Social Networks, 1 (1978/79)5–51

Contacts and Influence

Ithiel de Sola Pool
*Massachusetts Institute of Technology**

Manfred Kochen
*University of Michigan***

This essay raises more questions than it answers. In first draft, which we have only moderately revised, it was written about two decades ago and has been circulating in manuscript since then. (References to recent literature have, however, been added.) It was not published previously because we raised so many questions that we did not know how to answer; we hoped to eventually solve the problems and publish. The time has come to cut bait. With the publication of a new journal of human network studies, we offer our initial soundings and unsolved questions to the community of researchers which is now forming in this field. While a great deal of work has been done on some of these questions during the past 20 years, we do not feel that the basic problems have been adequately resolved.

1. Introduction

Let us start with familiar observations: the "small world" phenomenon, and the use of friends in high places to gain favors. It is almost too banal to cite one's favorite unlikely discovery of a shared acquaintance, which usually ends with the exclamation "My, it's a small world!". The senior author's favorite tale happened in a hospital in a small town in Illinois where he heard one patient, a telephone lineman, say to a Chinese patient in the next bed: "You know, I've only known one Chinese before in my life. He was —— from Shanghai." "Why that's my uncle," said his neighbor. The statistical chances of an Illinois lineman knowing a close relative of one of (then) 600 000 000 Chinese are minuscule; yet that sort of event happens.

The patient was, of course, not one out of 600 000 000 random Chinese, but one out of the few hundred thousand wealthy Chinese of Westernized families who lived in the port cities and moved abroad. Add the fact that the Chinese patient was an engineering student, and so his uncle may well have been an engineer too – perhaps a telecommunications engineer. Also there were perhaps some geographic lines of contact which drew the members of one family to a common area for travel and study. Far from surprising, the encounter seems almost natural. The chance meetings that we have are a clue to social structure, and their frequency an index of stratification.

*MIT, Center for International Studies, 30 Wadsworth Street, Cambridge, Mass. 02139, U.S.A.
**Mental Health Research Institute, The University of Michigan, Ann Arbor, Mich. 48104, U.S.A.

Less accidental than such inadvertent meetings are the planned contacts sought with those in high places. To get a job one finds a friend to put in a good word with his friend. To persuade a congressman one seeks a mutual friend to state the case. This influence is peddled for 5%. Cocktail parties and conventions institutionalize the search for contacts. This is indeed the very stuff of politics. Influence is in large part the ability to reach the crucial man through the right channels, and the more channels one has in reserve, the better. Prominent politicians count their acquaintances by the thousands. They run into people they know everywhere they go. The experience of casual contact and the practice of influence are not unrelated. A common theory of human contact nets might help clarify them both.

No such theory exists at present. Sociologists talk of social stratification; political scientists of influence. These quantitative concepts ought to lend themselves to a rigorous metric based upon the elementary social events of man-to-man contact. "Stratification" expresses the probability of two people in the same stratum meeting and the improbability of two people from different strata meeting. Political access may be expressed as the probability that there exists an easy chain of contacts leading to the power holder. Yet such measures of stratification and influence as functions of contacts do not exist.

What is it that we should like to know about human contact nets?

– For any *individual* we should like to know how many other people he knows, *i.e.* his acquaintance volume.

– For a *population* we want to know the distribution of acquaintance volumes, the mean and the range between the extremes.

– We want to know what kinds of people they are who have many contacts and whether those people are also the influentials.

– We want to know how the lines of contact are stratified; what is the structure of the network?

If we know the answers to these questions about individuals and about the whole population, we can pose questions about the implications for *paths* between pairs of individuals.

– How great is the probability that two persons chosen at random from the population will know each other?

– How great is the chance that they will have a friend in common?

– How great is the chance that the shortest chain between them requires two intermediaries; *i.e.*, a friend of a friend?

The mere existence of such a minimum chain does not mean, however, that people will become aware of it. The surprised exclamation "It's a small world" reflects the shock of discovery of a chain that existed all along.[1] So another question is:

---

[1] In the years since this essay was first written, Stanley Milgram and his collaborators (Milgram 1967; Travers and Milgram 1969; Korte and Milgram 1970) have done significant experiments on the difficulty or ease of finding contact chains. It often proves very difficult indeed.

– How far are people aware of the available lines of contact? A friend of a friend is useful only if one is aware of the connection. Also a channel is useful only if one knows how to use it. So the final question is, what sorts of people, and how many, try to exert influence on the persons with whom they are in contact: what sorts of persons and how many are opinion leaders, manipulators, politicists (de Grazia 1952; Boissevain 1974; Erickson and Kringas 1975)?

These questions may be answered at a highly general level for human behavior as a whole, and in more detail for particular societies. At the more general level there are probably some things we can say about acquaintanceship volume based on the nature of the human organism and psyche. The day has 24 hours and memory has its limits. There is a finite number of persons that any one brain can keep straight and with whom any one body can visit. More important, perhaps, there is a very finite number of persons with whom any one psyche can have much cathexis.

There are probably some fundamental psychological facts to be learned about the possible range of identifications and concerns of which a person is capable (Miller 1956).

These psychic and biological limits are broad, however. The distribution of acquaintanceship volumes can be quite variable between societies or social roles. The telephone makes a difference, for example. The contact pattern for an Indian villager *sans* radio, telephone, or road to his village is of a very different order from that of a Rotarian automobile dealer.

There is but little social science literature on the questions that we have just posed.[2] Even on the simplest question of the size of typical acquaintanceship volumes there are few data (Hammer, n.d.; Boissevain 1967). Some are found in anecdotal descriptions of political machines. In the old days there was many a precinct captain who claimed to know personally every inhabitant of his area. While sometimes a boastful exaggeration, there is no doubt that the precinct worker's success derived, among other things, from knowing 300 - 500 inhabitants of his neighborhood by their first names and family connections (Kurtzman 1935). At a more exalted level too, the art of knowing the right people is one of the great secrets of political success; James Farley claimed 10 000 contacts. Yet no past social science study has tested how many persons or what persons any politician knows. The estimates remain guesswork.

There exists a set of studies concerning acquaintanceship volume of delinquent girls in an institutional environment: J. L. Moreno and Helen Jennings asked girls in a reform school (with 467 girls in cottages of 23 or 24 apiece) to enumerate all other girls with whom they were acquainted (Jennings 1937). It was assumed they knew all the girls in their own cottage.

---

[2] In the last few years, however, the literature on human networks has started proliferating. There are articles dealing with information and help-seeking networks in such fields as mental health (Saunders and Reppucci 1977; Horowitz 1977; McKinlay 1973). There is also some anthropological literature on networks in different societies (Nutini and White 1977; Mitchell 1969; Jacobson 1970).

Computed that way, the median number of acquaintances was approximately 65. However, the range was tremendous. One girl apparently knew 175 of her fellow students, while a dozen (presumably with low I.Q.s) could list only four or fewer girls outside of their own cottage.

These figures have little relevance to normal political situations; but the study is valuable since it also tested the hypothesis that the extent of contact is related to influence. The girls were given sociometric tests to measure their influence. In each of two separate samples, a positive correlation (0.4 and 0.3) was found between contact range and influence.

One reason why better statistics do not exist on acquaintanceship volume is that they are hard to collect. People make fantastically poor estimates of the number of their own acquaintances (Killworth and Russell 1976). Before reading further, the reader should try to make an estimate for himself. Define an acquaintance as someone whom you would recognize and could address by name if you met him. Restrict the definition further to require that the acquaintance would also recognize you and know your name. (That excludes entertainment stars, public figures, *etc.*) With this criterion of acquaintance, how many people do you know?

The senior author tried this question on some 30 colleagues, assistants, secretaries and others around his office. The largest answer was 10 000; the smallest was 50. The median answer was 522. What is more, there seemed to be no relationship between the guesses and reality. Older or gregarious persons claimed no higher figures than young or relatively reclusive ones. Most of the answers were much too low. Except for the one guess of 10 000 and two of 2000 each, they were all probably low. We don't know that, of course, but whenever we have tried sitting down with a person and enumerating circles of acquaintances it has not taken long before he has raised his original estimate as more and more circles have come to mind: relatives, old school friends, merchants, job colleagues, colleagues on former jobs, vacation friends, club members, neighbors, *etc.* Most of us grossly underestimate the numer of people we know for they are tucked in the recesses of our minds, ready to be recalled when occasion demands.

Perhaps a notion of the order of magnitude of acquaintanceship volume can be approached by a *gedankenexperiment* with Jennings' data on the reform school. The inmates were young girls who had not seen much of the world; they had but modest I.Q.s and memories; they had come from limited backgrounds; and in the recent past they had been thoroughly closed off from the world. We know that the average one knew 65 inmates. Is it fair to assume that we may add at least 20 teachers, guards, and other staff members known on the average? Somewhere the girls had been in school before their internment. Perhaps each knew 40 students and 10 teachers from there. These girls were all delinquents. They were usually part of a delinquent gang or subculture. Perhaps an average of 30 young people were part of it. They had been arrested, so they knew some people from the world of lawyers, judges, policemen, and social workers. Perhaps there were 20 of them. We have not yet mentioned families and relatives; shall we say another 30? Then

there were neighbors in the place they had lived before, perhaps adding up to 35. We have already reached 250 acquaintances which an average girl might have, based solely on the typical life history of an inmate. We have not yet included friends made in club or church, nor merchants, nor accidental contacts. These might add another 50. Nor have we allowed for the girls who had moved around – who had been in more than one school or neighborhood or prison. Perhaps 400 acquaintances is not a bad guess of the average for these highly constricted, relatively inexperienced young girls. Should we not suspect that the average for a mature, white collar worker is at least double that?

Perhaps it is, but of course we don't know. All we have been doing so far is trying to guess orders of magnitude with somewhat more deliberation than was possible by the respondents to whom we popped the question "How many people do you know?". There has been no real research done to test such estimates.

It could be done by a technique analogous to that used for estimating a person's vocabulary. In any given time period during which we observe, a person uses only some of the words he knows and similarly has contact with only some of the people he knows. How can we estimate from this limited sample how many others are known to him? In each case (words and friends) we can do it by keeping track of the proportion of new ones which enter the record in each given time period. Suppose we count 100 running words. These may contain perhaps 60 different words, with some words repeated as many as 6 or 7 times, but most words appearing once. Adding a second 100 running words may add 30 new ones to the vocabulary. A third hundred may add 25 new ones, and so on down. If we extrapolate the curve we reach a point where new words appear only every few thousand running words, and if we extrapolate to infinity we have an estimate of the person's total vocabulary. In the same way, on the first day one may meet 30 people. On the second day one may meet another 30 but perhaps only 15 of them are new, the other 15 being repeaters. On the third day perhaps the non-repeaters may be down to 10, and so on. Again by extrapolating to infinity an estimate of the universe of acquaintances may be made.

Extrapolation to infinity requires strong assumptions about the number of very rarely seen acquaintances. If there are very many who are seen but once in a decade, then a much longer period of observation is required. If the number of people seen once in two decades is not significantly smaller than the number seen in a shorter period, then there are methodological difficulties in estimation.

Two further cautions are necessary. It turns out that the lumpiness in the schedules of our lives makes this technique unusable except over long periods. Perhaps we start on Thursday and go to work. Friday we go to work and see almost the same people. Saturday we go to the beach and have an entirely new set of contacts. Then Monday, perhaps, we are sent on a trip to another office. In short, the curves are highly irregular. Long and patient observation is called for.

Also note that at the end of a lengthy experiment (say after one year), it is necessary to check back over the early lists to determine who are forgotten and no longer acquaintances. Just as new persons enter the acquaintanceship sphere, old ones drop out of it. In one record, for example, a subject recorded 156 contacts in five successive days, with 117 different persons whom he then knew. Two years and ten months later, though still working in the same place, he could no longer recall or recognize 31 of these; *i.e.*, 86 (or 74%) were still acquaintances.

It is important to collect more such empirical information. Section 2 of this paper describes some empirical findings that we have obtained. But before we can decide what to collect we need to think through the logical model of how a human contact net works. We shall do that roughly and non-mathematically in this introduction. Section 3 of the paper deals with it more formally.

One question that quite properly is raised by readers is what do we mean by acquaintanceship, or friendship, or contact. For the mathematical model, the precise definition of "knowing" is quite irrelevant. What the mathematical model gives us is a set of points each of which is connected with some of the other points. As we look away from our model to the world for which it stands, we understand that each point somehow represents a person, and each connection an act of knowing. The model is indifferent to this, however. The points could stand for atoms, or neurons, or telephones, or nations, or corporations. The connections could consist of collisions, or electric charges, or letters written, or hearing about, or acquaintanceship, or friendship, or marriage. To use the model (and satisfy ourselves that it is appropriate) we shall have to pick definitions of person (*i.e.*, point) and knowing (*i.e.*, connectedness) related to the problem at hand. But we start with a model that is quite general. We do indeed impose some constraints on the points and on their connections. These constraints are the substance of our theory about the nature of human contacts.

One simplification we make in our model is to assume that the act of knowing is an all-or-none relationship. That is clearly not true and it is not assumed by Hammer (n.d.), Gurevich (1961) and Schulman (1976). There are in reality degrees of connectedness between persons. There are degrees of awareness which persons have of each other, and there are varied strengths of cathexis. But we cannot yet deal with these degrees. For the moment we want to say of any person, A, that he either does or does not know any given other person, B.

The criterion of human acquaintanceship might be that when A sees B he recognizes him, knows a name by which to address him, and would ordinarily feel it appropriate that he should greet him. That definition excludes, as we have noted, recognition of famous persons, since as strangers we do not feel free to greet them. It excludes also persons whom we see often but whose names we have never learned; *e.g.*, the policeman on the corner. It is, however, a useful operational definition for purposes of contact net studies, because without knowing a name it is hard to keep a record.

Alternatively, the criterion might be a relationship which creates a claim on assistance. In politics, that is often the important kind of knowing. One might well find that a better predictor of who got a job was a man's position in the network of connections defined by obligation than his position in the network of mere acquaintance connections.

For some anthropological studies the connection with which we are concerned might be kinship. As many societies operate, the most important fact in the dealings of two persons is whether they are kin or not. Kinship creates obligations and thus provides a protection of each by the other. Blood kinship is a matter of degree fading off imperceptibly; we are all ultimately related to everyone else. But society defines the limit of who are recognized as kin and who are unrelated. This varies from society to society, and sometimes is hard to establish. In many societies, Brazil and India for example, the first gambit of new acquaintances is to talk about relatives to see if a connection can be established. For such societies kinship is clearly an important criterion of connectedness.

Another criterion of connectedness, of considerable relevance in the United States, is the first-name index. This makes a sharp distinction between levels of knowing, just as does *Sie* and *du* in German or *vous* and *tu* in French.

Whatever definition of knowing we choose to use, our model proceeds by treating connectedness as an all-or-none matter. In short, we are trying to develop not a psychological model of *the* knowing relationship, but a model for treating data about knowing relationships (however defined) which can be applied using whatever knowing relationship happens to be of interest.

The political scientist, using an appropriate definition, may use a contact net model to study influence (Gurevich and Weingrod 1976; n.d.). He asks the number of "connections" of a political kind a person has. The sociologist or anthropologist, using an appropriate definition, may use such a model to study social structure. He asks what kinds of persons are likely to be in contact with each other. The communications researcher may use such a model to study the channels for the flow of messages. Psychologists may use it to examine interrelationships within groups.

So far we have imposed only one restriction on the knowing relationship in our model, namely, that it be all-or-none. There are a few further things we can say about it. When a mathematician describes a relationship he is apt to ask three questions about it: Is it reflexive? Is it symmetric? Is it transitive? The "equals" (=) relationship is reflexive, symmetric, and transitive.

The knowing relationship about which we are talking is clearly not an equality relationship. Anything equals itself; *i.e.,* the equals relation is reflexive. Acquaintanceship is reflexive or not as one chooses to define it. The issue is a trivial one. One could say that by definition everyone knows himself, or one could say that by definition the circle of acquaintances does not include oneself. (We have chosen in our examples below to do that latter and so to define the knowing relation as nonreflexive.)

There is no reason why the knowing relation has to be symmetric. Many more people knew the film star Marilyn Monroe than she knew. If we use the definition of putting a face together with a name then, clearly, persons with good memories know persons with bad memories who do not know them. Similarly, it has been found in some studies that persons are more apt to know the names of persons with higher than lower social status. Thus, privates know each others' names *and* the names of their officers. Officers know each others' names and the names of those they serve, but not necessarily those of privates. Those served may only know servants categorically as, for example, "the tall blond waitress". All in all, to define any knowing relationship as a symmetric one is a great constraint on reality, but it is one which simplifies analysis enormously. It helps so much that for the most part we are going to make that assumption in the discussion below. And, for many purposes, it is largely correct. A kinship relationship is clearly symmetric; if A is a kin to B, B is a kin to A. Also the recognition relationship is mostly symmetric. Most of the time if A can recognize and greet B, B can recognize and greet A. It is generally convenient in our model to define away the minority of cases where this does not hold.

On the other hand, the assumption of transitivity is one that we cannot usefully make. If A knows B, and B knows C, it does not follow that A knows C. If it did follow, then all of society would decompose into a set of one or more cliques completely unconnected with each other. It would mean that everyone you knew would know everyone else you knew, and it follows that you could not know anyone who was outside the clique (*i.e.*, not known to all your friends).[3] Clustering into cliques does occur to some extent and is one of the things we want to study. We want to measure the extent to which these clusters are self-contained, but they are not that by definition.

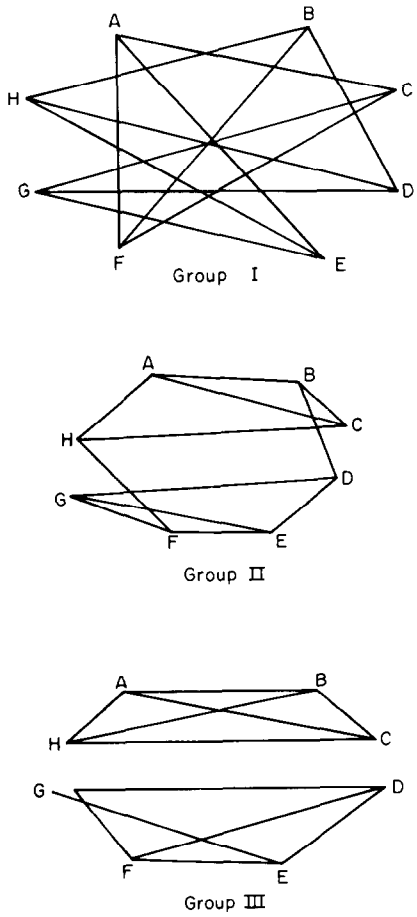
Thus one useful model of a contact network consists of a set of individuals each of whom has some knowing relationships with others in the set of a kind which we have now defined: all-or-none, irreflexive, symmetric, not necessarily transitive.

We would like to be able to describe such a network as relatively unstructured or as highly structured. Intuitively that is a meaningful distinction, but it covers a considerable variety of strictly defined concepts. Figure 1 describes three hypothetical groups of eight people each, in which each individual has three friends. In the first there are no cliques, in the third there are two completely disjoint cliques, and the second group is intermediate. In the first any two people can be connected by at most one intermediary; in the second some pairs (*e.g.*, A and E) require two intermediaries to be connected; in the third some individuals cannot be connected at all. We are inclined to describe the third group as the most stratified or structured and the first as least so,

[3] Most sociometric literature deals with "liking" rather than "knowing". Preference relationships do tend to be transitive (Hallinan and Felmlee 1975).

and in some senses that is true. But, of course, the first graph is also a rigid structure in the sense that all individuals are alike. In general, however, when we talk of a network as showing more social stratification or clustering in this paper, we mean that it departs further from a random process in which each individual is alike except for the randomness of the variables. The clustering in a society is one of the things which affects who will meet whom and who can reach whom.[4] Any congressman knows more congressmen than average for the general populace; any musician knows more musicians.

Figure 1. *Networks of different structuredness.*



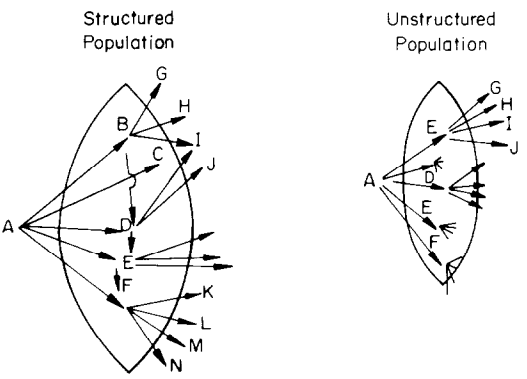
[4] A growing literature exists on structures in large networks (Boorman and White 1976; Lorrain 1976; Lorrain and White 1971; Rapoport and Horvath 1961; Foster *et al.* 1963; Foster and Horvath 1971; Wolfe 1970; McLaughlin 1975; Lundberg 1975; Alba and Kadushin 1976).

The simplest assumption, and one perhaps to start with in modelling a large contact net, is that the number of acquaintances of each person in the population is a constant. We start then with a set of N persons each of whom knows n persons from among the N in the universe; n is the same for all N persons.

If in such a population we pick two persons at random and ask what is the probability that they know each other, the answer can quickly be given from knowing N and n (or, if n is a random variable, the mean n). We know nothing about A and B except that they are persons from a population of size N each of whom on the average knows n other persons in that population. The probability that B is one of the n persons in the circle of acquaintances of A is clearly n/N. If we were talking of a population of 160 000 000 adults and each of them knew, on the average, 800 persons, the chances of two picked at random knowing each other would be one in 200 000.

Suppose we pick A and B who do not know each other, what is the probability of their having an acquaintance in common? The answer to that question, even with random choice of A and B, no longer depends just on n and N. The results now depend also on the characteristic *structure* of interpersonal contacts in the society, as well as on the size of the population and the number of acquaintances each person has. To see the reason why, we turn to an example which we outline diagrammatically in Fig. 2. This Figure represents parts of two networks in which $n = 5$; *i.e.*, each person knows five others in the population. We start with A; he knows B, C, D, E, and F; this is his circle of acquaintances. Next we turn to B; he also knows five people. One of these, by the assumption of symmetry, is A. So, as the acquaintanceship tree fans out, four persons are added at each node.

Figure 2. *Structure in a population.*



However, here we note a difference between the structured and the unstructured population. In a large population without structure the chance of any of A's acquaintances knowing each other is very small (one in 200 000 for the U.S.A. figures used above). So, for a while at least, if there is no

structure the tree fans out adding four entirely new persons at each node: A knows five people; he has 20 friends of friends, and 80 friends of friends of friends, or a total of 125 people reachable with at most two intermediaries. That unstructured situation is, however, quite unrealistic. In reality, people who have a friend in common are likely to know each other (Hammer, n.d.). That is the situation shown in the slightly structured network on the left side of Fig. 2. In that example one of D's acquaintances is B and another is E. The effect of these intersecting acquaintanceships is to reduce the total of different people reached at any given number of steps away from A. In the left-hand network A has five friends, but even with the same n only 11 friends of friends.

So we see, the more cliquishness there is, the more structure there is to the society, the longer (we conjecture) the chains needed on the average to link any pair of persons chosen at random. The less the acquaintanceship structure of a society departs from a purely random process of interactions, in which any two persons have an equal chance of meeting, the shorter will be the average minimum path between pairs of persons.[5] Consider the implications, in a random network, of assuming that n, the mean number of acquaintances of each person, is 1000. Disregarding duplications, one would have 1000 friends, a million (1000^2) friends-of-friends, a billion (1000^3) persons at the end of chains with two intermediaries, and a trillion (1000^4) with three. In such a random network two strangers finding an acquaintance in common (*i.e.*, experiencing the small-world phenomenon) would still be enjoying a relatively rare event; the chance is one million out of 100 or 200 million. But two intermediaries would be all it would normally take to link two people; only a small minority of pairs would not be linked by one of those billion chains.

Thus, in a country the size of the United States, if acquaintanceship were random and the mean acquaintance volume were 1000, the mean length of minimum chain between pairs of persons would be well under two intermediaries. How much longer it is in reality because of the presence of considerable social structure in the society we do not know (nor is it necessarily longer for all social structures). Those are among the critical problems that remain unresolved.

Indeed, if we knew how to answer such questions we would have a good quantitative measure of social structure. Such an index would operationalize the common sociological statement that one society is more structured than

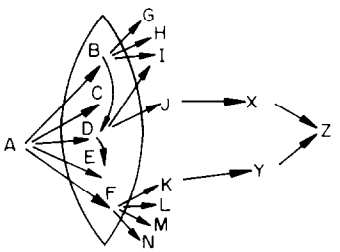
---

[5] Let us state this more carefully for a network of n nodes and m links, in which $n! \gg m$, but all nodes are reachable from all nodes. In that case, m pairs know each other. The question is what structure will minimize the average number of steps between the $n! - m$ remaining pairs. Whenever the m pairs who know each other are also linked at two steps, then the two-step connection is wasted. The same is true for pairs linked by more than one two-step route. Such wastage occurs often when there are dense clusters of closely related nodes in a highly structured network. It happens rarely (because $n! \gg m$) in a random network structure – but it does happen. The minimum average chain would occur not in a random structure, but in one designed to minimize wasted links. However, when $n! \gg m$, the random structure will depart from that situation only to a small extent.

another. The extent to which the mean minimum chain of contacts departs from that which would be found in a random network could be a convenient index of structuredness.

There are all sorts of rules for the topology of a network that can make its graph depart from random linkages. Perhaps the simplest and most important structure is that of triangular links among a given person's friends. If two persons both know person A, the odds are much better than otherwise that they will know each other; if they do know each other the acquaintanceship links form a triangle. For an example see Fig. 3. Disregarding the symmetric

Figure 3. *Effect of structure.*



path (*i.e.*, A knows B so B knows A), let us ask ourselves how many links it takes to go from A out to each of his acquaintances and back to A *via* the shortest path. If we start out on the path from A to B, we can clearly return to A *via* a triangle, A,B,D,A. We can also return by a triangle if we go from A to D or A to E. On the other hand, there is no triangle which will take one back if one starts on the path from A to F. Sooner or later there will be a path back, in this instance a path of eight links. (The only instance in which there would be no path back would be if the society were broken into two cliques linked at no point (see Fig. 1), or at only one point.) Clearly, the number of triangles among all the minimum circular chains is a good index of the tightness of the structure, and one that is empirically usable. It is perfectly possible to sample and poll the acquaintances of A to estimate how many of them know each other. That figure (which measures the number of triangles) then provides a parameter of the kind for which we are looking (Hammer, n.d., Wasserman 1977).

The fact that two persons have an acquaintance in common means that to some extent they probably move in the same circles. They may live in the same part of the country, work in the same company or profession, go to the same church or school, or be related. These institutions provide a nucleus of contacts so that one acquaintance in common is likely to lead to more. One way to describe that situation can be explained if we turn back to Fig. 3. Suppose we inquire of a person whether he knows A. If the answer is yes, then the chances of his knowing B are better than they would otherwise have been. Conversely if the answer is no, that reduces the chances of his knowing B. If he has told us that he does not know either A or B the

chances of his knowing C are still further reduced. And so on down the list. This fact suggests that a second measure of structuredness would be the degree to which the chance of knowing a subsequent person on the list of acquaintances of A is reduced by the information that a person does not know the previous person on the list. In a society that is highly segmented, if two persons have any acquaintances in common they will have many, and so each report of nonacquaintanceship reduces more markedly than it otherwise would the chances of finding one common acquaintance on the list.

We require a measure, such as one of those two we have just been discussing, of the degree of clusteredness in a society, to deal with the question with which we started a few pages back, namely, the distribution of length of minimum contact chains: how many pairs of persons in the population can be joined by a single common acquaintance, how many by a chain of two persons, how many by a chain of three, *etc.*?

The answer depends on three values: N, n, and a parameter measuring structuredness. Increased social stratification reduces the length of chains between persons in the same stratum and at the same time lengthens the chains across strata lines. Thus, for example, two physicians or two persons from the same town are more likely to have an acquaintance in common than persons who do not share such a common characteristic. While some chains are thus shortened and others are lengthened by the existence of clusters within a society, it seems plausible to conjecture that the mean chain averaged over pairs of persons in the population as a whole is lengthened. Two persons chosen at random would find each other more quickly in an unstructured society than in a structured one, for most of the time (given realistic values of N, n, and clustering) persons chosen at random will not turn out to be in the same strata.

We might conjecture, for example, that if we had time series data of this kind running over the past couple of decades, we would find a decline in structuredness based on geography. The increased use of the long-distance telephone (and in the future of computer networks), and also of travel, probably has made acquaintanceship less dependent on geographic location than ever in the past.

In the final section of this paper we turn to an exploration of some of the alternative ways of modelling a network of the kind just described. The central problem that prevents an entirely satisfactory model is that we do not know how to deal with the structuredness of the population. Because of its lovely mathematical simplicity, there is an almost irresistable tendency to want to assume that whenever we do not know how the probability of acquaintanceship within two pairs of persons differs, we should treat it as equal; but it is almost never equal (Hunter and Shotland 1974; White 1970a). The real-world population lives in an n-dimensional space distributed at varying social distances from each other. But it is not a Euclidean space. Person A may be very close to both B and C and therefore very likely to know them both, but B and C may be very far from each other.

In the hope of getting some clues as to the shape of the distribution of closeness among pairs in real-world populations, we undertook some research on the actual contact networks of some 27 individuals. These data we shall describe in Part 2 of this paper. While we learned a lot from that exercise, it failed to answer the most crucial questions because the most important links in establishing the connectedness of a graph may often be not the densely travelled ones in the immediate environment from which the path starts, but sparse ones off in the distance. How to go between two points on opposite sides of a river may depend far more critically on where the bridge is than on the roads near one's origin or destination. The point will become clear as we examine the data.

2. Empirical estimates of acquaintanceship parameters

One is awed by the way in which a network multiplies as links are added. Even making all allowances for social structure, it seems probable that those whose personal acquaintances range around 1000, or only about 1/100 000 of the U.S. adult population, can presumably be linked to another person chosen at random by two or three intermediaries on the average, and almost with certainty by four.

We have tried various approaches to estimating such data. We start with *gedankenexperiments,* but also have developed a couple of techniques for measuring acquaintance volume and network structure.

Consider first a rather fanciful extreme case. Let us suppose that we had located those two individuals in the U.S. between whom the minimum chain of contacts was the longest one for any pair of persons in the country. Let us suppose that one of these turned out to be a hermit in the Okefenokee Swamps, and the other a hermit in the Northwest woods. How many intermediaries do we need to link these two?

Each hermit certainly knows a merchant. Even a hermit needs to buy coffee, bread, and salt. Deep in the backwood, the storekeeper might never have met his congressman, but among the many wholesalers, lawyers, inspectors, and customers with whom he must deal, there will be at least one who is acquainted with his representative. Thus each of the hermits, with two intermediaries reaches his congressman. These may not know each other, though more likely they do, but in any case they know a congressman in common. Thus the maximum plausible minimum chain between any two persons in the United States requires no more than seven intermediaries.

This amusing example is not without significance. Viewed this way, we see Congress in a novel but important aspect, that of a communication node. The Congress is usually viewed as a policy choosing, decision-making instrument, which selects among pre-existing public opinions which are somehow already diffused across the country. Its more important function, however, is that of a forum to which private messages come from all corners, and within which a public opinion is created in this process of confrontation of attitudes

and information. Congress is the place which is quickly reached by messages conveying the feelings and moods of citizens in all walks of life. These feelings themselves are not yet public opinion for they are not crystallized into policy stands; they are millions of detailed worries concerning jobs, family, education, old age, *etc.* It is in the Congress that these messages are quickly heard and are revised and packaged into slogans, bills, and other policy formulations. It is these expressions of otherwise inchoate impulses that are reported in the press, and which become the issues of public opinion. Thus the really important function of the Congress, distinguishing it from an executive branch policy making body, is as a national communication center where public reactions are transformed into public opinion. Its size and geographically representative character puts it normally at two easily found links from everyone in the country. Its members, meeting with each other, formulate policies which express the impulses reaching them from outside. Through this communication node men from as far apart as the Okefenokee Swamps and the north woods can be put in touch with the common threads of each other's feelings expressed in a plank of policy. A body of 500 can help to weld a body of 100 000 000 adults into a nation.

While thinking about such matters has its value, it is no substitute for trying to collect hard data.

Empirical collection of contact data is possible but not easy:

First of all, people are not willing to reveal some or all of their contacts.

Second, it is hard to keep track of such massive and sequential data.

Third, because contacts run in clusters and are not statistically independent events, the statistical treatment of contact data is apt to be hard.

Reticence is probably the least serious of the difficulties. It is certainly no more of a problem for studies of contacts than for Kinsey-type research or for research on incomes or voting behavior, all of which have been successfully conducted, though with inevitable margins of error. As in these other areas of research, skill in framing questions, patience, proper safeguards of confidence, and other similar requirements will determine success, but there is nothing new or different about the difficulties in this field. Reticence is less of an obstacle to obtaining valid information about contacts than are the tricks played by our minds upon attempts at recall.

Indeed it is usually quite impossible for persons to answer questions accurately about their contacts. We noted above the bewilderment which respondents felt when asked how many people they knew, and how most gave fantastic underestimates. Over one day, or even a few hours, recall of contacts is bad. Given more than a very few contacts, people find it hard to recall whom they have seen or conversed with recently. They remember the lengthy or emotionally significant contacts, but not the others. The person who has been to the doctor will recall the doctor, but may neglect to mention the receptionist. The person who has been to lunch with friends may forget about contact with the waiter. In general, contacts which are recalled are demonstrably a highly selected group.

Most importantly, they are selected for prestige. A number of studies have revealed a systematic suppression of reports of contacts down the social hierarchy in favor of contacts up it (Warner 1963; Festinger *et al.* 1950; Katz and Lazarsfeld 1955). If one throws together a group of high status and low status persons and later asks each for the names of the persons in the group to whom he talked, the bias in the outcome is predictably upward. Unaided recall is not an adequate instrument for collecting contact data except where the problem requires recording only of emotionally meaningful contacts. If we wish to record those, and only those, we can use the fact of recall as our operational test of meaningfulness. Otherwise, however, we need to supplement unaided recall.

Some records of contacts exist already and need only be systematically noted. Noninterview sources of contact information include appointment books, committee memberships, and telephone switchboard data. The presidential appointment book is a fascinating subject for study.

Telephone switchboard data could be systematically studied by automatic counting devices without raising any issues of confidence. The techniques are already available and are analogous to those used for making load estimates. They could have great social science value too. A study, for example, of the ecology of long-distance telephone contacts over the face of the country would tell us a great deal about regionalism and national unity. A similar study of the origin and destination of calls by exchange could tell us a great deal about neighborhoods, suburbanism, and urbanism in a metropolitan region. This would be particularly interesting if business and residential phones could be segregated. The pattern of interpersonal contact could be studied by counting calls originating on any sample of telephones. (What proportion of all calls from any one phone are to the most frequently called other phone? What proportion to the 10 most frequently called others?) How many different numbers are called in a month or a year? Would the results on such matters differ for upper and lower income homes, urban and rural, *etc.*?

In similar ways mail flows can tell us a good deal (Deutsch 1956, 1966). The post office data are generally inadequate, even for international flows, and even more for domestic flows. Yet sample counts of geographic origins and destinations are sometimes made, and their potential use is clear.

Not all the information we want exists in available records. For some purposes interviews are needed for collection of data. Various devices suggest themselves for getting at the range of a person's contacts. One such device is to use the telephone book as an *aide-memoire*. We take a very large book, say the Chicago or Manhattan book. We open it to a page selected by a table of random numbers. We then ask our respondent to go through the names on that page to see if they know anyone with a name that appears there or a name that would appear there if it happened to be in that book. Repeat the operation for a sample of pages. One can either require the subject to think of all the persons he knows with such names, which is both tedious and, therefore, unreliable, or assume that the probability of a second, third, or

fourth known person appearing on a single page is independent of the previous appearance of a known name on the page. Since that is a poor assumption we are in a dilemma. Depending on the national origins of our respondent, he is apt to know more persons of certain names; he may know more Ryans, or Cohens, or Swansons according to what he is. Nationality is a distorting factor in the book, too. The Chicago phone book will contain a disproportionate number of Polish names, the Manhattan phone book a disproportionate number of Jewish ones. Also if the subject knows a family well he will know several relatives of the same name. In short, neither the tedious method of trying to make him list all known persons of the name, nor the technique in which one simply counts the proportion of pages on which no known name occurs (and uses that for p, $1 - p = q$, and then expands the binomial), gives a very satisfactory result. Yet with all those qualifications, this technique of checking memory against the phone book gives us a better estimate of approximate numbers of acquaintances than we now have.

One of the authors tried this technique on himself using a sample of 30 pages of the Chicago phone book and 30 pages of the Manhattan phone book. The Chicago phone book brought back names of acquaintances on 60% of the pages, yielding an estimate that he knows 3100 persons. The Manhattan phone book, with 70% of the pages having familiar names, yielded an estimate of 4250 acquaintances. The considerations raised above suggested that the estimate from the Manhattan phone book should be higher, for the author is Jewish and grew up in Manhattan. Still the discrepancy in estimates is large. It perhaps brings us closer to a proper order of magnitude, but this technique is still far from a solution to our problem.

To meet some of these problems we developed a somewhat better method which involves keeping a personal log of all contacts of any sort for a number of sample days. Each day the subject keeps a list (on a pad he carries with him) of all persons whom he meets and knows. The successive lists increasingly repeat names which have already appeared. By projecting the curve one hopes to be able to make estimates of the total size of the acquaintanceship volume, and from the lists of names to learn something of the character of the acquaintances.

The rules of inclusion and exclusion were as follows:

(1) A person was not listed unless he was already known to the subject. That is to say, the first time he was introduced he was not listed; if he was met again on a later day in the 100 he was. The rationale for this is that we meet many people whom we fail to learn to recognize and know.

(2) Knowing was defined as facial recognition and knowing the person's name – any useful name, even a nickname. The latter requirement was convenient since it is hard to list on a written record persons for whom we have no name.

(3) Persons were only listed on a given day if when the subject saw them he addressed them, if only for a greeting. This eliminated persons seen at a distance, and persons who the subject recognized but did not feel closely enough related to, to feel it proper to address.

Table 1. *100-day contacts of respondents*

Sex	Job	Age	*(a)* No. of different persons seen in 100 days	*(b)* No. of contact events	Ratio *b/a*
Blue collar					
M	Porter	50 - 60	83	2946	35.5
M	Factory labor	40 - 50	96	2369	24.7
M	Dept. store receiving	20 - 30	137	1689	12.3
M	Factory labor	60 - 70	376	7645	20.3
M	Foreman	30 - 40	510	6371	12.5
F	Factory labor and unemployed	30 - 40	146	1222	8.4
White collar					
F	Technician	30 - 40	276	2207	8.0
F	Secretary	40 - 50	318	1963	6.2
M	Buyer	20 - 30	390	2756	7.1
M	Buyer	20 - 30	474	4090	8.6
M	Sales	30 - 40	505	3098	6.1
F	Secretary	50 - 60	596	5705	9.5
Professional					
M	Factory engineer	30 - 40	235	3142	13.5
F	T.V.	40 - 50	533	1681	3.2
M	Adult educator	30 - 40	541	2282	4.2
M	Professor	40 - 50	570	2175	3.8
M	Professor	40 - 50	685	2142	3.1
M	Lawyer-politician	30 - 40	1043	3159	3.0
M	Student	20 - 30	338	1471	4.4
M	Photographer	30 - 40	523	1967	4.8
M	President*	50 - 60	1404**	4340**	3.1**
Housewives					
F	–	30 - 40	72	377	5.2
F	–	20 - 30	255	1111	4.4
F	–	20 - 30	280	1135	4.0
F	–	30 - 40	363	1593	4.4
F	–	30 - 40	309	1034	3.3
F	–	50 - 60	361	1032	2.9
Adolescent					
M	Student	10 - 20	464	4416	9.5

*Data estimated from Hyde Park records.
**Record for 85 days.

(4) Telephone contacts were included. So were letters written but not letters received. The rationale for the latter is that receiving a letter and replying to it is a single two-way communication such as occurs simultaneously in a face-to-face contact. To avoid double counting, we counted a reply as only half the act. Of course, we counted only letters written to people already known by the above criterion.

(5) A person was only listed once on a given day no matter how often he was seen. This eliminated, for example, the problem of how many times to count one's secretary as she walked in and out of the office.

The task of recording these contacts is not an easy one. It soon becomes a tedious bore. Without either strong motivation or constant checking it is easy to become forgetful and sloppy. But it is far from impossible; properly controlled and motivated subjects will do it.

The data on 27 persons were collected mostly by Dr. Michael Gurevich (1961) as part of a Ph.D. dissertation which explored, along with the acquaintanceship information itself, its relation to a number of dependent variables. As Table 1 shows us, the respondents, though not a sample of any defined universe, covered a range of types including blue collar, white collar, professional, and housewives.

Among the most important figures in the Table are those found in the right-hand column. It is the ratio between the number of different persons met and the number of meetings. It is what psychologists call the type–token ratio. It is socially very indicative, and is distinctive for different classes of persons.

Blue collar workers and housewives had the smallest number of different contacts over the 100 days. They both lived in a restricted social universe. But in the total number of interpersonal interactions the blue collar workers and housewives differed enormously. Many of the blue collar workers worked in large groups. Their round of life was very repetitive; they saw the same people day in and day out, but at work they saw many of them. Housewives, on the other hand, not only saw few different people, but they saw few people in the course of a day; they had small type–token ratios. They lived in isolation.

In total gregariousness (*i.e.*, number of contact events) there was not much difference among the three working groups. Blue collar workers, white collar workers, and professionals all fell within the same range, and if there is a real difference in the means, our small samples do not justify any conclusions about that. But in the pattern of activity there was a great difference. While blue collar workers were trapped in the round of a highly repetitive life, professionals at the other extreme were constantly seeing new people. They tended to see an average acquaintance only three or four times in the hundred days. One result of this was that the professionals were the persons whose contacts broke out of the confines of social class to some extent. They, like the others (see Table 2) tended to mix to a degree with people like themselves but, to a slightly greater degree than the other classes, they had a chance to meet people in other strata of society.

The tendency of society to cluster itself as like seeks like can also be seen in Tables on contacts by age, sex, and religion (see Tables 3, 4 and 5). These data reflect a society that is very structured indeed. How can we use the data to estimate the acquaintanceship volume of the different respondents? We found that over 100 days the number of different persons they saw ranged between 72 for one housewife and 1043 for one lawyer-politician. Franklin

Table 2. *Number of acquaintances by occupation*

Acquaintances' occupation	Subject's occupation				
	Blue collar (%)	Housewife (%)	White collar (%)	Professional (%)	Entire group (%)
Professional	11	24	20	45	24
Managerial	9	7	19	14	14
Clerical	13	7	13	7	11
Sales worker	5	6	19	4	11
Craftsman, foreman	15	5	6	5	7
Operative	25	1	3	5	8
Service worker	9	2	2	1	3
Laborer	4	1	1	–	1
Housewife	4	35	10	12	13
Student	2	3	1	5	3
Farmer	–	–	–	–	–
Dont' know	4	10	8	3	6
	100*	100*	100*	100*	100*

*Figures may not add up to 100% because of rounding.

Table 3. *Subject's age compared with his acquaintance's age*

Acquaintance's age	Subject's age			
	20 - 30 (%)	31 - 40 (%)	41 - 50 (%)	Over 50 (%)
Under 20	7	2	2	1
20 - 30	21	19	11	15
31 - 40	30	39	33	20
41 - 50	21	22	27	32
Over 50	21	19	27	33
	100*	100*	100*	100*

*Figures may not add up to 100% because of rounding.

Roosevelt's presidential appointment book, analyzed by Howard Rosenthal (1960), showed 1404 different persons seeing him. But that leaves us with the question as to what portion of the total acquaintance volume of each of these persons was exhausted.

One of the purposes of the data collection was to enable us to make an estimate of acquaintance volume in a way that has already been described above. With each successive day one would expect fewer people to be added, giving an ogive of persons met to date such as that in Fig. 4. In principle

Table 4. *Sex of subject and sex of acquaintance*

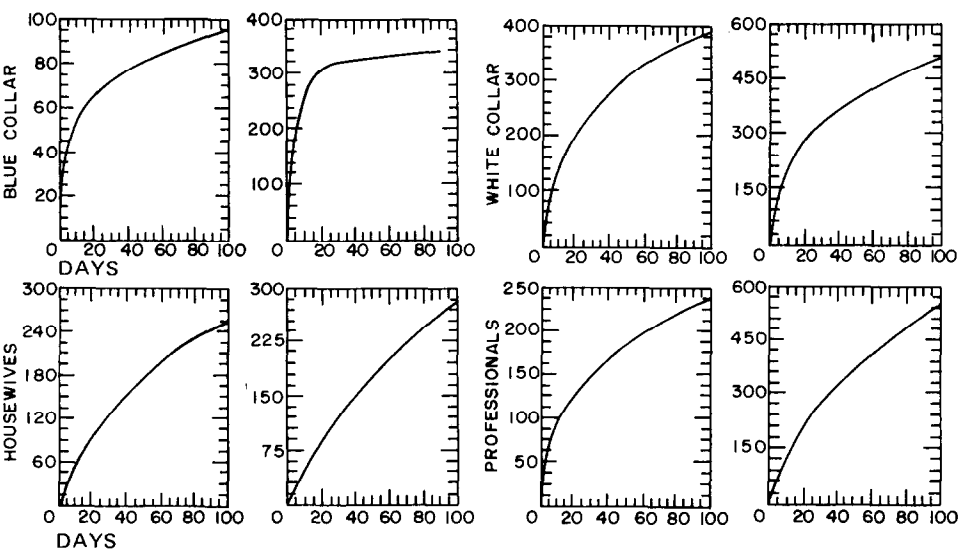
Subject	Acquaintances		
	Male (%)	Female (%)	Total (%)
Blue collar			
Male	83	17	100
White collar			
Male	65	35	100
Female	53	47	100
Professional			
Male	71	29	100
Housewife			
Female	45	55	100

Table 5. *Religion of subject and religion of acquaintance*

Subject's religion	Acquaintance's religion				
	Protestant	Catholic	Christian (didn't know denomination)	Jewish	Religion known
	(%)	(%)	(%)	(%)	(%)
Protestant	46	25	25	4	100*
Catholic	15	57	23	5	100*
Jewish	9	16	27	47	100*

*Figures may not add up because of rounding and omission of other religions.

Figure 4. *Acquaintanceship ogives.*



one might hope to extrapolate that curve to a point beyond which net additions would be trivial.

Fitting the 100-day curve for each subject to the equation (acquaintanceship volume) = At^x gave acquaintanceship volumes over 20 years ranging from 122 individuals for a blue collar porter in his fifties to 22 500 persons for Franklin Roosevelt.

However, that estimation procedure does not work with any degree of precision. The explanation is that the estimate of the asymptote is sensitive

Table 6. *Frequency distribution of contacts with acquaintances*

Frequency of contact over 100 days	Blue collar group				
	Case A (%)	Case B (%)	Case C (%)	Case D (%)	Case E (%)
1	4.8	23.9	29.0	9.3	23.5
2	2.4	11.4	11.6	5.0	10.7
3	–	4.1	6.5	3.9	8.4
4	–	4.1	4.3	3.4	4.7
5	1.2	3.1	3.6	3.4	4.9
6 - 10*	2.4	0.4	1.7	3.4	2.2
11 - 20*	0.8	0.5	1.2	2.1	1.3
21 - 30*	1.0	0.6	1.0	1.3	1.0
31 - 40*	1.8	0.6	0.6	0.9	0.7
41 - 50*	1.7	0.3	0.5	0.5	0.4
51 - 60*	1.7	1.4	0.1	0.4	0.2
61 - 70*	0.6	1.1	–	0.7	0.1
71 - 80*	0.1	0.1	0.07	–	0.02
81 - 90*	–	–	–	–	–
91 - 100*	0.2	0.2	0.07	0.05	0.02
	100%	100%	100%	100%	100%

Frequency of contact over 100 days	White collar group					
	Case G (%)	Case H (%)	Case I (%)	Case J (%)	Case K (%)	Case L (%)
1	43.4	44.3	27.2	30.8	47.7	37.7
2	11.5	16.9	20.0	12.4	13.1	12.9
3	7.9	7.5	10.7	9.0	6.5	7.5
4	4.3	3.7	6.1	6.9	7.1	4.5
5	3.2	3.4	6.1	4.0	3.2	3.0
6 - 10*	1.9	1.8	2.3	2.8	1.9	2.3
11 - 20*	0.7	0.8	0.7	1.1	0.6	0.9
21 - 30*	0.4	0.3	0.4	0.4	0.2	0.3
31 - 40*	0.3	–	0.2	0.2	0.2	0,3
41 - 50*	0.5	0.09	0.1	0.2	0.1	0.3
51 - 60*	0.1	0.1	0.2	0.2	0.1	0.4
61 - 70*	–	0.2	–	0.1	0.06	0.1
71 - 80*	–	–	–	–	–	–
81 - 90*	–	–	–	–	–	–
90 - 100*	0.04	0.03	0.03	0.02	0.02	0.02
	100%	100%	100%	100%	100%	100%

(continued on facing page)

Table 6. *(continued)*

Frequency of contact over 100 days	Professionals				Housewives		
	Case M (%)	Case O (%)	Case P (%)	Case Q (%)	Case V (%)	Case W (%)	Case X (%)
1	39.5	53.0	43.3	49.6	56.0	54.6	47.9
2	7.7	12.3	17.5	18.5	18.8	18.9	16.5
3	4.3	7.5	12.2	10.9	7.8	7.8	8.8
4	3.9	4.2	5.9	4.7	1.5	3.2	6.8
5	3.0	3.6	5.2	3.8	3.9	2.5	4.4
6 - 10*	1.2	2.3	1.8	1.3	1.1	1.3	1.6
11 - 20*	1.6	0.4	0.5	0.3	0.3	0.4	0.3
21 - 30*	0.4	0.09	0.07	0.09	0.04	0.04	0.2
31 - 40*	0.4	0.07	0.02	0.06	0.08	0.04	0.03
41 - 50*	0.3	0.05	0.05	0.01	0.04	–	0.1
51 - 60*	0.7	0.07	0.02	0.01	0.08	0.1	–
61 - 70*	0.1	–	–	–	–	–	–
71 - 80*	–	–	–	–	0.04	0.07	–
81 - 90*	–	–	–	–	–	–	0.03
91 - 100*	0.1	0.02	0.02	0.01	0.08	0.04	0.03
	100%	100%	100%	100%	100%	100%	100%

*The percentages in each entry are average percentages for a single day, not for the 5- or 10-day period.

to the tail of the distribution (Granovetter 1976). Such a large proportion of the respondent's acquaintances are seen only once or twice in 100 days that any estimate which we make from such data is very crude. Table 6 shows the figures. Except for blue collar workers, half or more of the acquaintances were seen only once or twice in the period.

One may think that the way around this problem would be to rely more heavily on the shape of the curve in its more rugged region where contact events are more frequent. The problem with that is that the nature of the contacts in the two parts of the curve are really quite dissimilar. To explain that perhaps we should look more closely at a single case; we shall use that of one of the author's own contact lists.

In 100 days he had contact with 685 persons he knew. On any one day the number of contacts ranged from a low of two other persons to a high of 89, the latter in the Christmas season. The mean number of acquaintances with whom he dealt on a day was 22.5. The median number was 19. There were several discreet typical patterns of days, resulting in a multimodal distribution. There was one type of day, including most weekend days, when he would typically meet 7 - 9 people, another type of day with typically around 17 contacts, and a third type of day of highly gregarious activity which involved dealing with about 30 people.

Only about half of the 685 persons were seen more than once in the 100 days. The mean frequency was 3.1 times per person. The distribution, however, is highly skewed (Table 7).

Table 7. *Contact frequency distribution for one person*

Number of days on which contact was had during the 100 days	Number of persons with that frequency of contact	Days	Persons	Days	Persons
1	335	11	4	24	1
2	125	12	4	26	2
3	74	13	1	30	1
4	32	14	2	33	2
5	26	15	4	34	1
6	12	16	2	36	1
7	16	18	1	45	1
8	5	19	1	51	1
9	8	20	4	92	1
10	4	23	2		

These figures, however, are somewhat misleading. It seems that we are actually dealing with two distributions: one which includes those persons living in the author's home and working in his office whom he saw during his regular daily routine, and the other including all his other acquaintances in the seeing of whom all kinds of chance factors operatored. All individuals seen 19 or more times are in the former group; so are all but two individuals seen 13 or more times. Removing 51 such family members and co-workers gives us the data that are really relevant to estimating the large universe of occasional contacts, but in that sample more than half the persons listed were seen only once and 91% five times or less. No easily interpretable distribution (such as Poisson which would imply that there is no structure among these contacts) fits that distribution, and with such small frequencies the shape of the distribution is unstable between respondents. It is possible that the projection of the 100-day data for this author to a year's time could come out at anywhere between 1100 and 1700 persons contacted. That is not a very satisfactory estimate, but it is far better than the estimates we had before.

This estimate is way below our telephone book estimates, which it will be recalled ranged from 3100 to 4250 acquaintances. The discrepancy is more revealing than disturbing. It suggests some hypotheses about the structure of the universe of acquaintances. It suggests that there is a pool of persons with whom one is currently in potential contact, and a larger pool in one's memory, which for the senior author is about 2 - 3 times as large. The active pool consists of acquaintances living in the areas which one frequents, working at the activity related to one's occupation, belonging to the groups to which one belongs. Random factors determine in part which persons out of this pool one happens to meet, or even meet several times during any set period. But in one's memory there are in addition a considerable number of other persons whose names and faces are still effectively stored, but who are

not currently moving in the same strata of contacts as oneself. These are recorded by the telephone book measure; they will not appear in the record of meetings except for the rarest kind of purely chance encounter. Needless to say, these two pools are not clearly segregated, but merge into each other. Yet, our data would suggest that they are more segregated than we would otherwise have suspected. The probabilities of encounter with the two types of persons are of quite different orders of magnitude.

We have now established plausible values for some of the parameters of the contact net of one of the authors. He typically deals with about 20 people in a day. These are drawn from a set of some 1500 persons whom he actively knows at the present time. At the same time he remembers many other persons and could still recognize and name perhaps 3500 persons whom he has met at some point in the past. (Incidentally, he has never regarded himself as good at this.)[6]

The remaining parameter which we would wish to estimate is the degree of structuredness in this acquaintanceship universe. The indicator that we proposed to use was the proportion of the acquaintances of the list-keeper who knew each other; *i.e.*, the proportion of triangles in the network graph. When the 100-day data collection was finished, we took the lists of some of the respondents and turned them into a questionnaire. To a sample of the people who appeared on the respondent's list of contacts, we sent a sample of the names on the list and asked, regarding each, "Do you know that person?". This provided a measure of the degree of ingrowth of the contact net. It can be expressed as the percentage of possible triangles that are completed (Wasserman 1977). The values for five subjects from whom we got the data ranged from 8 to 36%, and we would speculate that a typical value lies toward the low end of this range.

We have indicated above that the degree of structure affects how much longer than chance the minimum chain between a pair of randomly chosen persons is apt to be. We can go no further in specifying the effect of structure on the chains in this qualitative verbal discussion. Any more precise conclusion depends on the treatment of this subject in a much more formal mathematical way. We turn, therefore, to a restatement of our presentation in a mathematical model.

3. Mathematical models of social contact

To describe with precision the structure of human acquaintance networks, and the mechanisms by which social and political contacts can be established within them, it is necessary to idealize the empirical situation with a model. Models have been used effectively in a number of related fields. Rapoport

[6]The $n = At^X$ fitted curve for this author's ogive reached that level in just 5 years, but without taking account of forgetting.

and others have modelled the flow of messages in a network (Rapoport and Horvath 1961; Foster *et al.* 1963; Foster and Horvath 1971; Rapoport 1963; Kleinrock 1964). Related models use Markov chains, queuing theory and random walks (White 1970b, 1973). Most such models, however, depend critically upon an assumption that the next step in the flow goes to other units in the model with a probability that is a function of the present position of the wanderer. The problem that we are addressing does not lend itself to that kind of model; the probability of contact between any two persons is a function of a long-established continuing relationship that inheres in them. The model required for our purposes must be one which retains a characterization of the relationship of each pair of individuals.

Nonetheless, it is useful to begin our analysis with the simplest models in order to develop the needed framework within which to formulate the essential problems. Two extreme situations are relatively easy to analyze. The first is one in which the number of individuals is sufficiently small so that combinatorial methods are still feasible. The second is one in which there are so many individuals that we can treat it as an infinite ensemble, applying methods similar to those used in statistical mechanics. The hard problems deal with conditions between these two extremes.

Graph-theoretic models

Let P denote a group of N people. We shall represent the individuals by integers $1,...,i,...,N$. We draw a directed line or arrow from individual i to individual j to indicate that i knows j. This can be presented as a directed graph, shown in Fig. 5 for $N = 5$, and also represented by an incidence matrix in Fig. 6, where a one is entered in the cell of row i and row j if i knows j and a zero otherwise. If we assume the knowing relation to be symmetric, then every arrow from i to j is side by side with an arrow from j to i – and the incidence matrix is symmetric as well – and we may as well use undirected edges. Let M be the total number of edges or mutual knowing-bonds.

Figure 5. *A directed graph.*

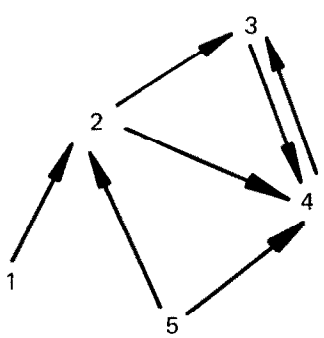


Figure 6. *An incidence matrix.*

	1	2	3	4	5
1	–	1	0	0	0
2	0	–	1	1	0
3	0	0	–	1	0
4	0	0	1	–	0
5	0	1	0	1	0

The incidence matrix has N rows and N columns, but only $(N^2 - N)/2$ of its elements can be chosen freely for a symmetric irreflexive (or reflexive) knowing relation. Thus, there can be at most $(N^2 - N)/2$ pairs or edges. Generally, $0 \leqslant M \leqslant (N^2 - N)/2$. If M takes the largest value possible, then every individual knows every other; if $M = 0$, then no individual knows any other. There is just one structure corresponding to each of these extreme cases. If $M = 1$, there are $(N^2 - N)/2$ possible structures, depending on which pair of people is the one. If $M = 2$, there are $\binom{N^2 - N/2}{2}$ possible structures, and there are altogether $2^{(N^2 - N)/2}$ possible structures corresponding to $M = 0,1,2,\ldots,(N^2 - N)/2$. The number of possible structures is largest when $M = (N^2 - N)/4$.

Let **U** denote the symmetric incidence matrix, and let u_{ij} be (0 or 1) its element in row i, column j. Let $u_{ij}^{(k)}$ denote the corresponding element in the symmetric matrix $\mathbf{U}^k$. This represents the number of different paths of exactly k links between i and j (Luce 1950; Doreian 1974; Peay 1976; Alba 1973). A path is an adjacent series of links that does not cross itself. Two paths are called distinct if not all the links are identical. Thus, there are exactly two 2-step paths from 5 to 3 in Fig. 5, one *via* 4 and one *via* 2; multiplying **U** by itself (with 0 in the diagonals) gives

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 1 & 0 \end{bmatrix}$$

and the element in row 5, column 3 is clearly 2, since matrix multiplication calls for the sum of the products of the elements in row 5, (0 1 0 1 0), and the elements in column 3, (0 1 0 1 0), which is $0 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 = 2$.

It follows that $u_{ii}^{(3)}$ is the number of triangles that start and end with individual i. Each individual could be the start–end point of as many as $\binom{N-1}{2}$ different triangles, or as few as 0. If $u_{ii}^{(3)} = 0$ for all i, then there cannot be any tightly knit cliques; if $u_{ii}^{(3)} \geqslant 1$ for all i, then there is a considerable degree of connectedness and structure.

Let n denote the number of others each individual knows. This is the number of 1's in each row and each column of the incidence matrix or the number of edges incident on each node of the graph. Let α_k be the sum of all the elements in $\mathbf{U}^k$. It follows that $\alpha_1 = 2M$, and α_2 is twice the number of length-2 paths, which could serve as an index of clustering.

If each of a person's n acquaintances knew one another, **U** would consist of N $(n + 1) \times (n + 1)$ matrices consisting of all 1's (except for the diagonal) strung out along the diagonal, assuming that $n + 1$ divides N. Here no individual in one cluster knows anyone in a different cluster.

Such combinatorial, graph-theoretic approaches are intuitively appealing and have considerable descriptive power. There is also a number of theorems for counting the number of different configurations, such as Polya's theorem,

as well as computer-based techniques for eliminating structures, such as Lederberg's creation of a language, DENDRAL, for representing the topology of molecules. Graph-theoretic theorems, however, have to ignore reality to introduce assumptions leading to mathematically interesting applications or else follow the scientifically unnatural approach of starting with strong but far-fetched assumptions and relaxing them as little as possible to accommodate reality. The limitations of combinatorial methods become clearest when their computational complexity is studied. Multiplying matrices is of polynomial complexity, requiring of the order of N^3 multiplications; for sparse incidence matrices this can be reduced. But tracing out various configurations or finding a specified path can be much more complex, so that it cannot even be done by computer. Moreover, there is no realistic way that data can be obtained to fill in the elements of U for a nation,[7] and different ways of representing acquaintanceship among millions of people must be found. Even storing who knows whom among millions is a non-trivial problem, and more efficient ways of processing such data than are provided by conventional ways of representing sets such as P by ordering its elements $1,\ldots,N$ must be used. The problems of processing data about social networks and drawing inferences from them have received considerable attention, but still face serious obstacles (Wasserman 1977; Holland and Leinhardt 1970; Breiger *et al.* 1975; Granovetter 1974; Newcomb 1961).

Statistical models with independence and no structure

We now take advantage of the large size of N, typically 10^8 or greater, corresponding to the population of a country such as the U.S. We select any two individuals A and B at random from such a large population P. We would like to estimate the distribution of the shortest contact chain necessary for A and B to get in touch.

Let $k = 0$ mean that A and B know one another, that a direct link exists. We have a chain of one link with $k = 0$ intermediaries. But $k = 1$ means that A and B do not know one another, yet have a common acquaintance. It is a chain with two links and one intermediary. $k = 2$ means that A and B do not even have a mutual acquaintance but A knows someone who knows B. It is a chain with three links and two intermediaries.

Let p_k be the probability of a chain with exactly k intermediaries, $k = 1,2,\ldots$. We approximate p_0 by n/N, the ratio of each person's total number of acquaintances to the total population size. Thus, if A knows 1000 people

[7]The use of bibliometric data – for example, who co-authored with whom, who cited whom, which can be obtained in computerized form from the Institute for Scientific Information in Philadelphia for much of the world's scientific literature – may be a practical source. Mathematicians have for some time used the term "Erdös number", which is the distance between any author and Paul Erdös in terms of the number of intermediary co-authors; *e.g.*, A may have co-authored with B who co-authored with C who co-authored with Erdös, making the Erdös distance 2 from A. The use of co-citation and similar data also appears promising (Griffith *et al.* 1973).

out of 100 000 000 Americans (other than A) then the probability of his knowing a randomly chosen person B among the 100 000 000 is $10^3/10^8 = 10^{-5}$.

Let q_0 be the probability that B does not know A. This is $q_0 = 1 - p_0$. It is the probability that one of A's acquaintances is not B. If we now make the strong assumption that the corresponding probability of a second of A's acquaintances is also not B, nor is it affected by knowledge of the probability of the first of A's friends not being B, then the probability that none of the n of A's acquaintances is B is q_0^n. This corresponds to a random or unstructured acquaintance net.

The probability p_1 that A and B are not in direct contact but have at least one common acquaintance is $q_0(1 - q_0^n)$. This assumes that B not being in direct contact with A is also independent of B not being in direct contact with each of the n people whom A knows.

Similarly, we estimate: $p_2 = q_0 q_0^n (1 - q_0^{n^2})$.

This uses another simplifying assumption: each of A's n acquaintances has n *new* acquaintances that will not include any of A's n acquaintances nor any acquaintance of his acquaintances. Thus, there are altogether n^2 different people who are the friends of A's friends. Thus if A knows 1000 people, their friends number a million people not assumed to be counted so far.

If we extend these assumptions for the general case, we have

$$p_k = q_0 q_0^n q_0^{n^2} \ldots q_0^{n^{k-1}} (1 - q_0^{n^k})$$

$$= (1 - p_0)^{(n^k - 1)/(n-1)} [1 - (1 - p_0)^{n^k}] \qquad k = 1, 2, 3, \ldots \qquad (1)$$

Table 8. *Distribution of contact in an unstructured net*

	$n = 500$	$n = 1000$	$n = 2000$
p_0	0.00000500	0.00001000	0.00002000
p_1	0.00249687	0.00995012	0.03921016
p_2	0.71171102	0.98999494	0.96076984
$\sum_{k=3}^{n-1} p_k$	0.28578711	0.00004495	0.00000000
Mean	2.28328023	1.99007483	1.96074984
Variance	0.20805629	0.00993655	0.03774959

Table 8 shows some typical numbers for $N = 10^8$. The numbers were computed using equation 1 on the University of Michigan 470/V6. Note that the average number of intermediaries is 2 (when $n = 1000$), and the average chain is three lengths, with very little variation around that mean. Nor is that average sensitive to n, a person's acquaintance volume. This is not implau-

sible for, according to the above assumptions, if a person knows 1000 people (in one remove), then in two removes he reaches 1000×1000, and in three removes 10^9, which exceeds a population of 10^8, according to a simple and intuitive analysis. This result is, however, very sensitive to our independence assumption. The probability of a randomly chosen person C knowing A, given that he knows a friend of A, is almost certainly greater than the unconditional probability that C knows A. (The latter should also exceed the conditional probability of C knowing A, given that C does not know any friend of A.) We turn next to models that do not depend on this independence assumption.

The number of common acquaintances

The independence assumptions of the last section imply that the probability of A having exactly k acquaintances in common with randomly chosen B is $\binom{n}{k} p_0^k q_0^{n-k}$. Here, p_0^k is the probability that k out of the n acquaintances of A each knows B and that each of his remaining $n - k$ acquaintances do not know B; there are $\binom{n}{k}$ ways of selecting these k from the n people whom A knows. The mean of this binomial distribution is np_0 and the variance np_0q_0.

If $n = 10^3$ and $N = 10^8$ then $p_0 = 10^{-5}$, $q_0 = 1 - 10^{-5}$ and the average number of common acquaintances is approximately 0.01 with a variance of 0.01. This is far too small to be realistic, and it points out the weakness of the independence assumption.

One way to replace it is to define p_0', the conditional probability that a randomly chosen friend of A knows randomly chosen person B', given that B' also knows A. This should exceed p_0 or n/N. A plausible estimate for the probability that two of A's friends know each other is $1/(n - 1)$, because there are $n - 1$ people from whom a friend of A could be chosen with whom to form an acquaintance bond. The probability that k of A's friends each knows another friend could now be estimated to be $(p_0')^k$ or $(n - 1)^{-k}$, if we assume independence of acquaintance among A's friends. Similarly, $(1 - p_0')^{n-k}$ is an estimate of the probability that $n - k$ of A's friends do not know another of A's friends. As before, the mean number of common acquaintances is np_0', which is $n/(n - 1)$, or close to 1, with a variance of $np_0'q_0'$, which is close to 0. This, too, is too small for realism, however.

Consider next an approach that relates recursively the average number of acquaintances common to k individuals chosen at random. Call this m_k and assume that

$$m_{k+1} = am_k, \quad m_1 = n, \quad k = 2, 3, \ldots \tag{2a}$$

This means that the average number of acquaintances common to four people is smaller than the average number common to three by a fraction, a, which is the same proportion as the number of friends shared by three is to the number shared by two. This constant a is between 0 and 1 and would have to be statistically estimated. It is assumed to be the same for all $\binom{n}{k}$ groups of k people.

p_0, the probability of A knowing a randomly chosen person B, is n/N or m_1/N, as before. If $m_2 = am_1$, then $n/N = (m_2/a)/N$ and $a = m_2/n$. Thus, if we could estimate the number of acquaintances shared by two people, we could estimate a. Thus, we can set the number of common acquaintances, m_2, to any value we please, and use it to revise the calculation of p_k from what it was in the last section.

p_1, the probability that A does not know randomly chosen B but knows someone who knows B, is $(1 - p_0) \times$ Prob {A and B have at least one common acquaintance}. The latter is the number of ways of choosing a person out of the n people A knows so that he is one of the m_2 common acquaintances, or m_2/n. Thus,

$$p_1 = (1 - p_0)m_2/n$$

and

$$p_2 = (1 - p_0)(1 - p_1)p_2'$$

To calculate p_2', the probability that B knows someone who is a friend of one of A's n acquaintances, we need n', the number of different persons known to the n acquaintances of A. Then we could estimate p_2' by

$$p_2' = \binom{n'}{1}\frac{m_1}{N} - \binom{n'}{2}\frac{m_2}{N} + \binom{n'}{3}\frac{m_3}{N} - \binom{n'}{4}\frac{m_4}{N} + \ldots \pm \binom{n'}{n'}\frac{m_{n'}}{N} \tag{2b}$$

Here $\binom{n'}{k}m_k$ is the number of ways that B could be one of the m_k acquaintances common to some k of the n' friends of A's friends. It follows from eqn. (2a) that

$$m_2 = am_1$$

$$m_3 = am_2 = a(am_1) = a^2m_1$$

and generally that

$$m_k = a^{k-1}n \tag{2c}$$

Substituting into eqn. (2b), we can show that

$$p_2' = \frac{n}{aN}\left[1 - (1 - a)^{n'}\right]$$

To estimate n', we note that of all A's n friends, m_2 are also known to one other person, m_3 to two others, *etc.* Thus,

$$n' = \binom{n}{1}m_1 - \binom{n}{2}m_2 + \binom{n}{3}m_3 - \binom{n}{4}m_4 + \ldots \pm \binom{n}{n}m_n$$

$$= n\left[\binom{n}{1} - \binom{n}{2}a + \binom{n}{3}a^2 - \binom{n}{4}a^3 + \ldots \pm \binom{n}{n}a^{n-1}\right]$$

$$= \frac{n}{a}\left[\binom{n}{1}a - \binom{n}{2}a^2 + \binom{n}{3}a^3 - \ldots \pm \binom{n}{n}a^n\right]$$

$$= \frac{n}{a}[1 - (1 - a)^n] \text{ by the binomial theorem} \tag{2d}$$

Hence,

$$p_2 = (1 - p_0)(1 - p_1)(n/aN)[1 - (1 - a)^{n'}]$$

and

$$p_3 = (1 - p_0)(1 - p_1)(1 - p_2)(n/aN)[1 - (1 - a)^{n''}]$$

where

$$n'' = (n/a)[1 - (1 - a)^{n'}]$$

We can set up a recursive equation for p_k in general. We can also require it to hold for $k = 1$, in which case we should expect that

$$m_2/n = (n/aN)[1 - (1 - a)^n] = a \tag{2e}$$

If $n = 10^3$ and $N = 10^8$, then a should be such that $(10^{-5}/a)[1 - (1 - a)^{1000}] = a$. This is a transcendental equation to be solved for a, and the value of $a = 0.003$ is an approximate solution because $10^{-5}[1 - (1 - 0.003)^{1000}]$ is approximately $(0.003)^2$ or 9×10^{-6}, which is reasonably close. A value for $a = 0.003$ or $m_2 = 3$ is no longer so unreasonable for the number of acquaintances common to two people chosen at random. The assumption expressed in eqn. (2a) now implies that m_3, the number of acquaintances common to three people, is $(0.003) \times 3$ or 0.009, which is effectively zero. This is too small to be realistic. Using these values, we obtain,

$p_0 = 0.00001$, as before

$p_1 = 0.003$ compared with 0.009949

$p_2 = 0.00332$ compared with 0.99001

$p_3 = 0.00330$

$n' = 381\,033$, $n'' = 333\,333$

The distribution of k is now considerably flattened, with chains of short length no less improbable than chains of greater length. This is due to a value of a greater than 10^{-5}, as specified by a chosen value of m_2 and eqn. (2e).

The above analysis, though more realistic, is still limited by an independence assumption and the low value of $m_3, m_4, \ldots$. Yet it may be fruitful to explore it further by exploiting the sensitivity of these results to m_2, or replacing eqn. (2a) by one in which a is not constant. We now proceed, however, to replace this approach by defining the following conditional probabilities.

Let K_A be A's circle of acquaintances, with $\bar{K}_A$ its complement. Let $A_1, \ldots, A_n$ denote the individuals in it. Consider:

$\text{Prob}(B \in \bar{K}_{A_1})$, $\text{Prob}(B \in \bar{K}_{A_2} | B \in \bar{K}_{A_1})$, $\text{Prob}(B \in \bar{K}_{A_3} | B \in \bar{K}_{A_2}, B \in \bar{K}_{A_1})$, *etc.* The product of these probabilities is $\text{Prob}(B \in \bar{K}_{A_1} \cap \bar{K}_{A_2} \cap \bar{K}_{A_3} \cap \ldots)$, the probability that a randomly chosen B is not known to each of A's acquaintances.

A simple and perhaps plausible assumption other than independence is that of a Markov chain:

$$\text{Prob}(B \in \bar{K}_{A_k} | B \in \bar{K}_{A_{k-1}}, \ldots, B \in \bar{K}_{A_1}) = \text{Prob}(\bar{K}_{A_k} | \bar{K}_{A_{k-1}}) = b$$

where b is a constant to be statistically estimated.

Thus,

$$\text{Prob}(\bar{K}_{A_n}, \bar{K}_{A_{n-1}}, \ldots, \bar{K}_{A_1}) = \text{Prob}(\bar{K}_{A_1}) b^{n-1} = (1 - n/N) b^{n-1}$$

For $k = 2$,

$$\text{Prob}(\bar{K}_{A_2}, \bar{K}_{A_1}) = (1 - n/N)b = 1 - 2n/N + m_2/N$$

Hence

$$b = \frac{1 - 2n/N + m_2/N}{1 - n/N}$$

This gives more freedom to choose m_2. If $m_2 = 10$, $n = 10^3$, $N = 10^8$, then $b = 0.9999900999$.

Now

$$p_0 = n/N = 0.00001 \quad \text{as before}$$

and

$$p_1 = (1 - p_0)[1 - (1 - n/N)b^{n-1}] = 0.001$$

$$p_2 = (1 - p_0)(1 - p_1)p_2'$$

where

p_2' = Prob(B knows at least one of the n' friends of A's friends)

$$= 1 - (1 - n/N)b^{n'-1}$$

$$n' = \binom{n}{1} m_1 - \binom{n}{2} m_2 + \binom{n}{3} m_3 - \binom{n}{4} m_4 + \ldots \pm \binom{n}{n} m_n \quad \text{as before}$$

To estimate m_k we need $\text{Prob}(K_1, \ldots, K_k)$, the probability of B being known to k randomly chosen people, and we shall assume this to be $\text{Prob}(K_1) \cdot c^{k-1}$, where $c = \text{Prob}(K_k | K_{k-1})$. If $k = 2$, then

$$\text{Prob}(K_1, K_2) = m_2/N = \text{Prob}(K_1) \cdot c = (n/N)c$$

so that $c = m_2/n$. Hence,

$$m_k = N \cdot (n/N)\,(m_2/n)^{k-1} = n(m_2/n)^{k-1} \qquad k = 1, 2, \ldots$$

Therefore,

$$n' = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} \cdot n(m_2/n)^{k-1}$$

$$= \frac{n}{m_2/n} \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} (m_2/n)^k$$

$$= (n^2/m_2)[1 - (1 - m_2/n)^n]$$

If $m_2 = 10$, then $c = 10/1000 = 0.01$ and $n' = 10^5(1 - e^{-10}) = 99996$. Thus

$$p_2' = 1 - (1 - 0.00001)(0.9999900999)^{99996}$$
$$\simeq 1 - (0.99999)(0.3716)$$
$$\simeq 0.6278$$

and

$$p_2 = (0.99999)(0.999)(0.6278) \simeq 0.627$$

To compute p_3 we shall need p'', the number of different people who are the friends of the acquaintances of the n people whom A knows.

$$n'' = \sum_{k=1}^{n'} (-1)^{k-1}\binom{n'}{k} n(m_2/n)^{k-1} = (n^2/m_2)[1 - (1 - m_2/n)^{n'}]$$

$$\simeq (10^6/10)[1 - (1 - 10^{-2})^{10^5}] \simeq 10^5(1 - e^{-1000}) \simeq 10^5 \simeq n'$$

$$p_3' = 1 - (1 - n/N)b^{n''-1} = 1 - (1 - 10^{-5})(0.9999900999)^{10^5} \simeq 0.6278$$

$$p_3 = (1 - p_0)(1 - p_1)(1 - p_2)p_3' \simeq (0.999)(0.373)(0.6278) = 0.234$$

This calculation leads to more plausible results, but it still does not have an underlying rationale to warrant attempts to fit data.

Contact probabilities in the presence of social strata

In a model of acquaintanceship structure it is desirable to be able to characterize persons as belonging to subsets in the population which can be interpreted as social strata. We show how the distribution for the length of minimal contact chains can be computed when strata are introduced. We begin by partitioning the entire population into r strata, with the ith stratum containing m_i members. Let h_{ij} denote the mean number of acquaintances which a person who is in stratum i has in stratum j. The mean number of acquaintances of a person in stratum i is then $n_i = \Sigma_{j=1}^{r} h_{ij}$. The conditional probability p_{ij} that a person picked at random in stratum j is known to someone in stratum i, given j, is $h_{ij}/m_j = p_{ij}$. The $r \times r$ matrix (p_{ij}) is symmetric and doubly stochastic because we have assumed that the "knowing" relation is symmetric.

We now select two people, A and C, with A in stratum i and C in stratum j. To obtain the probability that there is no 2-link contact chain from A to C, with the intermediary being in a specified stratum k, let K_i be the set of A's h_{ik} friends in stratum k. Combinatorially, Prob$\{K_i \cap K_j = \phi\}$ is the number of ways of selecting h_{ik} and h_{jk} out of m_k elements such that $K_i \cap K_j = \phi$,

divided by the total number of ways of selecting h_{ik}, h_{jk} out of m_k elements, assuming independent trials without replacement. Thus,

$$\text{Prob}\{K_i \cap K_j = \phi\} = \frac{m_k!/[h_{ik}!h_{jk}!(m_k - h_{ik} - h_{jk})!]}{\binom{m_k}{h_{ik}}\binom{m_k}{h_{jk}}}$$

$$= \frac{(m_k - h_{ik})!(m_k - h_{jk})!}{m_k!(m_k - h_{ik} - h_{jk})!} \tag{3}$$

The probability that there is no chain from A in stratum i to C in stratum j *via* some mutual acquaintance in any stratum is

$$\prod_{k=1}^{r} \frac{(m_k - h_{ik})!(m_k - h_{jk})!}{m_k!(m_k - h_{ik} - h_{jk})!} \equiv q_{ij}'$$

While data about all the elements of (h_{ij}) are not likely to be readily obtainable, the variables m_i, n_i and h_{ii} for $i = 1, \ldots, r$ may be estimable. We now make a methodological simplification and assume these variables equal for all i,-with $m_i = m = N/r$, $n_i = n$, $h_{ii} = h$ and

$$h_{ij} = \frac{n - h}{r - 1} = h' \text{ for all } i \neq j \tag{4}$$

To compute q_1', the probability that there is no chain of length 1 – or that there is *no* mutual acquaintance – between two individuals A and C, it is necessary to consider two cases:

(1) that in which A and C are in the same stratum;

(2) that in which A and C are in different strata.

In the first case, $q_1' = uv^{r-1} \equiv q_1'(1)$ (the number in parentheses refers to case 1), where u is the probability that B, the intermediary between A and C, fails to be in the same stratum as A and C, and v is the probability that he fails to be in a different stratum. Using eqn. (3), it is readily seen that

$$u = \frac{(m - h)!^2}{m!(m - 2h)!} \tag{5}$$

$$v = \frac{(m - h')!^2}{m!(m - 2h')!} \tag{6}$$

By similar reasoning,

$$q_1'(2) = w^2 v^{r-2}$$

where w is the probability that the stratum of B is the same as that of A but not of C; this is equal to the probability that the stratum of B is the same as that of C but not of A. This is, by eqn. (3),

$$w = \frac{(m - h)!(m - h')!}{m!(m - h - h')!} \tag{7}$$

With the help of Stirling's formula and series expansions we can derive a useful approximation for w. It is

$$w \simeq (1 + hh'/m^2)e^{-hh'/m} \simeq e^{-hh'/m} \tag{8}$$

As before, let p_1 denote the probability that A and C do not know each other, but that they have at least one common acquaintance. Then

$$p_1(i) \simeq (1 - p_0)[1 - q_1'(i)] \qquad i = 1, 2$$

To estimate p_1, we could take a weighted average,

$$p_1 = (1/r)p_1(1) + (1 - 1/r)p_1(2)$$

The above relation is written as an approximation, because $q_1'(i)$ is not a conditional probability given that A and C do not know each other, but the error is negligible. The number in the parentheses, 1 or 2, refers to whether or not A and C are in the same stratum, respectively. Thus,

$$p_1(1) \simeq (1 - n/N)(1 - uv^{r-1})$$

Because u can also be approximated by $\exp(-h^2/m)$ and v by $\exp(-h'^2/m)$, we can approximate $p_1(1)$ by

$$1 - \exp[-(h^2/m) - (h'^2/m)(r - 1)]$$

Substituting $m = N/r$, this becomes

$$p_1(1) \simeq 1 - \exp\{-(r/N)[h^2 + h'^2(r - 1)]\} \tag{9}$$

If A has more friends in a given stratum not his own than he has in his own stratum, then $h' > h$. If almost all of A's friends are in his own stratum, then $h' \ll h$, and $h \simeq n$. If r is large enough, $p_1(1)$ can be very close to 1. For instance, if $N = 10^8$, $h = 100$, $n = 1000$ and $r = 10$, we have that $h' = 900/9 = 100$, and $p_1(1) \simeq 0.00995$, as in the case of independence.

Next,

$$p_1(2) \simeq (1 - n/N)(1 - w^2v^{r-2})$$

$$\simeq 1 - \exp\{-(2/N)[2hh' + (r - 2)h'^2]\} \tag{10}$$

For the same numerical values as above,

$$p_1(2) \simeq 1 - e^{-10^{-2}} \simeq 0.00995 \text{ also}$$

We now wish to compute p_2^*, the joint probability that A and C do not know each other, *and* that they have no common friends, *and* that A has some friends, at least one of whom knows some friend of A. As before, we shall compute the conditional probability that A has some friends, at least one of whom knows some friend of C, given that A and C neither know each other, nor have a common acquaintance. We shall denote this conditional probability by $p_2'^*$, so that $p_2^* = (1 - p_0)(1 - p_1'^*)p_2'^*$. To say that A has some friends, at least one of whom knows some friend of C, is to say that there is at least one person, B, who knows A *and* who has at least one friend,

D, in common with C. By the assumed symmetry of the knowing relation, this is the same as saying: there exists $B \in K_C$, where K_C is the set of all people who can be linked to C by a minimal chain of length 1 (one intermediary). Select B at random and consider the choice fixed. Prob($B \in \bar{K}_C$) = $1 - p_1^*$, averaged over all strata. Assuming independence, the probability that any n B's, and in particular the n friends of A, all fail to be connected to C by a minimal chain of length 1 is $(1 - p_1^*)^n$. Hence, neglecting a small correction due to the condition in the definition of $p_1'^*$, we can estimate:

$$p_2'^* = 1 - (1 - p_1^*)^n \simeq 1 - \exp(-p_1^* n)$$

for p_1^* very small.

To obtain a more precise estimate of $p_2'^*$ we proceed as follows. Let s(A) denote the stratum of A. Consider first the case $i = 1$, where s(A) = s(C). Now suppose that s(B) = s(A). Then the probability that no chain of length 1 links B and C is uv^{r-1} as before. If s(B) $\neq$ s(A), however, B can be in any one of $r - 1$ strata, and for each stratum the probability that no chain of length 1 links B and C is $w^2 v^{r-2}$. Hence the probability that no chain of length 2 links A and C with s(A) = s(C) is

$$\begin{aligned} q_2'(1) &= u^h v^{h(r-1)} (w^2 v^{r-2})^{(r-1)h'} \\ &= u^h v^{(r-1)h + (r-1)(r-2)h'} w^{2(r-1)h'} \end{aligned} \tag{11}$$

Consider next the case $i = 2$, where s(A) $\neq$ s(C). If s(B) = s(A), the probability that no chain of length 1 links B and C is $(w^2 v^{r-2})^h$. If s(B) $\neq$ s(A), this probability is the product of:

(a) the probability of no 1-chain linking B and C when s(B) = s(C) – this is $(uv^{r-1})^{h'}$; and

(b) the same probability when s(B) $\neq$ s(C), *i.e.* $(w^2 v^{r-2})^{(r-2)h'}$. Hence, the probability that no chain of length 2 links A and C when s(A) $\neq$ s(C) is

$$\begin{aligned} q_2'(2) &= (w^2 v^{r-2})^h (uv^{r-1})^{h'} (w^2 v^{r-2})^{(r-2)h'} \\ &= u^{h'} v^{h(r-2) + h'(r-1) + (r-2)^2 h'} w^{2h + 2(r-2)h'} \\ &= u^{h'} v^{h(r-2) + h'(r^2 - 3r + 3)} w^{2[h + (r-2)h']} \end{aligned} \tag{12}$$

As before, we may estimate the conditional probability that A and C are linked by at least one 2-chain given that A does not know C or any friend of C by

$$\begin{aligned} 1 - p_2'^* = q_2'^* &= (1/r) u^h v^{(r-1)h + (r-1)(r-2)h'} w^{2(r-1)h'} + \\ &+ (1 - 1/r) u^{h'} v^{h(r-2) - h'(r^2 - 3r + 3)} w^{2[h - (r-2)h']} \end{aligned}$$

Note that effects due to the two conditions have been neglected and that independence has been assumed throughout.

Observe also that we could have written

$$q_2'(1) = [q_1'(1)]^h [q_1'(2)]^{h'(r-1)}$$

$$q_2'(2) = [q_1'(1)]^{h'} [q_1(2)]^h [q_2(2)]^{h'(r-2)}$$

$$q_2'^* = (1/r)q_2'(1) + (1 - 1/r)q_2'(2)$$

The above relation suggests a recursive scheme of generalizing the calculation. That is:

$$p_k = (1 - p_0)(1 - p_1'^*)(1 - p_2'^*) \ldots (1 - p_{k-1}'^*)(1 - q_k'^*)$$

$$q_k'^* = (1/r)q_k'(1) + (1 - 1/r)q_k'(2)$$

$$q_k'(1) = [q_{k-1}'(1)]^h [q_{k-1}'(2)]^{h'(r-1)}$$

$$q_k'(2) = [q_{k-1}'(1)]^{h'} [q_{k-1}'(2)]^{h+h'(r-2)} \qquad k = 2, 3, 4, \ldots$$

Using the cruder method suggested in the first paragraph of the above section,

$$p_k'^* = 1 - (1 - p_{k-1}'^*)^n \qquad k = 2, 3, 4, \ldots$$

There is another iterative method that could be used to compute p_k^*. If k is odd (*e.g.*, $k = 3$), compute $q_k'(1)$ and $q_k(2)$ using formulas (9) and (10) but substituting $p_{k-1}'(1)m$ for h and $p_{k-1}'(2)m$ for h'. Similarly, if k is even, use formulas (11) and (12) with the same substitutions for h and h'.

In the Appendix we develop further approximations to facilitate the calculation of p_0^*, p_1^*, and p_2^*, which we find to be 0.00001, 0.00759, and 0.9924, respectively, with the parameters used previously.

Note the departure from the model without strata is not very great. That is a significant inference. Structuring of the population may have a substantial effect on p_1. (It has no effect, of course, on p_0.) However, in a connected graph (which we believe any society must be) the nuclei get bridged by the longer chains quite effectively, and so the mean length of chains between randomly chosen pairs is only modestly affected by the structuring. We would therefore conjecture that, despite the effects of structure, the modal number of intermediaries in the minimum chain between pairs of Americans chosen at random is 2. We noted above that in an unstructured population with $n \simeq 1000$ it is practically certain that any two individuals can contact one another by means of at least two intermediaries. In a structured population it is less likely, but still seems probable. And perhaps for the whole world's population probably only one more bridging individual should be needed.

Monte-Carlo simulation models

To achieve greater understanding of the structural aspects of acquaintance nets, we approached an explanation of the dynamics of how acquaintance

bonds are formed with the help of a stochastic model that was simulated by computer. We regarded each individual to be located as a point in a social space, which we regarded as a square region in the two-dimensional Euclidean plane, to start with. As before, we let N be the number of individuals. Each individual can change his position in time t to time $t + 1$ by $(\Delta x, \Delta y)$ where

$$\Delta x = \begin{cases} s & \text{with probability } p \\ -s & \text{with probability } q \\ 0 & \text{with probability } r \end{cases} \quad \text{where } p + q + r = 1$$

and with Δy defined similarly, and statistically independent of Δx. Each individual is confined to remain in a $D \times D$ square, so that if his location at t is $\mathbf{z}(t) = [x(t), y(t)]$, then in the next simulation cycle it is

$$[x(t) + \Delta x \bmod D,\ y(t) + \Delta y \bmod D] = [x(t + 1), y(t + 1)]$$

We now define e_{AB} to be 1 if the line connecting $[x_A(t), y_A(t)]$ and $[x_A(t + 1), y_A(t + 1)]$ intersects the line from $[x_B(t), y_B(t)]$ to $[x_B(t + 1), y_B(t + 1)]$, and $e_{AB} = 0$ if these paths do not intersect. The event E_{AB} corresponding to $e_{AB}(t) = 1$ at time t is interpreted as a contact between A and B on day t. $(1/t)\Sigma_{\tau=1}^{t} e_{AB}(\tau)$ denotes the frequency with which A and B have met during the first t days.

Next, let $K_A(t)$ be the set of all people whom A has met by day t, or $\{\text{all B}: e_{AB}(\tau) = 1 \text{ for } \tau \leqslant t\}$. We now extend $K_A(t)$ to include A and define the center of that group or cohort on day t as follows:

$$\mathbf{c}_A(t) = [\bar{x}_A(t), \bar{y}_A(t)]$$

with

$$\bar{x}_A(t) = \frac{x_A(t) + \sum_{B \in K_A(t)} x_B(t) \sum_{\tau=1}^{t} e_{AB}(\tau)}{1 + \sum_{B} \sum_{\tau} e_{AB}(\tau)}$$

and $\bar{y}_A(t)$ is similarly defined. The x-coordinate of the center is the average of the x-coordinates of A and all the people he has met, weighted by how frequently they were contacted.

The probabilities p and q also vary with time and with each individual, as follows with $\mathbf{z}_A(t) = (x_A(t), y_A(t))$.

$$\text{If } \mathbf{c}_A(t) > \mathbf{z}_A(t), \text{ then } \begin{aligned} p_A(t + 1) &= p_A(t) + e \\ q_A(t + 1) &= q_A(t) - e/2 \\ r_A(t + 1) &= r_A(t) - e/2 \end{aligned}$$

$$\text{If } \mathbf{c}_A(t) < \mathbf{z}_A(t), \text{ then } \begin{aligned} p_A(t + 1) &= p_A(t) - e/2 \\ q_A(t + 1) &= q_A(t) + e \\ r_A(t + 1) &= r_A(t) - e/2 \end{aligned}$$

If $\mathbf{c}_A(t) = \mathbf{z}_A(t)$, then the probabilities do not change. Initially, $[p(0), q(0),$

$r(0)] = (1/3, 1/3, 1/3)$ and no probability must ever fall outside $[\delta, 1 - \delta]$ to ensure that the system remains stochastic; when these values are reached, the probabilities stay there until the z's and c's change.

After considerable experimentation with several values of the different parameters, we chose:

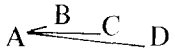
Number of individuals	$N = 225$
Size of one side of square grid	$D = 15$
Social responsiveness or elasticity	$e = 0.2$
Lower bound on probability of position change	$\delta = 0.01$
Unit increment in position change	$s = 1$

Well before the 10th iteration, clustering begins and by the 20th iteration it clusters into a single group. For realism, we would expect several clusters to emerge (corresponding to social strata) that exhibit both local and global structure, which are not too rigidly determined by the Euclidean structure of the social space. We have not explored the model sufficiently to determine if it has these properties, if small changes in the model could provide it with these properties, or if this approach should be abandoned. Computation cost increases as N^2 and the number of iterations, and took a few minutes per iteration on the MIT 370-186 system in 1973. This cost could be reduced by sampling, resulting in a fractional decrease that is the sample size divided by N. After enough iterations have produced what appears to be a realistic but scaled-down acquaintance net in such an idealized social space, a second program (also written by Diek Kruyt) to compute the distribution of chain lengths is then applied. Its cost varies as N^3.

Our present decision – held since 1975 – is to explore the use of a computer program that constructs an acquaintance net according to a simulation that uses the data we obtained from the 100-day diaries kept by our 27 respondents (see § 2). The basic inputs to this program are:

The total number of individuals	$N = 1000$
The number of people seen by person A on any f days in 100	$Y(f) =$ data
The number of different people that A did not see in 100 days	$Y_A(0)$
The number of people, each of whom has exactly k acquaintances in common with A	$M_A(k) =$ data

Outputs include the distribution of chain length. The program starts by selecting A and linking to him all the $Y(100)$ people he sees daily (chosen at random from the $N - 1$ in the program). This might, for example, be the nucleus of his circle of acquaintances consisting of $Y(100) = 3$ people. Call them B, C, and D, and we have



so far.

Next proceed with the first of A's friends just chosen, say B. Link to him all $Y(100)$ others chosen randomly from $N - 1$, but including A. This might generate the following list of B's friends: A, C, F. Repeat this for all people labeled so far, *e.g.*, C, D, F, *etc.*, until there are no more new "target" people. Then repeat this procedure for $Y(99)$ in place of $Y(100)$, but eliminating certain randomly chosen links if they do not satisfy the following constraint.

Our data suggested that there are fewer people who have one acquaintance in common with A than there are who have two acquaintances in common with A, *etc.*, but that only a few people have very many acquaintances in common with A. Thus, there is a value, M, for which $M(k)$ is greatest, where $M = M(K)$. For example, if $M(1) = 2$, $M(2) = 3$, $M(4) = 5$, $M(5) = 4$, *etc.*, then $M = 5$, $K = 4$. We must ensure that M people among those chosen so far each have K acquaintances in common with A, also with the people he sees daily. We then repeat these steps with $Y(98)$ in place of $Y(99)$ and replace the constraint that M friends have K acquaintances in common with A, *etc.*, by one requiring that $M(K - 1)$ people have $K - 1$ acquaintances in common with A, B, *etc.* This is continued until $Y(0)$ and $M(1)$ replace $Y(1)$ and $M(2)$, respectively.

Effective and efficient algorithms for making these selections subject to the given constraints have yet to be developed. The computational complexity of this algorithm must also be determined, and hopefully is a polynomial in N. Hopefully also, such a program can be run for N large enough so that distribution of chain length does not change significantly as N is increased. Fruitful next steps seem to us to be the further development and analysis of the models sketched in this section. When these are found to have properties we consider realistic for large social contact nets and are the result of plausible explanatory inferences, then some difficult problems of statistical estimation must be solved. Hopefully, then we will have reached some understanding of contact nets that we have been seeking.

Appendix

Some approximations using Stirling's formula have already been derived and analyzed.

There is another very useful approximation based on a slightly different model in the general case.

Let q'_{ij} be defined as in eqn. (2), but rewrite it as

$$\frac{(m_k - h_{ik})!(m_k - h_{jk})!}{m_k!(m_k - h_{ik} - h_{jk})!} =$$

$$\frac{(m_k - h_{ik})!(m_k - h_{jk})(m_k - h_{jk} - 1)...(m_k - h_{jk} - h_{ik} + 1)(m_k - h_{jk} - h_{ik})!}{m_k(m_k - 1)...(m_k - h_{ik} + 1)(m_k - h_k)!(m_k - h_{jk} - h_{ik})!}$$

$$= \left(1 - \frac{h_{jk}}{m_k}\right)\left(1 - \frac{h_{jk}}{m_k - 1}\right)\left(1 - \frac{h_{jk}}{m_k - 2}\right) \cdots \left(1 - \frac{h_{jk}}{m_k - h_{ik} + 1}\right)$$

(h_{ik} terms)

It is easily seen that this represents the probability of failing to draw a sample of h_{ik} red balls from an urn having m_k balls of which h_{jk} are red, but sampling without replacement. If we sample with replacement, the above formula becomes $q_{jk}^{h_{ik}}$, where $q_{jk} = (1 - h_{jk}/m_k)$. This represents the probability that none of A's h_{ik} friends in stratum k is known to C (s(A) = i, s(C) = j), where it is possible to count the same friend more than once. The fractional error committed by this assumption is

$$\epsilon = \left[\prod_{k=1}^{r} q_{jk}^{h_{ik}} - \prod_{k=1}^{r} \prod_{l=0}^{h_{ik}-1} \left(1 - \frac{h_{jk}}{m_k - l}\right)\right] \Bigg/ \prod_{k=1}^{r} q_{jk}^{h_{ik}}$$

This will be estimated later. Now,

$$\log q'_{ij} \simeq \sum_{k=1}^{r} h_{ik} \log \left(1 - \frac{h_{jk}}{m_k}\right)$$

If $h_{jk} \ll m_k$ for all k, we can further approximate this by

$$-\sum_{k=1}^{r} h_{ik} \frac{h_{jk}}{m_k} = -\sum_{k=1}^{r} h_{ik} \frac{h_{kj}}{m_j} = -\frac{1}{m_j} \sum_{k=1}^{r} h_{ik} h_{kj}$$

with a fractional error of about $h_{jk}/2m_k$, which is less than $(h + h')^2/2m$, as in the previous approximation. Furthermore, this approximation permits matrix multiplication and greater generality than only two values of h_{ij}. If we denote the matrix (h_{ij}) by **H** and $(\log q_{ij})$ by **L**, then **L** = **H**$\bar{\mathbf{H}}$, $\bar{\mathbf{H}}$ being the transpose of **H**.

To estimate the error, we take

$$\epsilon = 1 - \prod_{k=1}^{r} \prod_{l=0}^{h_{ik}-1} \left[\frac{1 - h_{jk}/(m_k - l)}{1 - h_{jk}/m_k}\right]$$

The term in brackets is approximated by the series

$$\left(1 + \frac{h_{jk}}{m_k} + \frac{h_{jk}^2}{m_k^2} + \cdots\right) - \frac{h_{jk}}{m_k - l}\left(1 + \frac{h_{jk}}{m_k} + \cdots\right)$$

$$= 1 + h_{jk}\left(\frac{1}{m_k} - \frac{1}{m_k - l}\right) + \frac{h_{jk}^2}{m_k}\left(\frac{1}{m_k} - \frac{1}{m_k - l}\right) + \cdots$$

$$= 1 - \frac{l}{m_k(m_k - l)}\left(h_{jk} + \frac{h_{jk}^2}{m_k} + \ldots\right)$$

$$\simeq 1 - \frac{h_{jk}}{m_k^2}\, l$$

$$\epsilon \simeq 1 - \prod_{k=1}^{r} \exp\left(\frac{-h_{jk}}{m_k^2} \sum_{l=0}^{h_{ik}-1} l\right)$$

$$= 1 - \prod_{k=1}^{r} \exp\left[\frac{-h_{jk}}{m_k^2} \frac{(h_{ik} - 1)h_{ik}}{2}\right]$$

$$\simeq 1 - \prod_{k=1}^{r} \exp\left(-\frac{h_{jk}h_{ik}^2}{2m_k^2}\right)$$

$$= 1 - \exp\left(-\frac{1}{2m_j^2} \sum_{k=1}^{r} h_{ik}^2 h_{kj}\right)$$

According to this estimate, the approximation is good only when

$$\sum_{k=1}^{r} h_{ik}^2 h_{kj} < m_j^2$$

To compare this with the exponential approximation, let $h_{ik} = h$ if $i = k$, h' if $i \neq k$, so that

$$\sum_k h_{ik}^2 h_{kj} = h^2 h' + hh'^2 (r - 2)h'^3 \qquad i \neq j$$

$$= h^3 + (r - 1)h'^3 \qquad i = j$$

Hence, it would be required that $(h + h')^3 r < m^2$ or $(h + h')^{3/2} \sqrt{r} < m$, compared with $(h + h')^2 < m$.

For the above simplified situation, the replacement model gives

$$q'_{ii} \simeq \exp\left[\frac{-h^2 + (r - 1)h'^2}{m}\right]$$

$$q'_{ij} \simeq \exp\left[\frac{-2hh' + (r - 2)h'^2}{m}\right] \qquad i \neq j$$

As an example where the departure from the results obtained when stratification was disregarded becomes more pronounced than in the illustrations chosen so far, let $N = 10^8$, $n_i = n = 10^3$ for all i, $m_j = m = 10^4$ for all j, $r = 10^4$, $h_{ii} = h = 500$, $h_{ij} = h' = 500/(10^4 - 1) = 5 \times 10^{-2}$ for all $i \neq j$.

(1) $p_0^* = n/N = 10^{-5}$

(2) $p_1^* = (1 - p_0)p_1'^* = (1 - p_0)(1 - q_1'^*)$

$$q_1'^* = \frac{1}{r} q_1'(1) + \left(1 - \frac{1}{r}\right) q_1'(2)$$

$$q_1'(1) = \exp\left(-\frac{25 \times 10^4}{10^4} + \frac{25 \times 10^{-4}}{10^4} \times 10^4\right) \simeq e^{-25} \simeq 0$$

$$q_1'(2) \simeq \exp\left(-\frac{2 \times 500 \times 5 \times 10^{-2}}{10^4} + \frac{10^4 \times 25 \times 10^{-4}}{10^4}\right) \simeq 0.9925$$

$$q_1'^* \simeq 0.99241$$

$$p_1^* \simeq 0.00759$$

(3) Recall that $u \simeq \exp(-h^2/m)$, $v \simeq \exp(-h'^2/m)$, $w \simeq \exp(-hh'/m)$, so that

$$q_2'(1) = \exp\left(-\left[\frac{h^3}{m} + \frac{h'^2}{m}(r-1)h + (r-1)(r-2)h' + 2\frac{hh'^2}{m}(r-1)\right]\right)$$

$$= \exp\left(-\left[\frac{h^3}{m} + 3\frac{hh'^2}{m}(r-1) + \frac{h'^3}{m}(r-1)(r-2)\right]\right)$$

$$\simeq \exp\left(-\frac{1}{m}(h^3 + 3rhh'^2 + r^2h'^3)\right)$$

$$q_2'(2) = \exp\left(-\frac{1}{m}\{h^2h' + [h(r-2) + h'(r^2 - 3r + 3)]h'^2 + 2[h + (r-2)h']hh'\}\right)$$

$$\simeq \exp\left(-\frac{1}{m}[h^2h' + hrh'^2 + h'^3r^2 + 2h^2h' + 2(r-2)hh'^2]\right)$$

$$\simeq \exp\left(-\frac{1}{m}(3h^2h' + 3rhh'^2 + r^2h'^3)\right)$$

Then,

$$q_2'(1) = \exp[-10^{-4}(125 \times 10^6 + 3 \times 125 \times 10^{-2} \times 10^4 + 10^8 \times 125 \times 10^{-6})]$$

$$= \exp[-(12500 + 5)] \simeq 0$$

$$q_2'(2) = \exp[-10^{-4}(3 \times 125 \times 10^2 + 3 \times 10^4 \times 125 \times 10^{-2} + 10^8 \times 125 \times 10^{-6})]$$

$$= \exp(-8.75) \simeq 0.00016$$

Hence,

$$p_2^* \simeq (1 - 10^{-5})(1 - 0.00759)(1 - 0.00016)$$

$$\simeq 0.9924$$

References

Alba, R.
1973 "A graph-theoretic definition of a sociometric clique". *Journal of Mathematical Sociology 3*:113 - 126.

Alba, R. and C. Kadushin
1976 "The intersection of social circles: a new measure of social proximity in networks". *Sociological Methods and Research 5*:77 - 102.

Boissevain, J.
1974 *Friends of Friends: Networks, Manipulators, and Coalitions.* New York: St. Martin's Press.

Boorman, S. and H. White
1976 "Social structures from multiple networks". *American Journal of Sociology 81*:1384 - 1446.

Breiger, R., S. Boorman and P. Arabie
1975 "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling". *Journal of Mathematical Psychology 12*: 328 - 383.

de Grazia, A.
1952 *Elements of Political Science.* New York: Free Press.

Deutsch, K.
1956 "Shifts in the balance of communication flows". *Public Opinion Quarterly 20*:143 - 160.
1966 *Nationalism and Social Communication.* Cambridge, Mass.: MIT Press.

Doreian, P.
1974 "On the connectivity of social networks". *Journal of Mathematical Sociology 3*:245 - 258.

Erickson, B. and P. Kringas
1975 "The small world of politics, or, seeking élites from the bottom up". *Canadian Review of Sociology and Anthropology 12*:585 - 593.

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

Foster, C. and W. Horvath
1971 "A study of a large sociogram III: reciprocal choice probabilities as a measure of social distance". *Behavioral Science 16*:429 - 435.

Foster, C., A. Rapoport and C. Orwant
1963 A study of large sociogram II: elimination of free parameters". *Behavioral Science 8*:56 - 65.

Granovetter, M.
1974 *Getting a Job: A Study of Contacts and Careers.* Cambridge, Mass.: Harvard University Press.
1976 "Network sampling: some first steps". *American Journal of Sociology 81*:1287 - 1303.
Griffith, B., V. Maier and A. Miller
1973 *Describing Communications Networks Through the Use of Matrix-Based Measures.* Unpublished. Drexel University, Graduate School of Library Science, Philadelphia, Pa.
Gurevich, M.
1961 *The Social Structure of Acquaintanceship Networks.* Cambridge, Mass.: MIT Press.
Gurevich, M. and A. Weingrod
1976 "Who knows whom – contact networks in Israeli National élite". *Megamot 22*:357 - 378.
n.d. *Human Organization.* To be published.
Hallinan, M. and D. Felmlee
1975 "An analysis of intransitivity in sociometric data". *Sociometry 38*:195 - 212.
Hammer, M.
n.d. *Social Access and Clustering of Personal Connections.* Unpublished.
Holland, P. and S. Leinhardt
1970 "A method for detecting structure in sociometric data". *American Journal of Sociology 70*:492 - 513.
Horowitz, A.
1977 "Social networks and pathways to psychiatric treatment". *Social Forces 56*:81 - 105.
Hunter, J. and R. L. Shotland
1974 "Treating data collected by the small world method as a Markov process". *Social Forces 52*:321 - 332.
Jacobson, D.
1970 "Network analysis in East Africa; the social organization of urban transients". *Canadian Review of Sociology and Anthropology 7*:281 - 286.
Jennings, H.
1937 "Structure of leadership – development and sphere of influence". *Sociometry 1*:131.
Katz, E. and P. Lazarsfeld
1955 *Personal Influence.* Glencoe, Ill.: Free Press.
Killworth, P. and B. Russell
1976 "Information accuracy in social network data". *Human Organization 35*:269 - 286.
Kleinrock, L.
1964 *Communication Nets: Stochastic Message Flow and Delay*, New York: McGraw-Hill.
Korte, C. and S. Milgram
1970 "Acquaintanceship networks between racial groups: application of the small world method". *Journal of Personality and Social Psychology 15*:101 - 108.
Kurtzman, D. H.
1935 *Methods of Controlling Votes in Philadelphia.* Philadelphia: University of Pennsylvania.
Lorrain, F.
1976 *Social Networks and Classification.* Manuscript.
Lorrain, F. and H. White
1971 "Structural equivalence of individuals in social networks". *Journal of Mathematical Sociology 1*:49 - 80.
Luce, R.
1950 "Connectivity and generalized cliques in sociometric group structure". *Psychometrika 15*: 169 - 190.
Lundberg, C.
1975 "Patterns of acquaintanceship in society and complex organization: a comparative study of the small world problem". *Pacific Sociological Review 18*:206 - 222.
McKinlay, J.
1973 "Social networks, lay consultation and help-seeking behavior". *Social Forces 51*:275 - 292.
McLaughlin, E.
1975 "The power network in Phoenix. An application of the smallest space analysis". *The Insurgent Sociologist 5*:185 - 195.
Milgram, S.
1967 "The small world problem". *Psychology Today 22*:61 - 67.
Miller, G.
1956 "The magical number seven plus or minus two". *Psychological Review 63*:81 - 97.

Mitchell, J. C. (Ed.)
1969 *Social Networks in Urban Situations – Analysis of Personal Relationships in Central African Towns.* Manchester: University Press.
Newcomb, T.
1961 *The Acquaintance Process.* New York: Holt, Rinehart, and Winston.
Nutini, H. and D. White
1977 "Community variations and network structure in social functions of Compradrazgo in rural Tlaxcala, Mexico". *Ethnology 16*:353 - 384.
Peay, E.
1976 "A note concerning the connectivity of social networks". *Journal of Mathematical Sociology 4*:319 - 321.
Rapoport, A.
1963 "Mathematical models of social interaction". *Handbook of Mathematical Psychology.* New York: Wiley, pp. 493 - 579.
Rapoport, A. and W. Horvath
1961 "A study of a large sociogram". *Behavioral Science 6*:279 - 291.
Rosenthal, H.
1960 *Acquaintances and Contacts of Franklin Roosevelt.* Unpublished B.S. thesis: MIT.
Saunders, J. and N. Reppucci
1977 "Learning networks among administrators of human service institutions". *American Journal of Community Psychology 5*:269 - 276.
Schulman, N.
1976 "Role differentiation in urban networks". *Sociological Focus 9*:149 - 158.
Travers, J. and S. Milgram
1969 "An experimental study of the small world problem". *Sociometry 32*:425 - 443.
Warner, W. L.
1963 *Yankee City.* New Haven: Yale University Press.
Wasserman, S.
1977 "Random directed graph distributions and the triad census in social networks". *Journal of Mathematical Sociology 5*:61 - 86.
White, H.
1970a "Search parameters for the small world problem". *Social Forces 49*:259 - 264.
1970b *Chains of Opportunity.* Cambridge, Mass.: Harvard University Press.
1973 "Everyday life in stochastic networks". *Sociological Inquiry 43*:43 - 49.
Wolfe, A.
1970 "On structural comparisons of networks". *Canadian Review of Sociology and Anthropology 7*:226 - 244.