Interconnectedness and the Duration of Connections in Several Small Networks

DR. JOSEPH ZUBIN
BIOMETRICS RESEARCH
722 WEST 135 STREET
NEW YORK, N. Y. 10023

One aspect of a theory utilizing the concept of social network
is examined empirically for five small networks, with an emphasis
on hypothesized changes over time. An attempt is made to relate
the distribution of the connections in these networks, in terms
of a measure of distance, to their durations.

Introduction

The work described below is a preliminary test of one aspect of a social
theory against some relevant empirical material. The approach is based on
adopting a limited view of society as networks of connections among biological
points. Channels of connection may involve work, goods, information, or "pure"
social exchanges.

Social analysis in terms of networks has been used increasingly in the
past decade (for reviews, see Mitchell 1969, Bott 1971 and Whitten and Wolfe
1973) partly because it entails fewer cultural assumptions than does analysis
in terms of more specific group concepts like clan or class. It is both
descriptively convenient and amenable to interesting mathematical manipulation.
It is suggested here that in addition to these methodological advantages,
elementary laws of connection for networks of biosocial points may be posited,
which are potentially of broad theoretical relevance. The theory should be
applicable to populations of any size, although it will require much more
complex development for large populations than for the very small networks we
have studied.

It is posited that any connection that occurs between points in a social
system sets up a field of attraction which alters the probability of occurrence
of other connections. From this, in combination with certain other assumptions,
a wide range of theoretical socio-cultural propositions may be derived. Some
of these are apparently testable; others are for the time being purely
speculative.

The argument can briefly be summarized as follows:

The social realm may be conceived of in terms of points and connections
among them. If a connection exists between X and Y, it will tend to continue
and to increase in magnitude. If a connection also exists between Y and Z --
as well as X and Y -- there will be a tendency for the set to become
interconnected by the establishment of a connection between X and Z. For any
given set of conditions (involving terrain, technology, etc.) there are
limitations on the number of a given kind of connection that an individual, or
set of individuals, is capable of maintaining. The basic contradiction between
the tendency to expand connections, combined with the limitations on the
capacities for such expansion, may be resolved in several different ways. In
the broad sense of the word, communication is involved in any connection, and
the modes of communication may play a critical role in how this contradiction
is resolved, with different consequences for the groups involved, particularly
in terms of the relative importance of direct and indirect connections.

From the perspective of analysis of society in terms of sets of
connections, the most profound social change in human evolution has been a
consequence of the increased significance of indirect connections. This
change is the shift in relative importance of fairly small, interconnected,
 quasi-closed social groups as against a complex of direct and indirect
connections comprising a number of levels of group size. The organization
of indirect connections (which must always rest on a base of direct
connections) follows different principles from the organization of direct
connections. Direct connections have simpler limitations on number, time,
and space, and clearer transmission patterns, including more immediate
validation procedures between different sources of information. In addition, complex organization of indirect connections increases the sources and types of change possible for sets of direct connections. For example, such organization increases the channels of entry into the sets of direct connections. Because of increased size, it also increases the time span involved in having the effects of any event diffuse through the system.

The basis of the theory briefly suggested here is that sets of connections among biological points tend to attract connections with neighboring points. Since the research is directed to only one part of this postulate, only this aspect of the theory will be considered further: that connections tend to attract; and that the closer and stronger they are, the more they will attract. This much is almost self-evident; whether it is trivial or fruitful will depend in part upon what kinds and degrees of regularity it may show; in part on what may be built upon it. The choice of network characteristics specifically examined here was dictated first by the conception that simple measures of connection in relation to distance are fundamental to any analysis of networks; and second, by the expectation that theoretical and empirical extension to large populations will require, in addition, some information about changes over time, including estimates of rates of formation and severance of connections. We would consider our findings here valuable if they were to form a part of such an extension.

**Empirical material**

The empirical material analysed consists of five small networks located in New York, London, and rural Vermont, involving from 8 to 24 men and women in each. The networks differ considerably in terms of age and sex composition, occupations, religion, and other commonly used sociological categories. The
subjects have an age range from the late teens to the sixties; they are Protestant, Catholic and Jewish, primarily white, and slightly more than half of them are working class. The data were derived from interviews with each of the individuals included in the analysis. The material relevant here consisted in the informants' statements about who the people were that they were close to, and how long they had known each of them.

The networks were constructed on a branching scheme as diagrammed in ideal form in Figure 1. An initial individual was arbitrarily selected and interviewed. Information from this interview led to the selection of two other individuals from among the several people the first informant was closest to; each of them led to two others, and so on.

INSERT FIGURE 1 ABOUT HERE

Definitions and analyses

The connections used to derive the network are referred to here as "base connections," to distinguish them from additional close relationships that also exist among pairs of individuals in the network, but were not built in by our selection procedure. These additional close relationships we have called "extra connections," as indicated in Figure 2 between B and C, E and I, N and O.

INSERT FIGURE 2 ABOUT HERE

"Base distances" refer to the minimum number of steps necessary to complete a path along the base connections between any two points in the network. In Figure 2, A B is a base one-step or direct connection, A D is base two-step, A H is three, C H is four, and so on. It is important to note that although there is an actual direct connection -- an extra connection -- between
B and C, BC is a base 2-step connection, since BA and AC are both necessary base connections on the minimum base path between B and C. We have used the term "connection" only for one-step or direct ties. Figure 3 shows the diagrams for the five actual networks.

INSERT FIGURE 3 ABOUT HERE

To return now to the proposition under consideration: connections tend to attract each other, and in direct proportion to their magnitudes. The shorter the base distance between two members, the more likely it is that a direct connection will also be found to exist between them — i.e., an extra connection. In this form, the proposition needs little demonstration — it appears self-evident that two people with a friend in common are more likely to be friends than are two people chosen at random. However, it seems worthwhile to begin with this simple form, partly in order to give it some quantification (which is, of course, not self-evident), and partly to move towards a translation of the data into a temporal form.

Thus, the first specific hypothesis applied to the five empirical networks is that extra connections — those in addition to the base connections — are more likely to exist between pairs separated by shorter base distances. They are most likely between pairs at base 2-step distances, then base 3-step, and so on. This in fact turns out to be the case for all five networks: the proportion of actual extra connections to those that are structurally possible (i.e., all possible pairs with no base connection) is highest for base 2-step distances, diminishes sharply at base 3-step, and reaches zero in all these networks at base 4-step or base 5-step. In Table I, results for the five networks have been combined.
(The networks were, of course, analyzed separately, but the findings were consistent for all networks, and since the numbers in each are small, the result is clearer to see if the values are combined.)

**INSERT TABLE I ABOUT HERE**

The distributions of connections just considered are at a point in time, whereas the theoretical issue of interest here involves a development through time. To transform the hypothesis to project back in time, it is reformulated by making the assumption that all extra connections are at a 2-step distance just prior to becoming direct connections. We can then attempt to predict the distributions that were actually found by constructing connections drawn only from 2-step pairs, beginning with base 2-step pairs and continuing with pairs that become 2-step by the addition of the new connections. If we take the chain ABCD in Figure 4, the initial -- or base

**INSERT FIGURE 4 ABOUT HERE**

-- direct connections are AB, BC and CD; AC and BD are base 2-step pairs, and AD is base 3-step. If AC were to become a direct connection, base distances would be unchanged, but the resulting distance between A and D would be reduced to 2 steps, and the likelihood of AD becoming a direct connection would thus increase. Additionally, if BD as well as AC were to become direct, BD would become a 2-step distance by two independent routes, further increasing the probability of a direct connection. Specifically, the distribution of the number of extra connections to be expected as a function of the number of steps defining the base distances, may theoretically be obtained in the following manner: the first extra connection is randomly chosen from among the base 2-step pairs. This new connection provides a link
which may reduce the distance between other pairs in the network. Any newly
created 2-step pairs are added to the pool to be drawn from on the next
random choice, and the next extra connection is selected at random, with the
exception that any pair which becomes 2-step by more than one route will
automatically be considered an extra connection. This procedure is continued
until the same number of extra connections has been chosen as in fact are
found in the actual networks. Table II shows that the empirical distributions
of extra connections in the five networks are on the whole within the ranges
obtained from this theoretical procedure. (Theoretical ranges were
approximated by using three different sequences of random choices for each
network.)

INSEIT TABLE II ABOUT HERE

In this procedure, a hypothetical time dimension was utilized. The
original pool of pairs randomly drawn from for the first hypothetical extra
connection consisted only of base 2-step pairs. Thus, for the first extra
connection, the hypothetical proportion of all extra connections that are
base 2-step must be 1.00. The pool of 2-step pairs increases with the
formation of new extra connections; the pool of base 2-step pairs, however,
not only cannot increase, but in fact decreases as some of those pairs are
eliminated from the pool by already having been drawn as extra connections.
Thus, over hypothetical time, the proportion of extra connections that are
base 2-step pairs decreases. Since the distributions of extra connections
obtained by this theoretical procedure seem approximately to fit the
distributions of extra connections found empirically in the five networks,
the results suggest that this change in proportion may also occur in real
time. That is, the relative proportion of all extra connections that are
(original) base 2-step is highest when a network is relatively new, and decreases as the network "ages." The rank orders of this proportion and of real duration for the empirical networks do correspond to this inference. Table III shows this relationship, using as a measure of the "age" of the network (not hypothetically, but in real time) the mean duration in years of the base connections, as reported (reliably) by both parties to each relationship. (Other measures of duration, such as the geometric mean show the same pattern.)

If we then assume that the number of connections increases over time, and that the proportion of these that are base 2-step decreases over time, we can set up a family of curves for different rates of formation of connections for networks of different sizes. One such curve is developed below, on which we based an attempt to relate the distribution of connections to the actual relative durations of base connections for these networks.

For binary branching networks that are far from fully interconnected, the formation of one extra connection between a 2-step pair will on an average create two new 2-step pairs. (This varies with the size of the group, the departure from perfect binary branching, the stage of interconnectedness, and so on, but is a reasonable initial approximation.) If we assume, for convenience, a branching network with an initial population of 100 base 2-step pairs, and a rate of connection of 10% per unit time, the following theoretical table (Table IV) may be drawn:

The ratio of base 2-step connections to all extra connections over
hypothetical time based on the above assumptions (column f of Table IV) may be plotted as shown in Figure 5:

**INSERT FIGURE 5 ABOUT HERE**

Without knowing the unit of time relevant to the formation of a connection, we cannot "predict" back to the actual durations of base connections of the networks from their observed distribution of extra connections. However, since these networks are all similar in that they are unbounded networks of close connections, we tentatively assumed that they have similar time units, and predicted the actual relative durations of the five networks from the distributions of extra connections.

The actual proportion of base 2-step connections to all extra connections for each of the five networks was shown in Table 3. These values may be converted to hypothetical relative durations on the basis of Table IV or Figure 5.

Figure 6 then shows the relationship between the hypothetical relative durations which were predicted on the basis of the proportion of all observed extra connections that were base 2-step, and the actual durations, which were the arithmetic means of the durations of the base connections for each network. The solid diagonal is a hypothetical line showing where the points would fall if our prediction were perfect. The fit between the actual and the hypothetical seems very good.

**INSERT FIGURE 6 ABOUT HERE**

A qualification which is presumably obvious should nevertheless be made explicit. We are dealing here with only five networks, and these five are all quite small. Three of them have seven or fewer extra connections on
which to base the values used for estimating hypothetical time. With such low numbers, an error in our data on even one connection could alter these specific findings. In an attempt to cope with this problem, we first tried to expand the networks by including several peripheral individuals on whom we had auxiliary data for two of the networks. Our results remain substantially the same, but these additional network members add too few extra connections to provide a satisfactory solution. Our interpretation at this time is that the two networks with the largest number of extra connections are probably placed correctly relative to each other, but that the other three networks might, by error or chance variation, have shown a pattern somewhat different from what we obtained. As another limited check on one aspect of this problem, we utilized a differently defined set of base connections for one of the networks -- that is, we took the oldest set (longest actual duration) that would form a minimum branching network, rather than the set we had originally selected. Although the average base duration is of course somewhat larger than that originally obtained, there seems to be no substantial difference in the findings.

In addition to the need for working with larger sets of connections, two major assumptions involved in this analysis will require careful investigation in terms of additional data. It was assumed that all five networks had an identical rate of formation of new connections; and that any severance of connections that may have occurred was at a sufficiently low (or sufficiently regular) rate not to affect the results substantially. Our data are not adequate for any exploration of rates of severance of connections. We do, however, have data which permit initial consideration of some variation in rates of formation of connections. As may be seen from Figure 6, the predicted duration for the S-net is slightly high, indicating that the
assumed (identical) rate is a little too low, whereas the reverse discrepancy is found for the V-net and the P-net. Near identity of rates was a reasonable assumption because the networks were all constructed in the same way, as small open branching networks; quasi-closed groups should have different rates. Empirically, however, even these five networks show apparent differences in relative openness. To examine this as a possible source of the slight discrepancies between our predicted and actual durations, we developed a rough measure of this degree of openness.

The interviews elicited information about a number of social contacts in addition to those judged to be "close." On an average, each informant mentioned 29 other social contacts, or a total of about 2300 across all networks, about 15% of which were mentioned by at least two informants. This "name overlap" was used as an indication of the tendency of each network towards being a quasi-closed group. It was assumed that relatively closed groups would have more internal contact than more open groups, and therefore a faster rate of formation of new "close" connections.

Figure 7 shows the same five networks with the connections now defined in terms of a name overlap of at least two persons mentioned in common by the pair. (The first name in common was excluded since our selection procedure makes it virtually certain that all individuals in the network will give at least one name in common with at least one other network member.) For this Figure, the S-net, which has 24 members, was arbitrarily limited to the first 14 to make it the same size as the B, F, and V nets, which should make

\[ \text{INSERT FIGURE 7 ABOUT HERE} \]

it possible to see by inspection that the S-net has the greatest amount of overlap and the V-net the least, of these four. The P-net, with only 8
members, is less directly comparable, but seems also fairly low. These diagrams incidentally suggest that there are peripheral members who are probably more intensely connected into other networks than into the ones being studied. (With larger networks, this procedure might constitute a basis for defining relatively independent sub-networks.)

For each network, a measure of closedness was taken to be the average number of connections for each member included. Individuals were excluded who were not connected to the main set, or connected only by a single unsupported* connection. The resulting values are shown in Table V. If the average value is used as a standard, the direction of the discrepancy between the predicted durations and the actual durations (as plotted in Figure 6) is seen to correspond: S has a high rate of formation of connections and thus a longer predicted than actual duration, whereas V has a low rate and a shorter predicted than actual duration.

**Conclusion**

Tentatively then, analysis of these five networks supports the expectation that there is attraction among social connections, in the sense that direct connections are formed primarily between pairs at the least distance of an indirect route, and that this is systematic enough across similarly structured networks to permit the prediction of the distribution of connections among

---

*In the V-net in Figure 7, for example, 10 is connected with the main set by a single unsupported connection, whereas 01 and 03 support each other in their single connections through 02.*
points at differing periods of actual (though relative) time. We would also expect, although we cannot yet test it, that rates for differently structured networks should differ systematically.

On the basis of the assumption that social connections attract each other, and that they do so with a relative strength inversely related to their immediately prior social distance, it has been possible to predict for five empirical networks the distribution of whole sets of connections from defined sub-sets, and to estimate their durations.

On both theoretical and methodological grounds, we believe that the manipulations of connections at different step-distances, and their projection through hypothetical time, are warranted by both the specific material used and the more general properties of social networks. With respect to the projection to actual time, a number of additional parameters presumably become relevant, and although the "old" and "young" networks are likely to be properly placed, the nature of the data is such that we have less certainty about the relative placement of less extreme networks, particularly if they are very small. Larger numbers of larger networks of the same structural type should be used, and compared with other types of networks, for this kind of analysis. More important, direct rather than inferential data on changes over time are necessary.*

The analyses presented here are obviously crude, and deal with only one aspect of a larger theory. Nevertheless, we believe they give some support to the expectation that a theory at this level of abstraction can be used to

* We are currently collecting some longitudinal data on comparable networks.
generate propositions applicable with precision to the real social domain. The value of this kind of theory, in refined and developed form, would be its potential for dealing with a wide range of social phenomena in the same framework on the basis of invariant principles. No other kind of theory is capable of enabling us to predict new structures, for example, though intuitively one expects them to arise in consequence of present technological and population trends.
Notes

1 A shorter version of this paper was presented at the American

2 Biometrics Research Unit, New York State Department of Mental Hygiene,
and Columbia University.

3 Biometrics Research Unit, New York State Department of Mental Hygiene.
Currently at the Institute for Sociotherapy.
References Cited

Bott, E.
1971
Reconsiderations. In Family and Social Network.

Mitchell, J. C.
1969

Whitten, N. E. Jr.
and Wolfe, A. W.
1973
Table I

PROPORTION OF ACTUAL TO POSSIBLE EXTRA CONNECTIONS
BY BASE DISTANCE, FOR FIVE NETWORKS COMBINED

<table>
<thead>
<tr>
<th>BASE DISTANCE</th>
<th>POSSIBLE CONNECTIONS</th>
<th>ACTUAL EXTRA CONNECTIONS</th>
<th>PROPORTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 - step</td>
<td>86</td>
<td>28</td>
<td>.326</td>
</tr>
<tr>
<td>3 -</td>
<td>95</td>
<td>10</td>
<td>.105</td>
</tr>
<tr>
<td>4 -</td>
<td>95</td>
<td>3</td>
<td>.032</td>
</tr>
<tr>
<td>5 or more</td>
<td>232</td>
<td>0</td>
<td>-0-</td>
</tr>
</tbody>
</table>
Table II

ACTUAL AND HYPOTHETICAL EXTRA CONNECTIONS,

BY BASE DISTANCE, FOR FIVE NETWORKS

<table>
<thead>
<tr>
<th>DISTANCE</th>
<th>B net</th>
<th>F net</th>
<th>V net</th>
<th>S net</th>
<th>P net</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACTUAL EXTRA</td>
<td>EXPECTED</td>
<td>ACTUAL</td>
<td>EXPECTED</td>
<td>ACTUAL</td>
</tr>
<tr>
<td>2-step</td>
<td>8</td>
<td>6.5 - 9</td>
<td>4</td>
<td>4 - 5</td>
<td>5</td>
</tr>
<tr>
<td>3-</td>
<td>5</td>
<td>4 - 5.5</td>
<td>2</td>
<td>1 - 2</td>
<td>1</td>
</tr>
<tr>
<td>4-</td>
<td>1</td>
<td>1 - 2</td>
<td>0</td>
<td>0 - 1</td>
<td>1</td>
</tr>
<tr>
<td>5 or more</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### Table III

RELATIONSHIP OF PROPORTION OF EXTRA CONNECTIONS AT BASE 2-STEP TO DURATION OF BASE CONNECTIONS FOR FIVE NETWORKS

<table>
<thead>
<tr>
<th>NETWORKS</th>
<th>MEAN DURATION OF BASE CONNECTIONS (IN YEARS)</th>
<th>PROPORTION OF EXTRA CONNECTIONS THAT ARE BASE 2-STEP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Duration Rank Order</td>
<td>Proportion Rank Order</td>
</tr>
<tr>
<td>B</td>
<td>20.4 1</td>
<td>.57 5</td>
</tr>
<tr>
<td>F</td>
<td>14.8 2</td>
<td>.67 4</td>
</tr>
<tr>
<td>V</td>
<td>14.7 3</td>
<td>.71 3</td>
</tr>
<tr>
<td>S</td>
<td>9.4 4</td>
<td>.77 2</td>
</tr>
<tr>
<td>P</td>
<td>4.9 5</td>
<td>1.00 1</td>
</tr>
</tbody>
</table>
### Table IV

**THEORETICAL RATIOS OF BASE 2-STEP TO ALL EXTRA CONNECTIONS OVER TIME**

<table>
<thead>
<tr>
<th>Time Period</th>
<th>2-step Pairs</th>
<th>Extra Connections (Cumulative)</th>
<th>Ratio, Base 2 cx all extra Cx</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base (a)</td>
<td>Other (b)</td>
<td>Base 2 (c) Other (d)</td>
</tr>
<tr>
<td>t₀ - t₁</td>
<td>100</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>t₁ - t₂</td>
<td>90</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>t₂ - t₃</td>
<td>81</td>
<td>40</td>
<td>27.1</td>
</tr>
<tr>
<td>t₃ - t₄</td>
<td>72.9</td>
<td>60.2</td>
<td>34.4</td>
</tr>
<tr>
<td>t₄ - t₅</td>
<td>65.6</td>
<td>80.8</td>
<td>41.0</td>
</tr>
<tr>
<td>t₅ - t₆</td>
<td>59.0</td>
<td>102.1</td>
<td>46.9</td>
</tr>
<tr>
<td>t₆ - t₇</td>
<td>53.1</td>
<td>124.1</td>
<td>52.2</td>
</tr>
<tr>
<td>t₇ - t₈</td>
<td>47.8</td>
<td>147.1</td>
<td>57.0</td>
</tr>
<tr>
<td>t₈ - t₉</td>
<td>43.0</td>
<td>171.4</td>
<td>61.3</td>
</tr>
<tr>
<td>t₉ - t₁₀</td>
<td>38.7</td>
<td>197.1</td>
<td>65.2</td>
</tr>
<tr>
<td>t₁₀ - t₁₁</td>
<td>34.8</td>
<td>224.6</td>
<td>68.7</td>
</tr>
</tbody>
</table>
Table V

AVERAGE NUMBER OF CONNECTIONS BASED ON NAME OVERLAP,
OF INDIVIDUALS IN MAIN SUBSETS OF FIVE NETWORKS

<table>
<thead>
<tr>
<th>Network</th>
<th>Average Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>5.9</td>
</tr>
<tr>
<td>F</td>
<td>5.1</td>
</tr>
<tr>
<td>V</td>
<td>3.6</td>
</tr>
<tr>
<td>S</td>
<td>9.0</td>
</tr>
<tr>
<td>P</td>
<td>4.0</td>
</tr>
<tr>
<td>S*</td>
<td>6.9*</td>
</tr>
</tbody>
</table>

Average, using S at 24 5.5
Average, using S at 14 5.1

*Using only the first 14 net members
Figure 1
Figure 2
Figure 3: Diagrams of Five Empirical Networks
Figure 4
Figure 5
Figure 6
Figure 7: Connections based on name overlap of at least two. The S-Net has been limited to its first 14 members to give it the same number of points as B, F, and V.