Journal of Mathematical Sociology
1973 Vol. 3, pp 113-126

A GRAPH-THEORETIC DEFINITION OF A SOCIOMETRIC CLIQUE *

RICHARD D. ALBA

Columbia University

The intent of this paper is to provide a definition of a sociometric clique in the language of graph theory. This problem is viewed from two perspectives: maintaining fidelity to the intuitive notion of a clique; and providing adequate computational mechanics for large bodies of data. Luce's (1950) concept of an K-clique is used, but further qualifications are added. Two statistics or measures with associated probability distributions are defined for testing the adequacy of a subgraph which qualifies according to the definition.

1. INTRODUCTION

One of the factors which has most confused the discussion of sociometric clique identification in large bodies of data is the absence of a formal definition of a clique. Luce and Perry (1949) and Luce (1950) are the only well-known sources for such a formalization, and their attempts have not been followed out in the literature. Rather, more recent attempts to satisfactorily identify cliques have employed *ad hoc* clustering or taxonomic procedures which allow the data to suggest natural groupings. Notable among these are the approaches of MacRae (1960), Coleman and MacRae (1960), and Hubbel (1965).

One difficulty in the approach of Luce and Perry (1949) and Luce (1950) was that an adequate computational procedure to locate subsets of the data which satisfied their criteria was lacking. Recent computational literature, such as Bonner (1963) and especially Auguston and Minker (1970), contains algorithms which can be used to identify these subsets in reasonably large bodies of data (whose sizes are on the order of 100 to 1000 individuals).

Other difficulties, however, remain. The definition in Luce and Perry (1949), in which a clique is defined as a maximal complete subgraph, is too stringent for most purposes. The concept of *n*-clique, as presented in Luce (1950), may provide a suitable basis for a formalization but its properties must be explored further before any judgment can be made.

The intent of this paper is to provide a suitable definition of a sociometric clique.

*A version of this paper was read at the 1972 meetings of the American Sociological Association in New Orleans. This paper was prepared under National Science Foundation Grant GS 3231 to Professor Charles Kadushin of Teachers College, Columbia University. I am grateful for the comments of Professor Kadushin, as well as those of Dr. Kenneth Land of the Russell Sage Foundation and Professor Seymour Spilerman of the University of Wisconsin to a draft of this paper. I particularly want to thank Dr. Tom Louis of Columbia University for his help in formulating the applications of probability theory in this paper.

The concept of an *n*-clique provides a starting point. Upon close examination, it proves too weak for most purposes, and further criteria are added: connectedness, and a diameter of *n.* Two measures of group cohesiveness are then introduced along with associated probability distributions; each measures a different dimension of cohesiveness. It is then shown that groups which satisfy the above criteria for a clique but exhibit minimal cohesiveness in one particular dimension are trees. Hence, trees provide a trivial example of a sociometric clique, and procedures for locating all non-trivial cliques by the exclusion of trees are provided.

2. INTRODUCTORY GRAPH-THEORETICAL CONCEPTS

Before the discussion can begin, certain critical terms must be defined. In each case, the definitions are those of Harary (1969). The wordings are, in most cases, verbatim. Any theorems in later sections which have been taken from Harary are credited to him; all others are my own.

A *graph G* consists of a finite nonempty set V of *p points* together with a set X of q unordered pairs of distinct points of V. Each pair $X = \{u, v\}$ in X is a *line* of G, and if *{u, v*□ $\in X$, then u and v are said to be *adjacent* in G. Moreover, if $X = \{u, v\} \in X$, then x is said to be *incident* with the points u and v.

A *walk* of a graph G is an alternating sequence of points and lines $v_0, x_1, v_1, \ldots, v_{n-1}, x_n, v_n$ beginning and ending with points, in which each line is incident with the point immediately preceding it and with the point immediately following it. A walk is called a *path* if all points (and thus necessarily all the lines) are distinct. The *length* of a walk, and hence of a path, is the number of occurrences of lines in it.

A walk is *closed* if its first and last points are identical and is open otherwise. If a walk is closed, then it is a *cycle* provided its other n points are distinct and $n \geq 2$.

A graph is *connected* if every pair of points is connected by a path. The *distance* $d_G(u, v)$ between any pair of points u and v in G is the length of the shortest path, called a geodesic, in G connecting u and v; if there is no path connecting u and v, $d_G(u, v) = \infty$. The diameter, *d{G),* of a connected graph G is the length of any longest geodesic.

A *complete* graph has every pair of its points adjacent.

A *subgraph G'* of the graph G has a set of points V' and a set of lines X' such that $V' \subseteq V$ and $X' \subseteq X$ where $x = \{u, v\} \in X'$ iff $x \in X$ and $u, v \in V'$. In other words, the set of points of a subgraph G' of G is a subset of the set of points of G, and two points are adjacent in G' whenever they are adjacent in G. The subgraph G' may also be spoken of as the subgraph *of G induced by* the subset of points V'.

A subgraph is *maximal* with respect to some property whenever either it has the property or every pair of its points has the property and, upon the addition of any point, either it loses the property or there is some pair of its points which does not have the property.

3. THE PREVIOUS LITERATURE REVIEWED

Let us assume that we have some body of sociometric data for a given population. For simplicity, let us assume that we asked each person in the population to name those who have some particular relationship, say friendship, to him. To consider this body of data as a graph *G,* we will let the set of points, *VG),* correspond to the set of individuals. Then, a natural criterion for determining whether two points *u* and *v* are adjacent in *G* is the one of mutual choice; that is, the points *u* and *v* are adjacent if individual *u* names individual *v* and *v* names *u.*

It is important to note that this paper deals exclusively with graphs and undirected relations, unlike much of the recent sociometric literature (Holland and Leinhardt, 1970; Davis, 1970), in which directed graphs and directed relations are the primary formalization of a network. By and large, this literature is not concerned with constructing procedures for identifying various parts of a structure, but rather with verifying a theory of structure. This theory is expressed in various ways; we could crudely summarize it as the assertion that structure has two kinds of dimensions: one kind is hierarchical or ranked; the other is not. Davis (1970), for example, conceives of structure as a building, where the different floors are the different hierarchical levels and the different rooms on each floor are the different cliques on that level. Asymmetric or unreciprocated relations are associated with the dimensions of hierarchy; symmetric or mutual relations are associated with each particular level (Davis, 1970: 844).

Since this paper deals only with symmetric relations, it is consistent with the hierarchical notion of structure, in which direction has a meaning, only if we imagine the concepts presented here applied to a particular level of symmetric relations at a time. These concepts become particularly important when a given level is sufficiently extensive that it is useful to treat it as containing separate substructures. Referring back now to the criterion for adjacency presented above, absolute consistency with the hierarchical notion of structure is achieved by restricting the notion of adjacency to the situation of mutual choice. It is worthwhile to note in passing that, while much of the sociometric literature is concerned with relations which cannot be conceived of in a solely symmetrical fashion, there are situations (say, for example, the analysis of interlocking directorates) where the symmetric or undirected formalization is the natural one. In these latter cases, the concepts developed in this paper can be applied without restriction. Moreover, in some unbounded situations, in which respondents can give only approximate answers and in which they draw upon a potentially unlimited universe (for example, when they are asked, 'With whom did you have an interesting conversation last week?'), reifying the direction of choice may impose too strong a structure; in such cases it may be useful to force symmetry upon apparently directional data (Kadushin, 1970).

Let us now turn to the concept of a clique. In a very intuitive way, when we use the term sociometric clique, we mean a highly cohesive subgroup of individuals. Cohesiveness, however, has two possible dimensions. In one dimension, which we will call completeness, a cohesive subgroup would be one in which a high proportion of its pairs possess the appropriate relation. In theother, which we will call the centripetal-centrifugal dimension, a cohesive subgroup would be one in which relations among members of the subgroup are more important or more numerous than relations between members and non-members. We will return later to a formalization of these two dimensions. It is interesting to note that many informal definitions of cliques (Spilerman, 1966; Abelson, 1966) have referred solely to the centripetal-centrifugal dimension of cohesiveness, while Luce and Perry (1949) present a formalization of the completeness dimension of cohesiveness.

Utilizing our graph-theoretic terminology, the definition of a clique in Luce and Perry (1949) can be stated: a clique is a maximal complete subgraph. Hence, every pair of points in a clique is adjacent, and the addition of any point to the clique makes it incomplete. For example, if the relationship under consideration is friendship, a clique is a group of individuals, every two of whom are friends, which excludes no one who is a friend of everyone in the clique. Clearly, this definition is a formalization of the completeness dimension of cohesiveness.

This definition is quite stingy. A friendship group may lack but a few friendships to achieve completeness, and hence be found wanting as a clique. At the time when the Luce and Perry paper was written, a second difficulty presented itself: an adequate computational procedure to locate all maximal complete subgraphs was lacking.

With the widespread use of computers and the possibility of employing algorithms' which are not feasible for hand calculations due to the volume of calculations involved, this computational difficulty has been overcome. Interestingly enough, it has been overcome by researchers interested in developing fast, *ad hoc* clustering procedures for a variety of taxonomic problems. Bonner (1963) presents one algorithm for locating all maximal complete subgraphs; his algorithm, however, is very inefficient for large graphs. The algorithm of Bierstone, presented in Augustson and Minker (1970), represents a considerable computational improvement.

While the computational difficulty has been surmounted, the restrictiveness of the definition remains. The notion of completeness is crucial to the definition; that is, a relationship must exist between any two members of a clique. For the remainder of this discussion let us restrict the relationships under consideration to ones which require face-to-face interaction; again, friendship would provide an example. Then, the criterion of completeness would require face-to-face interaction between every two members of a clique. For many purposes, indirect interaction, that is, interaction accomplished through intermediaries, is sufficient. Social circles, as presented in Kadushin (1968), are a social unit which is closely related to the concept of clique and in which indirect interactions may predominate. However, even though we will allow for indirect interactions, we will want to restrict the maximum feasible social distance across which they can occur; this social distance will be measured by the number of intermediaries required for interaction. It is precisely to formalize such a concept that Luce (1950) introduced the notion of an *n-clique.* His discussion, however, is so different than that in this paper that the reader to whom the Luce paper is familiar may read this paper without fear of repition.

4.*N*-CLIQUES

In the ensuing discussion, $V(G)$ will mean the set of points associated with the graph G. Additionally, throughout the remainder of the paper, we will assume that the graph G is connected and has a diameter strictly greater than n. The *nth power* G^n of G is a graph with $V(G^n) = V(G)$ and such that u

and v are adjacent in G^n iff $d_G(u, v) \leq n;$ that is, two points are adjacent in G^n whenever the distance between them in G is n or less.

We define an *n-clique* of a graph G as a subgraph of G induced by a set of points V associated with a maximal complete subgraph of the power graph G^n. That is, each maximal complete subgraph of the power graph G^n has a set of points (or individuals) associated with it; this set of points together with the set of lines (or relations) connecting them in the original graph (or network) forms the subgraph we call an n-clique. Since each pair of points in an *n*-clique is adjacent in the power graph $G^n,$ the distance between the points in the original graph is less than or equal to $n.$ So we note an immediate correspondence between the concept of an n-clique and the concept of a clique as we outlined it in the previous section. Subsequent theorems will clarify this relationship.

It is worthwhile noting that there is a mathematical nicety which prevents us from defining an n-clique as a maximal complete subgraph of the power graph G^n. If it were so defined, the set of lines associated with it would be derived from the set of lines associated with the power graph, and every pair of points in an *n*-clique would therefore be adjacent. The intent of the concept is of course to consider a subgraph which is a substructure of the original graph or network; so we define it as a subgraph induced on the original graph.

The computational procedure for identifying all *n*-cliques of a given graph G is straightforward. First, the nth power graph of G is computed; this may be done by computing its adjacency matrix from the adjacency matrix of $G.$ Then all maximal complete subgraphs are identified using the Bierstone algorithm (Augustson and Minker, 1970). The sets of points associated with these maximal complete subgraphs together with the lines connecting the points in G are the desired *n*-cliques, We now prove some theorems concerning the basic properties of n-cliques.

Theorem 1.1: If G' is a subgraph of $G,$ then G' is an n-clique of G iff $V(G')$ is a maximal subset of $V(G)$ with the property that $d_G(u, v) \leq n$ whenever $u, v\in V(G').$

Proof. Let G' be an n-clique of $G.$ Let $u, v\in V\{G').$ Since $V(G')$ is derived from a complete subgraph of $G^n,$ u and v are adjacent in $G^n;$ hence, $d_G(u, v) \leq n.$ The result that $V\{G')$ is a maximal subset is a clear consequence of the fact that G' is a maximal subgraph.

Let us now suppose that $V\{G')$ is a maximal subset of $V\{G)$ with the property that $d_G(u, v) \leq n$ whenever $u, v \in V(G').$ Let G'' be the subgraph of G^n induced by $V(G').$ Clearly, G'' is a complete subgraph. Suppose that it is not maximal; that is, suppose there is some $w \notin V(G')$ such that for any $u \in V(G'),$ w and u are adjacent in $G''.$ Then $d_G(w, u) \leq n$ for all $u \in V(G')$ which contradicts our assumption that $V(G')$ is maximal. Hence, G'' is a maximal complete subgraph of $G^n;$ that is, G' is an *n*-clique of $G.$

It is clear from this theorem that an *n*-clique possesses the characteristics which we stated as desirable for a clique in the previous section. An *n*-clique allows for indirect interaction through a chain or chains of intermediaries, but places a limit on the maximum social distance across which indirect interaction may occur since the social distance between two members of an *n*-clique may not exceed $n.$ Moreover, since it is maximal, it omits no one who is within the required distance of the other member.

Theorem 1.2: If G' is an n-clique of G, then there is at least one pair, $u, v\ V(G')$ such that $d_G(u, v) = n$.

Proof: Since G' is an n-clique, $d_G(u, v) \leq n$ for any $u, v \in V(G')$. Suppose that for any $u, v \in V\{G'), d_G(u, v) < n$. Since $d(G) > n$, there must be some $w \notin V(G')$. Moreover, since G is connected, there must be some $w \notin V(G')$ and some $u \in V(G')$ such that u and w are adjacent in G. Hence $d_G(u, w) = 1$. But for any $v \in V(G'), d_G(u, v) < n$, Hence, for any $v \in V(G'), d_G(v, w) \leq n$. Then it must be that $w \in V(G')$, which is a contradiction.

The two lemmas in the discussion above are concerned with the properties that the subset of points of an n-clique has in the context of the larger graph, rather than with the properties of an n-clique as a graph itself. This distinction is important. As both Luce (1950) and Spilerman (1965) have noted, individuals in an n-clique may be connected through intermediaries who are non-members. Thus, in looking at the properties of the subset of points, we are looking at the properties which a social unit possesses by virtue of its occurrence in a larger social context. In looking at the properties of an n-clique as a subgraph, we are examining the properties which a social unit has by virtue only of relations among its members.

Imagine the following situation. We invite a group of people to a party, each of whom knows all the others or has at least one friend in common with each of those he doesn't know. If we exclude no one who satisfies these criteria, our party is a 2-clique. Suppose now that each person can talk only to those he already knows and to those to whom he is introduced by those he already knows; then the success or failure of the party will depend upon the properties of this 2-clique as a graph. One question is whether everyone will have at least one person at the party to talk to. Another is whether the party will subdivide into a number of factions where the members of one faction do not speak to the members of any other faction. Curiously, there may be people who have no one to talk to, and the party may divide into mutually exclusive factions. The answers to these questions rest upon the connectedness of the n-clique as a graph.

Assertion: An n-clique is not necessarily connected.

An example is provided by the graph in Figure 1. $\{A, B, C\}$ and $\{1, 2, 3\}$ are two

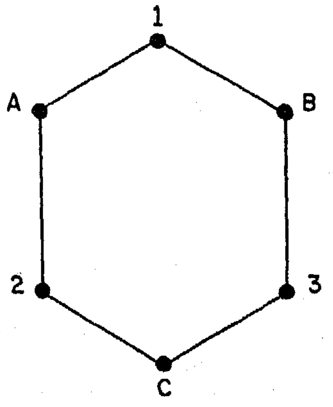
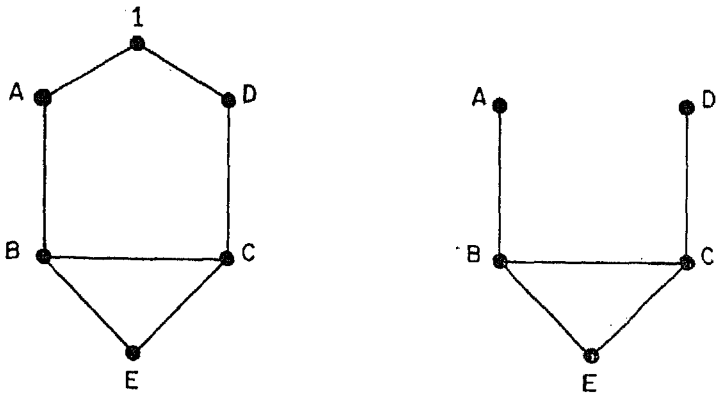


FIGURE 1

of the 2-cliques of this graph. Each of these 2-cliques is completely disconnected; there are no two points which are adjacent in either of the subgraphs.

Even when the *n*-clique is connected, however, its diameter may be greater than *n;* of course, there must be some path of length less than or equal to *n* joining every two points in an n-clique, but such a path may contain points which lie outside the *n*-clique; hence, when we restrict ourselves to paths which contain only points within the *n*-clique, there may be pairs which are joined only by paths which are longer than *n.* An example is provided in Figure 2. The subgraph in (b) is a 2-clique of the graph in (a). Although this 2-clique is connected, its diameter is 3, since it does not include the point 1 which lies on the only path of length 2 connecting *A* and D.



(a) (b)

FIGURE 2

Theorem 1.3: If G' is a connected n-clique, then $d(G') \geq n$.

Proof: Let G' be a connected *n*-clique of G such that $d(G') < n$.

Then, for any $u, ,v \in V(G')$, $d_{G'}(u, v) < n$. Clearly, then $d_G(u, v) < n$. But this contradicts the result of theorem 1.2.

The fact that an *n*-clique may have a diameter exceeding *n* provides difficulty in certain cases. If we restrict interaction to interaction which occurs only through intermediaries who are also clique members, then the social distance between some members may be greater than the maximum feasible social distance for interaction, *n.* This consideration suggests that we restrict ourselves to *n*-cliques which have a diameter of *n.*

5. DEFINING A SOCIOMETRIC CLIQUE

Let us now present a tentative definition of a sociometric clique. A *sociometric clique of diameter n* is an *n*-clique of diameter *n.* Thereby, sociometric cliques of diameter 1 are identical with the cliques provided by the definition of Luce and Perry (1949). We note in passing that a sociometric clique is necessarily connected since it has a finite diameter.

This definition has the virtue of computational clarity: first, all *n*-cliques are computed; second, the *n*-cliques with diameter *n* are isolated from among these. It seemingly contains a redundancy in that the parameter *'n'* appears twice. An equivalent, nonredundant way of stating the definition is: a sociometric clique of diameter *n* is a maximal subgraph of diameter *n.*

Theorem 2.1: *G'* is an *n*-clique of *G* with diameter *n* iff *G'* is a maximal subgraph of *G* with diameter *n.*

Proof: Left to the reader.

As a final remark on the immediate properties of this definition, we will note that' an individual may be assigned to more than one clique of diameter *n.* The graph in Figure 3 provides an example. It has two cliques of diameter 1: *{A, C, D}* and *{B, C, D}.* Points *C* and *D* lie in both of these cliques.

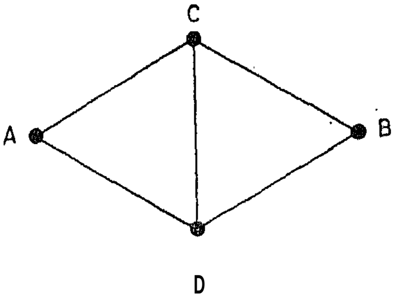


FIGURE 3

The possibility of non-unique assignment raises certain questions to which completely satisfactory answers cannot be given. Clearly, in certain situations, a unique assignment may be desired or necessary; in others, even where it is permissible for the same individual to be assigned to two or more cliques, the overlap between two cliques of a given diameter may be so great that treating them as separate cliques may be ludicrous. In each of these cases, there are two possible procedures; however, it is not clear which is more reasonable nor what the exact consequences of each procedure are. One procedure is simply to merge cliques which have a large overlap ('large' may of course mean any overlap at all). The effect of this procedure is to increase the diameter of the clique resulting from a merger in a non-systematic way; that is, the extended clique will have a diameter larger than *n,* but it may exclude individuals who would lie within a sociometric clique of the larger diameter. These considerations naturally suggest the second procedure: to increase the maximum permissible dia-meter for a clique; that is, to increase *n.* However, increasing the diameter may lead to a severe attenuation of the 'density' of the resulting cliques. Clearly, then, more experience and analysis is required in order to provide guidelines for the treatment of non-unique assignment of individuals and extensive overlap between cliques.

One possible way to resolve these difficulties is to combine subgraphs in accordance. with a procedure which preserves the overall 'density' of the resulting, combined: subgraphs. No such procedure can be presently described; it must await some future paper. However, these difficulties suggest an exploration of the concept of 'density' as well as of cohesiveness generally; moreover, some formalization of these concepts is required for an adequate evaluation of the application of the definition developed earlier. The next section develops these concepts in detail.

6. THE COHESIVENESS OF A CLIQUE

Informal definitions of a clique (Spilerman, 1966; Abelson, 1966; Davis, 1967) have characterized it as a highly cohesive subgroup. As mentioned above, there are at least two different interpretations which have been given to the concept of cohesiveness. One, which we have called the completeness dimension of cohesiveness, refers to the degree to which pairs of individuals within the clique possess the appropriate relation. The second, called the centripetal-centrifugal dimension of cohesiveness, refers to the degree to which members relate to each other rather than to non-members.

We would like now to formalize these dimensions by introducing appropriate measures. Completeness cohesiveness can clearly be measured by the degree to which the subgraph or clique approaches graph-theoretical completeness. A simple measure of the degree of completeness is the ratio of the number of lines joining clique members to the possible number of such lines. For any graph with p points, the number of possible lines is $p(p-1)/2$. If we let q stand for the number of actual lines, then the completeness measure for a clique with p members is defined as $2q/p(p-1)$.

In formalizing the centripetal-centrifugal dimension, we will make a slight alternation in its meaning. Rather than formalizing a measure of the degree to which members relate to each other rather than non-members, we will introduce a measure of the degree to which members relate to non-members. The reason for this alteration in meaning is that it permits a very simple probability function to be associated with the measure, as will be done below.

We now introduce a measure of centripetal cohesiveness (the alteration in name is appropriate for the alteration in meaning). A straightforward measure can be defined as the ratio of the count of lines joining members and non-members. For any subgraph or clique of p points drawn from a larger graph containing P points, the number of possible lines joining members and non-members is $p(P—p)$. Therefore, if the actual number of such lines is q, the centripetal measure for the subgraph is given by $\frac{q}{p(P-p)}$.

Ideally, a sociometric clique should be dense (we will use the term 'dense' interchangably with 'complete') and isolated; that is, its members should have many relations with each other and few with non-members. Hence, a satisfactory value for the completeness measure is a relatively high one, while a satisfactory value for the centripetal measure is a relatively low one. Somewhat more precisely, a satisfactory value for the completeness measure is one which is significantly higher than we would expect from the average density of the graph, while a satisfactory value for the centripetal measure is one which is significantly lower than we would expect. A natural way to measure the significance of a particular observed value is to determine the probability of finding subgraphs with the same value or better; the less probable the value, the more significant it is.

Hence our next step is to identify appropriate probability distributions for these measures. With respect to the completeness measure, our problem is precisely the following: what is the probability of observing a completeness greater than or equal to a fixed value in a subgraph drawn randomly from a graph with a given density? To answer it, we will make one slight change so that it reads: what is the probability of observing a number of lines equal to or greater than some fixed number in a subgraph with a fixed number of points which is randomly drawn from a graph with a given density? This problem is exactly analogous to a standard probability textbook problem: namely, that of sampling without replacement from an urn containing balls of two colors (Parzen, 1960: 33); to see the analogy, it is only necessary to imagine each pair of points as a ball which is one color if the points are joined by a line and another if they are not. The hypergeometric probability law can be applied to this situation to determine the probability of observing a fixed number of balls of one color in a sample of balls drawn without replacement (Parzen, 1960: 179); hence it can be used to determine the probability of observing a fixed number of lines in a randomly drawn subgraph.

The application of the hypergeometric probability law yields the following results in the case of the completeness measure. Suppose that we draw a subgraph containing w points from a larger graph containing *N* points and *Q* lines. We also will use the following notation: l(n) will stand for the observed number of lines in a subgraph with *n* points: *L(n)* will stand for the possible number of lines in a subgraph with *n* points, that is, $\frac{n(n-1)}{2}$. We wish to know first the probability of observing exactly

$$P[l(n) = q] = \frac{\binom{Q}{q}\binom{L(N)-Q}{L(n)-q}}{\binom{L(N)}{L(n)}}$$

Now, the probability of observing *q* or more lines is given by the sum of the probabilities of observing exactly *q* lines, exactly *q*+1 lines, etc.; the final term in this sum is either the probability of observing exactly *Q* lines or the probability of observing exactly *L(n)* lines, depending on whether *Q* or *L(n)* is smaller. Therefore, the probability of observing *q* or more lines is given by the expression:

$$P[l(n) \geq q] = \sum_{k=q}^{\min(L(n),\,Q)} \frac{\binom{Q}{k}\binom{L(N)-Q}{L(n)-k}}{\binom{L(N)}{L(n)}}$$

Similarly, with respect to the centripetal measure, we want to know the probability of observing a value of the centripetal measure less than or equal to a fixed value in a subgraph drawn randomly from a graph with a given density. We will use the following notation: *l(n, N)* will stand for the actual number of lines joining a subgraph of *n* points to the surrounding subgraph of *N* - *n* points; *L(n, N)* will stand for the possible number of such lines, that is, *n(N - n)*. Then, by a process of reasoning similar to that for the completeness measure, we can establish the following expression for that probability:

$$P[l(n, N) \leq q] = \sum_{k=0}^{q} \frac{\binom{Q}{k}\binom{L(N)-Q}{L(n, N)-k}}{\binom{L(N)}{L(n, N)}}$$

There is a refinement worth stating. There is a simple formula which exhibits the way that *Q,* the number of lines in the entire graph, is partitioned by the selection of a subgraph of *n* points:

$$Q = l(n)+l(n, N)+l(N-n)$$

Clearly, for fixed *Q,* the larger *l*(*n*) is, the stricter is the upper bound on *l(n, N).* Hence, a better test of how probable a particular value of *l(n, N)* is would involve computing the probability under the assumption that *l(n)* is fixed at its observed value; that is, a better test would involve computing the conditional probability $P[l(n, N) \leq q|l(n) = r]$. To compute this probability, it is sufficient to reduce the graph by *L(n)* pairs of points, of which *r* are lines, and then to compute the probability of observing *q* or fewer lines in a randomly drawn set containing *L(n), N* pairs of points. This probability is given by the expression:

$$P[l(n, N) \leq q|l(n) = r] = \sum_{k=0}^{q} \frac{\binom{Q-r}{k}\binom{L(N)-L(n)-(Q-r)}{L(n, N)-k}}{\binom{L(N)-L(n)}{L(n, N)}}$$

The use of these probabilities is like that of tests of significance. Their values determine whether a particular subgraph identified as a sociometric clique by the definition given earlier is in fact significantly different than any subgraph drawn randomly from a graph with the density of the original graph. If the subgraph is not significantly different, for example, if it is not significantly more dense, then it is not of interest to us. Thus, these probabilities provide a way of eliminating subgraphs which are not cliques in the sense that they are not sufficiently cohesive, but which have been selected by the criteria for our definition.

It should be noted that the use of the centripetal measure and its associated probability of occurrence is ambiguous. Clearly, when we have found a subgraph which is significantly more dense than we would expect based upon the average density of the graph and whose connections to the surrounding graph are significantly less frequent than we would expect, then we can rest content that we have identified an object worthy of being called a sociometric clique. However, the situation where we have found a subgraph which is significantly dense but yet has frequent connections to the surrounding graph is not entirely clear. Such a subgraph might, for example, be contained within a sociometric clique of a larger diameter which would have satisfactory values for both the completeness and centripetal measures. On the other hand, it might be the inner core of a network; that is, it might be the dense center of the network and thus have a significant number of connections to the less dense periphery. In the latter case, it might still be an analytically interesting object, one which we would not want to throw away. There is no obvious way to resolve the issues raised in this paragraph; clearly, as in the case of extensive overlap between cliques, guidelines cannot yet be established.

Finally, there is a refinement to make in the probability function associated with the completeness measure. In many situations, particularly those where the average density of the original graph is low, the simple fact that a subgraph is connected will be sufficiently improbable so that any connected subgraph will appear sufficiently complete or dense. Therefore, a finer test of the probability of occurrence of a given level of completeness would involve computing the probability of observing more than a specific number of lines in a randomly drawn subgraph under the assumption that the subgraph contains the minimal number of lines required for it to be connected. As we will see below, any connected graph of n points must contain at least n - 1 lines. Therefore, we wish to compute the probability that a randomly drawn subgraph of n points has q or more lines, given that it has at least n - 1 lines. This probability is given by the expression:

$$P[l(n) \geq q | l(n) \geq n-1] = \frac{\sum_{k=q}^{\min(L(n),\, Q)} \binom{Q}{k} \binom{L(N)-Q}{L(n)-k}}{\sum_{k=n-1}^{\min(L(n),\, Q)} \binom{Q}{k} \binom{L(N)-Q}{L(n)-k}}$$

We note that if a subgraph only contains the minimal number of lines required for connectedness, then its probability of occurrence according to the expression above is 1. Subgraphs with only the minimal number of lines required for connectedness are considered further in the next section.

7. TRIVIAL SOCIOMETRIC CLIQUES AND TREES

Consider any connected graph G with p points. Since G is connected, every pair of its points must be joined by a path. Presumably, G has a minimum number of lines when all its points lie on precisely one path. If all its points lie on one path, then the number of lines of G is p - 1; the length of this path *is* p - 1, and G is a path of length $p - 1$.

Since a sociometric clique is connected by definition, the minimum number of lines in a clique of p points *is* p - 1. Hence, the minimum value of the completeness measure for a clique of p points is $2/p$.

Any connected graph of p points which has p - 1 lines is called a *tree* (Harary, 1969). A path is one example of a tree. Trees have been extensively studied in graph theory. Some of their other properties (Harary, 1969) are:

(1) every two points of a tree are joined by a unique path;
(2) a tree has no cycles.

The reader should satisfy himself that there are graphs which have subgraphs that are sociometric cliques for some diameter n as well as trees.

A sociometric clique which is a tree we will call a *trivial* clique since it exhibits minimum completeness. We will then confine our interest to non-trivial cliques: those which are not trees. Therefore we look for axing criteria with which to cut away trees from consideration and hence simplify our computations. Of course, any clique of p points which is a tree will have a density whose probability of occurrence is 1. Computationally, however, this axe is not completely satisfying since it requires us to calculate the completeness measure and it associated probability. There is still simpler criterion: if a sociometric clique of diameter n has n + 1 points, then it is a tree.

Theorem 2: If G' is a sociometric clique of diameter n and has $n+1$ points, then it is a tree.

Proof: Since G' is a sociometric clique of diameter n, there are $x, y \in V\{G')$ such that $d_{G'}(x, y) = n$. Since the geodesic P in G' joining x and y is of length n and G' has n +1 points, all the points of G' must lie on P.

Suppose G' is not a tree. Then there are $u, v \in V(G')$ such that there are at least two distinct paths in G' joining u and v. Let us name these paths P_1 and P_2. Now one of these paths must contain a point which is neither u nor v and does not lie on the other path. Let us call this point u_1 and assume that it lies on path P_1. Since u and v must both lie on the geodesic joining x and y, let« be closer to x than is v, and hence v closer to y than is u. Consider the path P' composed of the following segments: the path of x to u along P, the path of u to v along P_2 and the path of v to y along P. Since this path does not contain u it is shorter than P, which is a contradiction. Hence, G' is a tree.

This theorem greatly simplified the computations for identifying non-trivial cliques of diameter n. The simplification is apparent from the following corollary.

Corollary 2.1: If G' is an n-clique with n+1 points, then either G' is not connected or G' is a tree.

Proof: Suppose G' is connected. Since it has n + 1 points, its diameter must be n. Therefore, it is a sociometric clique of diameter n. By the above theorem, it must be a tree.

Once we have calculated all n-cliques, we can dismiss those which contain only n + 1 points from further consideration.

8. CONCLUSION

The formalization in this paper, which is based in part on the work of Luce and Perry (1949) and Luce (1950), offers real promise. The definition makes strong intuitive sense and the subsets satisfying the definition can be computed for any body of empirical data.

The formalizations of the two dimensions of cohesiveness provide a way of evaluating the success of the definition in locating subgraphs which resemble our intuitive conceptions of a clique. Certain problems have been clearly indicated; among them are the possibilities of non-unique assignment of individuals to cliques and extensive overlap between cliques, as well as the ambiguous interpretation of the centripital measure. To aid researchers in evaluating the techniques proposed in this paper a computer program called COMPLT (Alba, 1971) has been written; this program embodies the techniques proposed here, and can be obtained from the author of the paper.

REFERENCES

Abelson, Robert P. (1966) Mathematical models in social psychology. *Advances in Experitnetal Social Psychology,* Volume III, ed., Leonard Berkowitz, New York.

Alba, Richard D. (1971) COMPLT—A program for analyzing sociometric data and clustering similarity matrices. Mimeo.

Augustson, Jack G. and Jack Minker (1970) An analysis of some graph theoretical cluster techniques. *Journal of the ACM* 17, 571-588.

Bonner, R. E. (1964) On some clustering techniques. *IBM Journal of Research and Development,* 22-32.

Coleman, J. and D. MacRae, Jr. (1960) Electronic processing of sociometric data for groups up to a thousand in size. *American Sociological Review* 25, 722-726.

Davis, James A. (1967) Clustering and structural balance in graphs. *Hitman Relations* 20,181-187.

Davis, James A. (1970) Clustering and hierarchy in interpersonal relations: testing two graph theoretical models on 742 sociomatrices. *American Sociological Review,* 35, 843-851.

Harary, Frank (1969) *Graph Theory*. Reading.

Holland, Paul W. and Samuel Leinhardt (1970) A method for detecting structure in sociometric data. *American Journal of Sociology 16,* 492-513.

Kadushin, Charles (1970) Sociometry and macro-sociology. Paper delivered at the 1970 meetings of the the International Sociological Association, Varna.

Luce, R. D. and A. Perry (1949) A method of matrix analysis of group structure. *Psychometrika 14,* 94-116.

Luce, R. D. (1950) Connectivity and generalized cliques in sociometric group structure. *Psychometrika* 15,169-190.

MacRae, D., Jr. (1960) Direct factor analysis of sociometric data. *Sociometry* 23, 360-371,

Parzen, Emanuel (1960) *Modem Probability Theory and Its Applications*. New York.

Spilerman, S. (1966) Structural analysis and the generation of sociograms. *Behavioral Science, 11,* 312-318.