 1962. *Family and Kinship in East London*. Baltimore: Penguin.