7
Cohesive Subgroups

One of the major concerns of social network analysis is identification of cohesive subgroups of actors within a network. Cohesive subgroups are subsets of actors among whom there are relatively strong, direct, intense, frequent, or positive ties. These methods attempt, in part, to formalize the intuitive and theoretical notion of social group using social network properties. However, since the concept of social group as used by social and behavioral scientists is quite general, and there are many specific properties of a social network that are related to the cohesiveness of subgroups, there are many possible social network subgroup definitions.

In this chapter and the next we discuss methods for finding cohesive subgroups of actors within a social network. In this chapter we discuss methods for analyzing one-mode networks, with a single set of actors and a single relation. In Chapter 8 we continue the discussion of cohesive subgroups and related ideas, but focus on affiliation networks. Affiliation networks are two-mode networks consisting of a set of actors and a set of events. Cohesive subgroups in one-mode networks focus on properties of pairwise ties, whereas cohesive subgroups in two-mode affiliation networks focus on ties existing among actors through their joint membership in collectivities. Thus, one major difference between this chapter and the next is whether one-mode or two-mode data are being analyzed.

We begin with an overview of the theoretical motivation for studying cohesive subgroups in social networks and discuss general properties of cohesive subgroups that have influenced network formalizations. We then discuss how to assess the cohesiveness of network subgroups, and extend subgroup methods to directional relations and to valued relations. The final section of this chapter briefly discusses alternative approaches for studying cohesiveness in networks using multidimensional scaling and factor analysis. Most of the methods discussed in this chapter are based on graph theoretic ideas, and use graph theoretic concepts and notation. Thus, it might be useful to review Chapter 4 before reading the rest of this chapter.

7.1 Background

In this section we discuss the theoretical background for social groups, briefly outline some ways to conceptualize cohesive subgroups, and review key notation and graph theoretic concepts that are used to study cohesive subgroups.

7.1.1 Social Group and Subgroup

Many authors have discussed the role of social cohesion in social explanations and theories (Burt 1984; Collins 1988; Erickson 1988; Friedkin 1984). Friedkin examines the use of network cohesion as an explanatory variable in sociological theories, especially for studying the emergence of consensus among members of a group:

> Structural cohesion models are founded upon the causal propositions that pressures toward uniformity occur when there is a positively valued interaction between two persons; that these pressures may occur by being "transmitted" through intermediaries even when two persons are not in direct contact; and that such indirect pressures toward uniformity are associated with the number of short indirect communication channels connecting the persons. (1984, page 236)

Consequently, according to this idea, one expects greater homogeneity among persons who have relatively frequent face-to-face contact or who are connected through intermediaries, and less homogeneity among persons who have less frequent contact (Friedkin 1984). In his review of sociological theory, Collins (1988) also states the importance of cohesion in social network analysis:

> The more tightly that individuals are tied into a network, the more they are affected by group standards .... (page 416)

Collins continues, noting that

> Actually, there are two factors operating here, which we can see from network analysis: how many ties an individual has to the group and how closed the entire group is to outsiders. Isolated and tightly connected groups make up a clique; within such highly cohesive groups, individuals tend to have very homogeneous beliefs. (page 417)

Cohesive subgroups are theoretically important according to these theories because of social forces operating through direct contact among subgroup members, through indirect conduct transmitted via intermediaries, or through the relative cohesion within as compared to outside the subgroup. Such theories provide motivation for cohesive subgroup methods for one-mode social networks (in which ties are measured between pairs of actors). These ideas are all used to study cohesive subgroups in social networks.

The notions of social group, subgroup, clique, and so on are widely used in the social sciences, particularly in social psychology and sociology. Although the notion of social group has received widespread attention in the social sciences, researchers often use the word without giving it a precise formal definition. As noted by Freeman (1984, 1992a) and Borgatti, Everett, and Freeman (1991) authors often assume that since "everybody knows what it means" it can be used without precise definition. Freeman reviews the history of the concept of group in sociology with special attention to network formalizations of this concept (Freeman 1992a).

Many network researchers who have developed or reviewed methods for cohesive subgroups in social networks have noted that these methods attempt to formalize the notion of social group (Seidman and Foster 1978a, 1978b; Alba and Moore 1978; Mokken 1979; Burt 1980; Freeman 1984, 1992a; Sailer and Gaulin 1984). According to these authors, the concept of social group can be studied by looking at properties of subsets of actors within a network. In social network analysis, the notion of subgroup is formalized by the general property of *cohesion* among subgroup members based on specified properties of the ties among the members. However, since the property of cohesion of a subgroup can be quantified using several different specific network properties, cohesive subgroups can be formalized by looking at many different properties of the ties among subsets of actors.

Although the literature on cohesive subgroups in networks contains numerous ways to conceptualize the idea of subgroups, there are four general properties of cohesive subgroups that have influenced social network formalizations of this concept. Briefly, these are:

- The mutuality of ties
- The closeness or reachability of subgroup members
- The frequency of ties among members

- The relative frequency of ties among subgroup members compared to non-members

Subgroups based on mutuality of ties require that all pairs of subgroup members "choose" each other (or are adjacent); subgroups based on reachability require that all subgroup members be reachable to each other, but not necessarily adjacent; subgroups based on numerous ties require that subgroup members have ties to many others within the subgroup; and subgroups based on the relative density or frequency of ties require that subgroups be relatively cohesive when compared to the remainder of the network. Successive definitions weaken the first notion of adjacency among all subgroup members. These general subgroup ideas lead to methods that focus on different social network properties. Thus, our discussion in this chapter is divided into sections, each of which takes up methods that are primarily motivated by one of these ideas.

In contrast to these ideas that focus on ties between pairs of actors in one-mode networks, some cohesive subgroup ideas are concerned with the linkages that are established among individuals by virtue of their common membership in collectivities. These ideas motivate methods for studying affiliation networks, which we discuss in Chapter 8.

Before we present the subgroup methods for one-mode networks, let us review some basic concepts and definitions from graph theory.

7.1.2 Notation

Our presentation of notation here is intentionally brief, since these ideas were covered in detail in Chapters 3 and 4. To start, we will limit our attention to graphs, and thus, to dichotomous nondirectional relations.

We begin with a graph, $\mathscr{G}$, consisting of a set of nodes, $\mathscr{N}$, and a set of lines, $\mathscr{L}$. Each line connects a pair of nodes in $\mathscr{G}$. Two nodes that are connected by a line are said to be *adjacent.* A node generated subgraph, $\mathscr{G}_s$, of $\mathscr{G}$, consists of a subset of nodes, $\mathscr{N}_s$, where $\mathscr{N}_s \subseteq \mathscr{N}$, along with the lines from $\mathscr{L}$ that link the nodes in $\mathscr{G}_s$. We will refer to a subset of nodes as a *subgroup* or *subset*, and the nodes along with the lines among them as a *subgraph.* A graph is *complete* if all nodes are adjacent; that is, if each pair of nodes is connected by an line. Similarly, a subgraph, $\mathscr{G}_s$, is complete if all pairs of nodes in it are adjacent.

A *path* connecting two nodes is a sequence of distinct nodes and lines beginning with the first node and terminating with the last. If there is a path between two nodes then they are said to be *reachable.* The *length* of a path is the number of lines in it. A shortest path between two nodes is called a *geodesic*, and the (geodesic) distance between two nodes, denoted by $d(i,j)$, is the length of a shortest path between them. The *diameter* of a graph is the length of the longest geodesic between any pair of nodes in the graph. In other words, the diameter of a graph is the maximum geodesic distance between any pair of nodes; $\max(d(i,j))$, for $n_i, n_j \in \mathcal{N}$. Similarly, the *diameter of a subgraph* can be defined as the longest geodesic between two nodes within the subgraph. The diameter of a subgraph is defined on the subset of nodes and lines that are present in the subgraph.

A graph is *connected* if there is a path between each pair of nodes in the graph. A subgraph is connected if there is a path between each pair of nodes in the subgraph, and the path contains only nodes and lines within the subgraph. The *degree* of a node, $d(i)$, is the number of nodes that are adjacent to it. The degree of node i in subgraph $\mathscr{G}_s$ is denoted by $d_s(i)$, and is defined as the number of nodes within the subgraph that are adjacent to node i.

A subgraph is said to be *maximal* with respect to some property (for example, completeness) if that property holds for the subgraph, but does not hold if additional nodes and the lines incident with them are added to the subgraph. If a subgraph is maximal with respect to a property, then that property holds for the subgraph, $\mathscr{G}_s$, but not for any larger subgraph that contains $\mathscr{G}_s$ (Mokken 1979). For example, a *component* of a graph is a maximal connected subgraph (Hage and Harary 1983). The presence of two or more components in a graph indicates that the graph is *disconnected*.

We can now define some interesting subgroup ideas using these graph theoretic concepts.

7.2 Subgroups Based on Complete Mutuality

The earliest researchers interested in cohesive subgroups gathered and studied sociometric data on affective ties, such as friendship or liking in small face-to-face groups, in order to identify "cliquish" subgroups. Network data on friendship nominations often give rise to directional dichotomous relations. Festinger (1949) and Luce and Perry (1949) argued that cohesive subgroups in directional dichotomous relations would be characterized by sets of people among whom all friendship choices were mutual. Specifically, Luce and Perry and Festinger proposed that a clique for a relation of positive affect is a subset of people among whom all choices are mutual, and no other people can be added to the subset who also have mutual choices with all members of the subset. This definition of a clique is appropriate for a directional dichotomous relation.

The clique is the foundational idea for studying cohesive subgroups in social networks. Graph theory provides a precise formal definition of a clique that is appropriate for a *nondirectional* dichotomous relation.

7.2.1 *Definition of a Clique*

A *clique* in a graph is a maximal complete subgraph of three or more nodes. It consists of a subset of nodes, all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique (Luce and Perry 1949; Harary, Norman, and Cartwright 1965). The restriction that the clique contain at least three nodes is included so that mutual dyads are not considered to be cliques. One can think of a clique as a collection of actors all of whom "choose" each other, and there is no other actor in the group who also "chooses" and is "chosen" by all of the members of the clique.

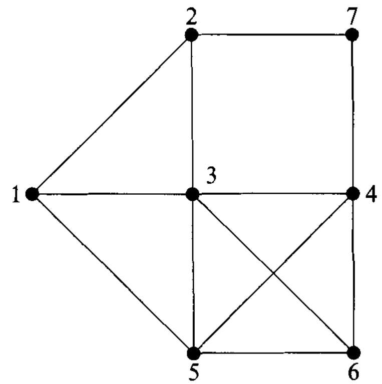
The clique definition is a useful starting point for specifying the formal properties that a cohesive subgroup should have. It has well-specified mathematical properties, and also captures much of the intuitive notion of cohesive subgroup; however, it has limitations, which we discuss below.

Figure 7.1 shows a graph and a listing of the cliques contained in it. The reader can verify that these subgraphs are in fact cliques, and that there are no remaining cliques in the graph.

Notice that cliques in a graph may overlap. The same node or set of nodes might belong to more than one clique. For example, in Figure 7.1 node 3 belongs in all three cliques. Also, there may be nodes that do not belong to any cliques (for example node 7 in Figure 7.1). However, no clique can be entirely contained within another clique, because if it were the smaller clique would not be maximal.

7.2.2 *An Example*

We will use the example of the relations of marriage and business among Padgett's Florentine families to illustrate cohesive subgroups throughout this chapter. Recall that both of these relations are dichotomous and nondirectional. We used the network analysis programs *GRADAP 2.0*



cliques: $\{1,2,3\}$, $\{1,3,5\}$, and $\{3,4,5,6\}$

Fig. 7.1. A graph and its cliques

(Sprenger and Stokman 1989) and *UCINET IV* (Borgatti, Everett, and Freeman 1991) to do the subgroup analyses described in this chapter.

First consider the relation of marriage among these families. For the marriage relation there are three cliques:

- Bischeri Peruzzi Strozzi
- Castellani Peruzzi Strozzi
- Medici Ridolfi Tornabuoni

Only seven of the sixteen families in this network belong to any clique on the marriage relation. Furthermore, the cliques are small; each clique contains only the minimum three families. By definition, there has been a marriage between all pairs of families in each clique. Notice that the first two cliques contain two members in common (Peruzzi and Strozzi), and differ only by a single member. However, the four families, Bischeri, Castellani, Peruzzi and Strozzi, do not form a clique because there is no marriage tie between Castellani and Bischeri.

For the business relation there are five cliques:

- Barbadori Castellani Peruzzi
- Barbadori Ginori Medici
- Bischeri Guadagni Lamberteschi
- Bischeri Lamberteschi Peruzzi
- Castellani Lamberteschi Peruzzi

Eight of the sixteen families belong to at least one clique on the business relation, and some families (for example Lamberteschi, Bischeri, and Peruzzi) belong to several cliques on this relation. As we saw with the marriage relation, the cliques are small (no more than three members) and there is considerable overlap among them. However, the cliques that are present in the business relation are different from the cliques that are present in the marriage relation.

7.2.3 Considerations

A clique is a *very* strict definition of cohesive subgroup. In fact, Alba (1973) calls it "stingy." The absence of a single line, or in sociometric terms, the absence of a single tie or "choice," will prevent a subgraph from being a clique. In a sparse network there may be very few cliques (as with the marriage relation among the Florentine families). In addition, the sizes of the cliques will be limited by the degree of the nodes. This can be a problem if the number of ties that an actor can have is limited by the data collection design. For example, in a sociometric study using a fixed choice design in which respondents are asked to list their three best friends, each person can be adjacent to at most three other people. Thus there can be no clique with more than four members. In general, if actors are restricted to k ties, then there can be no clique in the resulting data that has more than $k + 1$ members.

Early researchers were concerned with methods for detecting cliques in networks (Festinger 1949; Luce and Perry 1949; Luce 1950; Harary and Ross 1957). More recently, researchers have realized that cliques seldom are useful in analysis of actual data because the definition is too strict. Actual data rarely contain interesting cliques, since the absence of a single tie among subgroup members prevents the subgroup from meeting the clique definition. In addition, cliques that do occur are often quite small, and overlap one another (as we have seen in the analysis of Padgett's Florentine families).

An additional limitation of clique as a formalization of cohesive subgroup is that there is no internal differentiation among actors within a clique (Doreian 1969; Seidman and Foster 1978a, 1978b; Freeman 1992a, 1992b). Since a clique is complete, within the clique all members are graph theoretically identical. All clique members are adjacent to all other clique members, thus there are no distinctions among members based on graph theoretic properties within the clique. If we expect that the cohesive subgroups within a network should exhibit interesting internal structure, such as having some core actors who are more strongly identified with the subgroup and other peripheral actors who are less identified with it, then a clique might be an inappropriate definition of cohesive subgroup.

On the other hand, some researchers working with large network data sets (that include hundreds or even thousands of actors) have found that there may be numerous, but largely overlapping, cliques in the group (Alba and Moore 1978). In such cases, the cliques themselves might not be very informative. Instead, the researcher might study the overlap among the cliques. Studying how cliques overlap is one way to focus on the differentiation or internal structure of subgroups within the network. A recent paper by Freeman (1992b) describes how to use lattices (which we define in Chapter 8) to study the overlap among cliques in social network.

An active area of recent research is the development of methods to extend the definition of cohesive subgroup to make the resulting subgroups more substantively and theoretically interesting. These methods weaken the notion of clique so that the subgroups are less "stingy." There are obviously numerous ways to loosen the definition by removing required properties of a subgraph. These definitions describe subgraphs that are not cliques, but rather, are "clique-like" entities. The "trick" is to develop formal mathematical definitions that have known graph theoretic properties, and also capture important intuitive and theoretical aspects of cohesive subgroups. Two different structural properties have been used to relax the clique notion: first, Luce (1950), and later Alba (1973) and Mokken (1979), have used properties of reachability, path distance, and diameter to extend the clique definition; second, Seidman and Foster (1978a) and Seidman (1981b, 1983b) used nodal degree to propose alternative cohesive subgroup ideas. Both of these ideas take the clique as a starting point, and extend it by removing one or more restrictions. We will describe each of these in turn.

7.3 Subgroups Based on Reachability and Diameter

Reachability is the motivation for the first cohesive subgroup ideas that extend the notion of a clique. These alternative subgroup ideas are useful if the researcher hypothesizes that important social processes occur through intermediaries. For example, the diffusion of information has been hypothesized to occur in this way (Erickson 1988). Conceptually, there should be relatively short paths of influence or communication between all members of the subgroup. Subgroup members might not be adjacent, but if they are not adjacent, then the paths connecting them should be relatively short.

7.3.1 n-cliques

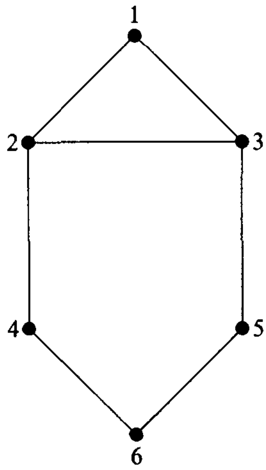
Recall that the geodesic distance between two nodes, denoted by $d(i,j)$, is the length of a shortest path between them. Cohesive subgroups based on reachability require that the geodesic distances among members of a subgroup be small. Thus, we can specify some cutoff value, n, as the maximum length of geodesics connecting pairs of actors within the cohesive subgroup. Restricting geodesic distance among subgroup members is the basis for the definition of an n-clique (Alba 1973; Luce 1950). An *n-clique* is a maximal subgraph in which the largest geodesic distance between any two nodes is no greater than n. Formally, an n-clique is a subgraph with node set $\mathcal{N}_s$, such that

$$d(i,j) \leq n \text{ for all } n_i, n_j \in \mathcal{N}_s \tag{7.1}$$

and there are no additional nodes that are also distance n or less from all nodes in the subgraph.

When $n = 1$, the subgraphs are cliques, since all nodes are adjacent. Increasing the value of n gives subgraphs in which longer geodesic distances between nodes are permitted. A value of $n = 2$ is often a useful cutoff value. 2-cliques are subgraphs in which all members need not be adjacent, but all members are reachable through at most one intermediary.

Let us look at an example to illustrate n-cliques. Figure 7.2, taken from Alba (1973) and Mokken (1979), contains a single clique, $\{1,2,3\}$, which, by definition, is a 1-clique. In this graph, there are two 2-cliques: $\{1,2,3,4,5\}$ and $\{2,3,4,5,6\}$. Notice that these two 2-cliques share four of their five members. In addition, it is important to note that even though we are using a maximum geodesic distance of $n = 2$ to find the 2-cliques, the first 2-clique ($\{1,2,3,4,5\}$) has a diameter of 3. The geodesic between nodes 4 and 5 includes node 6, which is not a member of this 2-clique. *Within* this 2-clique, the shortest path between 4 and 5 is the path 4, 2, 3, 5, which is of length 3. Thus, n-cliques can be found in which the intermediaries in a geodesic between a pair of n-clique members are not themselves n-clique members.



2-cliques: {1, 2, 3, 4, 5} and {2, 3, 4, 5, 6}
2-clan: {2, 3, 4, 5, 6}
2-clubs: {1, 2, 3, 4}, {1, 2, 3, 5}, and {2, 3, 4, 5, 6}

Fig. 7.2. Graph illustrating n-cliques, n-clans, and n-clubs

7.3.2 An Example

Let us return to the example of marriage and business relations among Padgett's Florentine families to illustrate n-cliques. We used the program *GRADAP 2.0* (Sprenger and Stokman 1989) for this analysis. There are thirteen 2-cliques in the marriage relation:

- Acciaiuoli Albizzi Barbadori Medici Ridolfi Salviati Tornabuoni
- Albizzi Bischeri Guadagni Lamberteschi Tornabuoni
- Albizzi Bischeri Guadagni Ridolfi Tornabuoni
- Albizzi Ginori Guadagni Medici
- Albizzi Guadagni Medici Ridolfi Tornabuoni
- Barbadori Castellani Medici Ridolfi Strozzi
- Barbadori Castellani Peruzzi Ridolfi Strozzi
- Barbadori Medici Ridolfi Strozzi Tornabuoni
- Bischeri Castellani Peruzzi Ridolfi Strozzi
- Bischeri Guadagni Peruzzi Ridolfi Strozzi
- Bischeri Guadagni Ridolfi Strozzi Tornabuoni

- Guadagni Medici Ridolfi Strozzi Tornabuoni
- Medici Pazzi Salviati

There are four 2-cliques on the business relation:

- Barbadori Bischeri Castellani Lamberteschi Peruzzi
- Barbadori Castellani Ginori Medici Peruzzi
- Barbadori Ginori Medici Pazzi Salviati Tornabuoni
- Bischeri Castellani Guadagni Lamberteschi Peruzzi

Notice that the 2-cliques are both larger and more numerous than the cliques found for both the marriage and business relations. Since the definition of an *n*-clique is less restrictive than the definition of a clique, when *n* is greater than 1 it is likely that a network will contain more *n*-cliques than cliques. It is also likely that the *n*-cliques will be larger than the cliques.

7.3.3 Considerations

There are several important properties of *n*-cliques, some of which limit the usefulness of this cohesive subgroup definition. Since *n*-cliques are defined for geodesic paths that can include any nodes in the graph, two problems might arise: first, an *n*-clique, as a subgraph, may have a diameter greater than *n*, and second, an *n*-clique might be disconnected. The first problem arises because the requirement that nodes be connected by paths of length *n* or less does not require that these paths remain within the subgroup (Alba 1973; Alba and Moore 1978). Geodesics connecting a pair of nodes in an *n*-clique may include nodes that lie outside of the *n*-clique. Thus, the diameter of the *subgraph* can be larger than *n*. The second problem is that an *n*-clique may not even be connected. Two nodes may be connected by a geodesic of *n* or less which includes nodes outside the *n*-clique, and these two nodes may have no path connecting them that includes only *n*-clique members. These problems indicate that *n*-cliques are not as cohesive as we might like for studying cohesive subgroups (Alba and Moore 1978; Mokken 1979).

7.3.4 n-clans and n-clubs

One idea to "improve" *n*-cliques is to restrict them so that the resulting subgroups that are identified are more cohesive, and do not have the problems of *n*-cliques. A useful restriction is to require that the diameter of an n-clique be no greater than n. Mokken (1979) has described two logical ways to do this. The first, which he calls an *n-clan*, starts with the n-cliques that are identified in a network and excludes those n-cliques that have a diameter greater than n. The second approach, called an *n-club*, defines a new entity, a *maximal n-diameter subgraph.*

An *n-clan* is an n-clique in which the geodesic distance, $d(i, j)$, between all nodes in the subgraph is no greater than n for paths *within* the subgraph. The n-clans in a graph can be found by examining all n-cliques and excluding those that have diameter greater than n. Any n-cliques that include pairs of nodes whose geodesics require non-subgroup members are excluded from consideration. The n-clans in a graph are those n-cliques that have diameter less than or equal to n (Alba 1973; Mokken 1979). All n-clans are n-cliques.

An *n-club* is defined as a maximal subgraph of diameter n. That is, an n-club is a subgraph in which the distance between all nodes *within the subgraph* is less than or equal to n; further, no nodes can be added that also have geodesic distance n or less from all members of the subgraph. n-clubs are not necessarily n-cliques, though they are always subgraphs of n-cliques.

Although conceptually similar, n-clans and n-clubs are somewhat different, as illustrated in Figure 7.2. This example is taken from Alba (1973) and Mokken (1979), and illustrates the difference between n-cliques, n-clans, and n-clubs. For this graph, taking $n = 2$ results in the following sets:

- 2-cliques: $\{1, 2, 3, 4, 5\}$ and $\{2, 3, 4, 5, 6\}$
- 2-clan: $\{2, 3, 4, 5, 6\}$
- 2-clubs: $\{1, 2, 3, 4\}$, $\{1, 2, 3, 5\}$, and $\{2, 3, 4, 5, 6\}$

First, consider the 2-cliques and 2-clans. Since the 2-clique $\{1, 2, 3, 4, 5\}$ has diameter greater than 2 (the distance from 4 to 5 is equal to 3) it is not an 2-clan. The 2-clique $\{2, 3, 4, 5, 6\}$ is a 2-clan since its diameter is not greater than 2. Now, consider the 2-clubs. The 2-clubs $\{1, 2, 3, 4\}$ and $\{1, 2, 3, 5\}$ both have diameter equal to 2, and are maximal, since no node can be added to either subgraph without increasing its diameter. Notice that each of these 2-clubs is a subgraph of the 2-clique $\{1, 2, 3, 4, 5\}$ (whose diameter is greater than 2). Finally, the 2-club $\{2, 3, 4, 5, 6\}$ has a diameter of 2 and is maximal.

As this example illustrates, 2-clubs are either 2-cliques, or are subgraphs of 2-cliques. Mokken (1979) demonstrates that all n-clans are also n-

cliques, and all n-clubs are contained within n-cliques. Furthermore, all n-clans are also n-clubs, though there can be n-clubs that are not n-clans.

As Sprenger and Stokman (1989) have noted, "hardly anybody" has used n-clans and n-clubs, and more research is needed on these cohesive subgroup ideas. The n-clans in a social network are relatively easy to find by examining the n-cliques, and eliminating those with diameter greater than n. The n-clubs are difficult to find, and often routines for n-clubs are not included in standard network analysis packages. Therefore, in the following example we restrict our attention to n-clans.

An Example. We will use the marriage and business relations for Padgett's Florentine families to illustrate n-clans. For the business relation, all of the four 2-cliques have a diameter that is 2 or less, and therefore these four 2-cliques are also 2-clans. For the marriage relation, five of the 2-cliques have diameter greater than 2, so they are excluded from the list of 2-clans. This leaves eight 2-clans:

- Acciaiuoli Albizzi Barbadori Medici Ridolfi Salviati Tornabuoni
- Albizzi Bischeri Guadagni Lamberteschi Tornabuoni
- Albizzi Ginori Guadagni Medici
- Albizzi Guadagni Medici Ridolfi Tornabuoni
- Barbadori Castellani Medici Ridolfi Strozzi
- Bischeri Castellani Peruzzi Ridolfi Strozzi
- Bischeri Guadagni Ridolfi Strozzi Tornabuoni
- Medici Pazzi Salviati

The difference between the 2-cliques and the 2-clans on the marriage relation is that the five 2-cliques with diameter greater than 2 are excluded. For example, the diameter of the 2-clique {Barbadori, Medici, Ridolfi, Strozzi, Tornabuoni} is greater than 2, since the geodesic between Strozzi and Barbadori (which is of length 2) includes Castellani (who is not in this 2-clique).

7.3.5 Summary

The three definitions of cohesive subgroups discussed in this section are primarily motivated by the property of reachability among the nodes in a subgraph. An n-clique simply requires that there is some short path (geodesic) between subgroup members, though this short path may go outside the subgraph. An n-clique may be seen as too loose a definition of cohesive subgroup, and restrictions requiring geodesic paths to remain within the subgroup can be applied by requiring the subgraph to have a given maximum diameter. *n*-clubs and *n*-clans are two possible definitions that have the desired restrictions.

As Erickson (1988) has noted, cohesive subgroup definitions based on reachability are important for understanding "processes that operate through intermediaries, such as the diffusion of clear cut and widely salient information" (Erickson 1988, page 108). In studying network processes such as information diffusion that "flow" through intermediaries, cohesive subgroups based on indirect connections of relatively short paths provide a reasonable approach.

A related cohesive subgroup idea is influence among subgroup members. This idea provides the motivation for Hubbell's (1965) adaptation of economic input-output models to sociometric data. Hubbell argues that ties between actors are "channels for the transmission of influence" (1965, page 377). Influence occurs both through direct contact and through indirect chains of contact via other actors. The goal is to identify subgroups of actors among whom there is a relatively strong mutual influence, whether the influence is direct or indirect. Hubbell's approach relies on measures of influence based on a weighting of adjacencies and paths of influence, and a partitioning of actors based on the degree to which subgroup members mutually influence each other.

In contrast, if one hypothesizes that network processes require direct contact among actors, and perhaps repeated, direct, contact to several actors, then a different cohesive subgroup definition is required. We turn now to subgroup methods that study cohesive subgroups by focusing on adjacency between actors, rather than on paths and geodesics.

7.4 Subgroups Based on Nodal Degree

In this section we describe cohesive subgroup ideas that are based on the adjacency of subgroup members. These approaches are based on restrictions on the minimum number of actors adjacent to each actor in a subgroup. Since the number of actors adjacent to a given actor is quantified by the degree of the node in a graph, these subgroup methods focus on nodal degree. Subgroups based on nodal degree require actors to be adjacent to relatively numerous other subgroup members. Thus, unlike the clique definition that requires all members of a cohesive subgroup to be adjacent to *all* other subgroup members, these alternatives require that all subgroup members be adjacent to some minimum number of other subgroup members.

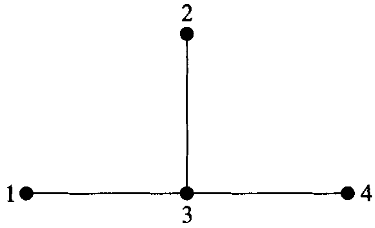


Fig. 7.3. A vulnerable 2-clique

Subgroups based on adjacency between members are useful for understanding processes that operate primarily through direct contacts among subgroup members. For example, Erickson hypothesizes that "multiple redundant channels of communication" will be related to the accuracy of information and the recognizability of subgroups (Erickson 1988, page 108).

These definitions arise in part because of the "vulnerability" of n-cliques. Seidman and Foster (1978) observed that n-cliques often are not robust. One measures robustness by considering "the degree to which the structure is vulnerable to the removal of any given individual" (Seidman and Foster 1978, page 142). Robustness is often assessed using measures of connectivity (see Chapter 4). Robust subgraphs are little affected by the removal of individual nodes. For example, consider the 2-clique in Figure 7.3 consisting of nodes $1, 2, 3$, and 4. Although all pairs of nodes are within path distance 2 of each other, these paths all contain node 3. Node 3 is critical for the connections between other nodes. Furthermore, $1, 2$, and 4 are not connected to each other through any paths that do not contain 3. This 2-clique is vulnerable to the removal of node 3.

The possible lack of robustness of n-cliques was one consideration that led to the proposal of an alternative subgroup definition. This alternative definition, the *k-plex*, builds on the notion that cohesive subgroups should contain sets of actors among whom there are relatively numerous adjacencies (Seidman 1978; Seidman and Foster 1978).

7.4.1 k-plexes

A *k-plex* is a maximal subgraph containing g_s nodes in which each node is adjacent to no fewer than $g_s - k$ nodes in the subgraph. In other words, each node in the subgraph may be lacking ties to no more than k subgraph members. We denote the degree of a node i in subgraph $\mathscr{G}_s$ by $d_s(i)$. A k-plex as a subgraph in which $d_s(i) \geq (g_s - k)$ for all $n_i \in \mathscr{N}_s$ and there are no other nodes in the subgraph that also have $d_s(i) \geq (g_s - k)$. That is, the k-plex is maximal.

Since there are g_s nodes in the subgraph, and we do not consider loops, the degree of a node within the subgraph cannot exceed $g_s - 1$. Thus, if $k = 1$, the subgraph is a clique (the "missing" line is the reflexive line from the node to itself). As k gets larger, each node is allowed more missing lines within the subgraph. Since nodes within a k-plex will be adjacent to many other members, a k-plex is more robust than an n-clique, and removal of a single node is less likely to leave the subgraph disconnected.

Seidman and Foster (1978) discuss properties of k-plexes. An important property of a k-plex is that the diameter of a k-plex is constrained by the value of k. Seidman and Foster prove that in a k-plex of g_s nodes, if $k < (g_s + 2)/2$, then the diameter of $\mathscr{G}_s$ is less than or equal to 2. Thus, if the value of k is small relative to the size of the k-plex, the k-plex will have a small diameter. They also note that if $\mathscr{G}_s$ is a k-plex with g_s nodes, then for any subgraph $\mathscr{G}_k$ of k nodes from $\mathscr{G}_s$, the set of nodes in $\mathscr{G}_k$ plus all nodes in $\mathscr{G}_s$ that are adjacent to the nodes in $\mathscr{G}_k$ constitute the node set of the k-plex $\mathscr{G}_s$. Thus, if you take any subset of k nodes in a k-plex, and then consider these k nodes along with the nodes adjacent to them, then all nodes in the k-plex (from which the subset is drawn) either will be in the original subset of k nodes or will be adjacent to one of these nodes (Seidman and Foster 1978).

An Example. Again, we return to the example of marriage and business relations for Padgett's Florentine families. We used the program *UCINET IV* (Borgatti, Everett, and Freeman 1991) for this analysis. Since 1-plexes are the same as cliques, we will examine the 2-plexes. Also, since k=2 means that two ties may be absent, we will restrict the size of the 2-plexes so that we only consider subgraphs with four or more members. For the marriage relation there are two 2-plexes, involving eight families:

- Albizzi Guadagni Medici Tornabuoni

- Bischeri Castellani Peruzzi Strozzi

Within each of these 2-plexes, each family is missing at most one marriage tie to one of the other families (since two ties can be missing, and one is the undefined reflexive tie). For the business relation there are three 2-plexes, involving six families:

- Barbadori Castellani Lamberteschi Peruzzi
- Bischeri Castellani Lamberteschi Peruzzi
- Bischeri Guadagni Lamberteschi Peruzzi

Notice that for both the marriage and the business relations there are relatively few 2-plexes, compared to fairly numerous 2-cliques.

Considerations. Choosing a useful value of k so that the resulting subgroups are both interesting and interpretable depends in part on the relationship between the sizes of the resulting subgroups and the chosen value of k. If the value of k is large relative to the size of a subgroup, then the k-plex can be quite sparse. For example, a 2-plex of size three might be meaningless, since all three nodes could be missing ties to $k = 2$ other nodes. A 2-plex of size five could also be quite sparse, since each node could have two lines present and two lines absent, and still meet the 2-plex requirement. Therefore, in practice the researcher should restrict the size of a k-plex so that it is not too small relative to the number of ties that are allowed to be missing.

7.4.2 k-cores

Another approach to cohesive subgroups based on nodal degree is the *k-core* (Seidman 1983b). A k-core is a subgraph in which each node is adjacent to at least a minimum number, k, of the other nodes in the subgraph. In contrast to the k-plex, which specifies the acceptable number of lines that can be *absent* from each node, the k-core specifies the required number of lines that must be *present* from each node to others within the subgraph. As before, we define the degree of node i within a subgraph, $d_s(i)$, as the number of nodes within the subgraph that are adjacent to i. We then define a k-core in terms of minimum nodal degree *within* the subgraph. A subgraph, $\mathscr{G}_s$, is a k-core if

$$d_s(i) \geq k \text{ for all } n_i \in \mathscr{N}_s.$$

A k-core is thus defined in terms of the minimum degree within a subgraph, or the minimum number of adjacencies that must be present. Seidman (1983b) notes that although k-cores themselves are not necessarily interesting cohesive subgroups, they are "areas" of a graph in which other interesting cohesive subgroups will be found.

7.5 Comparing Within to Outside Subgroup Ties

The three general cohesive subgroup approaches discussed so far in this chapter are based on properties of ties within the subgroup (adjacency, geodesic distance, or number of ties among subgroup members). However, as Seidman notes, cohesive subgroups "...in social networks have usually been seen informally as sets of individuals more closely tied to each other than to outsiders" (1983a, page 97). Thus, the intuitive notion of cohesive subgroup derives both from the relative strength, frequency, density, or closeness of ties within the subgroup, and the relative weakness, infrequency, sparseness, or distance of ties from subgroup members to nonmembers (Bock and Husain 1950; Alba 1973; Seidman 1983a; Sailer and Gaulin 1984; Freeman 1992a).

As Alba (1973) has noted, there are at least two different aspects to the concept of a cohesive subgroup: the concentration of ties within the subgroup, and a comparison of strength or frequency of ties within the subgroup to the strength or frequency of ties outside the subgroup. Alba has referred to the comparison of within to between subgroup ties as the "centripetal-centrifugal" dimension of cohesive subgroups. This idea has led to subgroup definitions that compare the prevalence of ties within the subgroup to the sparsity of ties outside the subgroup (Alba 1973; Bock and Husain 1950; Freeman n.d.; Sailer and Gaulin 1984). In this section we describe methods for analysis of subgroups based on comparison of ties within the subgroup to ties outside the subgroup.

The fourth cohesive subgroup idea is that cohesive subgroups should be relatively cohesive within compared to outside. Thus, instead of concentrating simply on properties of the ties among members within the subgroup, it is necessary to compare these to properties of ties to actors outside the subgroup.

It will be useful to define some additional graph properties before we describe these methods. Recall that a graph $\mathscr{G}$ consists of a set of nodes $\mathscr{N}$, and a set of lines $\mathscr{L}$. To start we will restrict our attention to dichotomous, undirected graphs. We will be interested in subsets of nodes $\mathscr{N}_s \subseteq \mathscr{N}$, and the subgraph $\mathscr{G}_s$ induced by node set $\mathscr{N}_s$. In addition, we can denote the subset of nodes that are in $\mathcal{N}$ but not in $\mathcal{N}_s$ as $\mathcal{N}_t = \mathcal{N} - \mathcal{N}_s$. $\mathcal{N}_t$ and $\mathcal{N}_s$ are mutually exclusive and exhaustive subsets. Now, there are three sets of lines in the graph: lines between nodes within the subset $\mathcal{N}_s$, lines between nodes in $\mathcal{N}_s$ and nodes in $\mathcal{N}_t$, and lines between nodes within $\mathcal{N}_t$. There are g nodes in $\mathcal{N}$, g_s nodes in $\mathcal{N}_s$, and $g_t = g - g_s$ nodes in $\mathcal{N}_t$. There are $g(g-1)/2$ possible lines in the entire graph, $g_s(g_s - 1)/2$ possible lines within $\mathcal{N}_s$, and $(g_s \times g_t)/2$ possible lines between members of $\mathcal{N}_s$ and "outsiders" belonging to $\mathcal{N}_t$.

Let us first consider an "ideal" type of subgraph which exhibits the most extreme realization of a cohesive subgroup in which there are ties within the subgroup but not between subgroup members and outsiders (Freeman n.d.). Such an ideal subgroup would consist of ties between *all* pairs of members within the subgroup, and *no* ties from subgroup members to actors not in the subgroup. In graph theoretic terms, such a subgraph is a *complete component* of the graph. All nodes in a complete component are adjacent, and there are no nodes outside the subgraph that are adjacent to any node in the component. Freeman has called such a subgraph a *strong alliance*. A strong alliance is also a clique, since it is complete and maximal. But, a strong alliance is a stricter subgroup definition than is a clique. There are many cliques that are not strong alliances.

A strong alliance is a stricter subgroup definition than a clique and is clearly too restrictive for data analytic purposes. However, there are natural graph theoretic relaxations of the strong alliance that define useful cohesive subgroup methods. Also a strong alliance provides a formal standard against which to compare observed cohesive subgroups to assess their cohesiveness.

7.5.1 LS Sets

An *LS* set is a subgroup definition that compares ties within the subgroup to ties outside the subgroup by focusing on the greater frequency of ties among subgroup members compared to the ties from subgroup members to outsiders (Luccio and Sami 1969; Lawler 1973; Seidman 1983a; Borgatti, Everett, and Shirey 1990). Seidman defines an *LS* set as follows:

> a set of nodes *S* in a social network is an *LS* set if each of its proper subsets has more ties to its complement within *S* than to the outside of *S*. (Seidman 1983a, page 98)

Consider the subgraph $\mathcal{G}_s$ with node set $\mathcal{N}_s$, and the subsets of nodes that can be taken from $\mathcal{N}_s$. We will define a subset of nodes taken from $\mathcal{N}_s$ as $\mathcal{Q}$, so that $\mathcal{Q} \subset \mathcal{N}_s$. The set of nodes, $\mathcal{N}_s$, is an *LS* set if any proper subset $\mathcal{Q} \subset \mathcal{N}_s$ has more lines to the nodes in $\mathcal{N}_s - \mathcal{Q}$ (other nodes in the subset) than to $\mathcal{N} - \mathcal{N}_s$ (nodes outside the subset) (see Seidman 1983a, page 97).

The definition of an *LS* set compares the frequency of ties within and between subsets. There are three basic sets to consider: $\mathcal{Q} \subset \mathcal{N}_s \subseteq \mathcal{N}$. The set $\mathcal{Q}$ is a "wild card" that stands for any possible subset of nodes that can be selected from $\mathcal{N}_s$ (the potential *LS* set). Next there are two additional sets that consist of nodes in one of these three sets but not in another: $\mathcal{N} - \mathcal{N}_s$ and $\mathcal{N}_s - \mathcal{Q}$. There are two kinds of lines to consider: lines from $\mathcal{Q}$ to $\mathcal{N}_s - \mathcal{Q}$ and lines from $\mathcal{Q}$ to $\mathcal{N} - \mathcal{N}_s$. Lines within the *LS* set, $\mathcal{N}_s$ (that is, from any subset of the nodes in the *LS* set to remaining *LS* set members), should be more numerous than lines from a subset of nodes in an *LS* set to non-*LS* set members.

Seidman (1983a) and Borgatti, Everett, and Shirey (1990) have described several important properties of *LS* sets. First, since all subsets of the *LS* set have more ties within than outside the subset, they are relatively robust, and do not contain "splinter" groups. This leads Borgatti, Everett, and Shirey (1990) to hypothesize that *LS* sets in a network will be relatively stable through time. An important relationship between the *LS* sets in a given graph is that any two *LS* sets either are disjoint (share no members) or one *LS* set contains the other (Borgatti, Everett, and Shirey 1990). Unlike cliques, *n*-cliques, and *k*-plexes, *LS* sets cannot overlap by sharing some but not all members. The fact that *LS* sets are related by containment means that within a graph there is a hierarchical series of *LS* sets.

7.5.2 Lambda Sets

Recently, Borgatti, Everett, and Shirey (1990) have extended the notion of an *LS* set. Their approach, which they call a *lambda set*, is motivated by the idea that a cohesive subset should be relatively robust in terms of its *connectivity*. That is, a cohesive subset should be hard to disconnect by the removal of lines from the subgraph. The extent to which a pair of nodes remains connected by some path, even when lines are deleted from the graph, is quantified by the *edge connectivity* or *line connectivity* of the pair of nodes (see Chapter 4). The line connectivity of nodes i and j, denoted $\lambda(i, j)$, is equal to the minimum number of lines that must be removed from the graph in order to leave no path between the two nodes. The line connectivity of two nodes is also equal to the number of paths between them that contain no lines in common (the number of line-disjoint or line-independent paths). The smaller the value of $\lambda(i, j)$, the more vulnerable i and j are to being disconnected by removal of lines. The larger the value of $\lambda(i, j)$, the more lines must be removed from the graph in order to leave no path between i and j.

Using the notion of line connectivity, Borgatti, Everett, and Shirey (1990) define a *lambda set*. The logic of the definition of a lambda set is similar to the definition of an *LS* set. Consider pairs of nodes in the subgraph $\mathcal{G}_s$, with node set $\mathcal{N}_s$. The set of nodes, $\mathcal{N}_s$, is a lambda set if any pair of nodes in the lambda set has larger line connectivity than any pair of nodes consisting of one node from within the lambda set and a second node from outside the lambda set. Formally, a *lambda set* is a subset of nodes, $\mathcal{N}_s \subseteq \mathcal{N}$, such that for all $i, j, k \in \mathcal{N}_s$, and $l \in \mathcal{N} - \mathcal{N}_s$, $\lambda(i, j) > \lambda(k, l)$.

Since high values of λ require high line connectivity within the lambda set, successively increasing values of λ gives rise to a series of lambda sets in a given network. These lambda sets do not overlap unless one lambda set is contained within another. An advantage of lambda sets is that they are more general than *LS* sets. Any *LS* set in a network will be contained within a lambda set, and a given network is more likely to contain lambda sets than it is to contain *LS* sets (Borgatti, Everett, and Shirey 1990).

One important property of lambda sets is that nodes within a lambda set are not necessarily cohesive in terms of either adjacency or geodesic distance, the two properties that are the basis for other kinds of cohesive subsets that we have discussed. Members of a lambda set do not need to be adjacent, and since there is no restriction on the length of paths that connect nodes within a lambda set, members of a lambda set may be quite distant from one another in the graph (Borgatti, Everett, and Shirey 1990).

So far we have described formal definitions of cohesive subgroups. Now we turn to some measures of how cohesive a subgroup is.

7.6 Measures of Subgroup Cohesion

Several researchers have proposed measures for the extent to which ties are concentrated within a subgroup, rather than between subgroups (Bock and Husain 1950; Alba 1973; Sailer and Gaulin 1984; Freeman n.d.). These measures are primarily descriptive, although Alba presents a probability model for his measure. The problem of assessing the "goodness" of an assignment of actors to cohesive subgroups within a network is related to issues we discuss in Chapter 16, under the topic of goodness-of-fit indices. In this section we present some descriptive measures, and leave the statistical approaches for later, after we have developed the necessary background (in Chapters 13 and 15).

Bock and Husain (1950) proposed that one way to search for cohesive subgroups in a social network is iteratively to construct subgroups so that the ratio of the strength of ties within the subgroup to ties between subgroups does not decrease appreciably with the addition of new members. They note the similarity of this analytic problem to the analysis of sets of test items to identify subsets of highly correlated items. If there are g members in the whole network, and g_s members in a subgroup $\mathcal{N}_s$, then a measure of the degree to which strong ties are within rather than outside the subgroup is given by the ratio:

$$\frac{\frac{\sum_{i\in\mathcal{N}_s}\sum_{j\in\mathcal{N}_s} x_{ij}}{g_s(g_s-1)}}{\frac{\sum_{i\in\mathcal{N}_s}\sum_{j\notin\mathcal{N}_s} x_{ij}}{g_s(g-g_s)}}. \tag{7.2}$$

The numerator of this ratio is the average strength of ties within the subgroup and the denominator is the average strength of the ties that are from subgroup members to outsiders. For a dichotomous relation the numerator is the density of the subgroup. For a valued relation the numerator is the average strength of ties within the subgroup. If the ratio is equal to 1, then the strength of ties does not differ within the subgroup as compared to outside the subgroup. If the ratio is greater than 1, then the ties within the subgroup are more prevalent (or stronger) on average than are the ties outside the subgroup. Bock and Husain suggest that cohesive subgroups of actors can be constructed by successively adding members to an existing subgroup, so long as the additional members do not greatly decrease the value of this ratio.

As we mentioned above, Alba (1973) views the measure in equation (7.2) in terms of two separate components. The numerator is a measure of the cohesiveness of a subgroup, and the denominator is a measure of sparsity of ties to actors outside the subgroup. Alba calls these the "centripetal" and "centrifugal" properties, respectively. Further, he presents formulas for the probability of obtaining the density of a subgroup equal to or greater than the observed density, given the density of the graph.

Alba (1973) uses the hypergeometric probability function to calculate the probability of observing exactly L_s lines in a subgraph of g_s nodes, taken from a graph with g nodes and L lines. Equivalently, this is the probability of drawing a random sample without replacement of $g_s(g_s - 1)/2$ dyads (the number of dyads within the subgroup) and observing exactly L_s ties present, from a graph of $g(g-1)/2$ dyads and $L = x_{++}/2$ ties. The probability that the observed number of lines in the subgraph is equal to q is given by the following hypergeometric probability (Alba 1973, page 122):

$$P(L_s = q) = \frac{\binom{L}{q}\binom{\frac{g(g-1)}{2} - L}{\frac{g_s(g_s-1)}{2} - q}}{\binom{\frac{g(g-1)}{2}}{\frac{g_s(g_s-1)}{2}}}. \tag{7.3}$$

Equation (7.3) is the probability of obtaining exactly q lines in the subgraph. The probability that we are interested in is the probability of *q or more* lines; that is, the probability of a subgraph that is as dense or denser than the one we observe. Thus, we must sum the probabilities from equation (7.3) for values of q from L_s, the observed number of lines in the subgraph, to its maximum possible, which is either $g_s(g_s - 1)/2$, the possible number of lines that could be present in the subgraph, or $L = x_{++}/2$, the observed number of lines in the graph, whichever is less. The formula for the probability of observing q or more lines in a subgraph of size g_s from a graph with L lines is:

$$P(L_s \geq q) = \sum_{k=q}^{\min(L, \frac{g_s(g_s-1)}{2})} \frac{\binom{L}{k}\binom{\frac{g(g-1)}{2} - L}{\frac{g_s(g_s-1)}{2} - k}}{\binom{\frac{g(g-1)}{2}}{\frac{g_s(g_s-1)}{2}}}. \tag{7.4}$$

If the calculated probability in equation (7.4) is small, then the observed frequency of lines within the subgraph is greater than would be expected by chance, given the frequency of lines in the graph as a whole. Thus, this probability can be interpreted as a p-value for the null hypothesis that there is no difference between the density of the subgraph and the density of the graph as a whole.

Freeman (n.d.) provides another approach to measuring the cohesiveness of a subgroup. Freeman's measure is based on his model of strict alliances (see discussion above) and the extent to which a given subgroup approaches that strictly defined property. Sailer and Gaulin (1984) discuss several alternative measures of cohesiveness of a subgroup, depending on how one conceptualizes the concentration of interactions within as opposed to outside the subgroup.

So far we have described cohesive subgroup methods for dichotomous nondirectional relations. We now discuss extensions of cohesive subgroups to relations that are valued or directional. These extensions allow the cohesive subgroup ideas discussed in the previous sections to be applied to a much wider range of social network data.

7.7 Directional Relations

Cohesive subgroup ideas can be extended to directional relations. We will continue to restrict our attention to dichotomous relations. Recall that a directional relation is one in which a tie has an origin and a destination. A directional relation can be represented as a directed graph. An arc in the directed graph is present from i to j if $i \rightarrow j$, or equivalently, if i "chooses" j. In a sociomatrix for a directional relation x_{ij} might not equal x_{ji}.

There are several ways to define cohesive subgroups for directional relations. The most straightforward way is to consider only the reciprocated ties that are present in the graph (Festinger 1949; Luce and Perry 1949; Luce 1950). More generally, it is possible to define properties of connectedness for directional relations, and then use these properties to define cohesive subgroups for directional relations. We will discuss each approach in turn.

7.7.1 Cliques Based on Reciprocated Ties

Recall that the definition of a clique originally proposed by Festinger (1949) and Luce and Perry (1949) focused on directional affective relations and required that all ties between all pairs of clique members be reciprocated. Thus, cliques can be found in a directional relation by focusing only on those ties that are reciprocated ($x_{ij} = x_{ji} = 1$). In analyzing a directional relation, this is equivalent to symmetrizing the sociomatrix by taking the minimum of the entries in corresponding off-diagonal cells. More precisely, we can define a new nondirectional relation, $\mathscr{X}^{min}$, where

$$x_{ij}^{min} = x_{ji}^{min} = \begin{cases} 1 & \text{if } x_{ij} = x_{ji} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The new relation $\mathscr{X}^{min}$ contains only ties that are reciprocated ($x_{ij} = x_{ji} = 1$) or null ($x_{ij} = x_{ji} = 0$). The sociomatrix representation for this relation is symmetric. The relation $\mathscr{X}^{min}$ can then be analyzed using methods for finding cliques or other cohesive subgroups in a nondirectional relation. However, if there are few reciprocated ties the resulting symmetric relation will be quite sparse, and might not yield many cohesive subgroups.

An Example. As an example of a clique analysis of a dichotomous directional relation we will consider the Friendship relation for Krackhardt's high-tech managers. Recall that each manager was asked, "Who are your friends?" Thus, a friendship tie is directed from one manager to another, and friendship choices need not be reciprocated. To find cliques (subsets of actors among whom all choices are reciprocated), it is necessary to analyze only those ties that are reciprocated. This is accomplished by symmetrizing the sociomatrix as described above. We analyzed the symmetrized sociomatrix for the friendship relation using *UCINET IV* (Borgatti, Everett, and Freeman 1991). There are six cliques, containing nine of the managers.

- 1 4 12
- 4 12 17
- 5 11 17
- 5 11 19
- 11 15 19
- 12 17 21

Notice that these cliques are small, containing only the minimum three members, and there is considerable overlap among them.

7.7.2 Connectivity in Directional Relations

A more flexible way to extend cohesive subgroup ideas to directional relations uses definitions of semipaths and connectivity for directed graphs. These ideas generalize the definitions of path, path distance, and connectivity from graphs to directed graphs, and were defined in Chapter 4. We will begin by briefly reviewing the two kinds of paths for digraphs and then use these kinds of paths to describe four ways to extend the notion of connectivity and *n*-cliques to directed graphs (Harary, Norman, and Cartwright 1965; Peay 1975a, 1980).

Recall that a *path* from node i to node j in a directed graph is a sequence of distinct nodes, where each arc has its origin at the previous node and its terminus at the subsequent node. Thus, a path in a directed graph consists of arcs all "pointing" in the same direction. The length of a path is the number of arcs in it. A *semipath* from node i to node j is a sequence of distinct nodes, where all successive pairs of nodes are connected by an arc from the first to the second, or by an arc from the second to the first. In a semipath the direction of the arcs is irrelevant. The length of a semipath is the number of arcs in it.

There are four different ways in which two nodes can be connected by a path, or semipath, of n arcs or fewer. Our definitions come from Peay (1980, pages 390–391). A pair of nodes, i, j, is:

(i) *Weakly n-connected* if they are joined by a *semipath* of length n or less

(ii) *Unilaterally n-connected* if they are joined by a *path* of length n or less from i to j, *or* a *path* of length n or less from j to i

(iii) *Strongly n-connected* if there is a *path* of length n or less from i to j, *and* a path of length n or less from j to i; the path from i to j may contain different nodes and arcs than the path from j to i

(iv) *Recursively n-connected* if they are strongly n-connected, and the path from i to j uses the same nodes and arcs as the path from j to i, in reverse order

These are increasingly strict connectivity definitions. A pair of nodes connected by a stricter kind of connectivity is also connected by weaker kinds.

7.7.3 *n-cliques in Directional Relations*

It is now possible to define four different kinds of cohesive subgroups based on the four types of connectivity (see Peay 1975, 1980). In each case, a cohesive subgroup is defined as a subgraph of three or more nodes that is maximal with respect to the specified property. The property is the kind of connectivity between the nodes in the subgraph. Since there are four kinds of connectivity in a directed graph, there are four definitions of cohesive subgroups. These are natural extensions of the definition of an n-clique described above.

(i) A *weakly connected n-clique* is a subgraph in which all nodes are weakly n-connected, and there are no additional nodes that are also weakly n-connected to all nodes in the subgraph.

(ii) A *unilaterally connected n-clique* is a subgraph in which all nodes are unilaterally n-connected, and there are no additional nodes that are also unilaterally n-connected to all nodes in the subgraph.

(iii) A *strongly connected n-clique* is a subgraph in which all nodes are strongly n-connected, and there are no additional nodes that are also strongly n-connected to all nodes in the subgraph.

(iv) A *recursively connected n-clique* is a subgraph in which all nodes are recursively n-connected, and there are no additional nodes that are also recursively n-connected to all nodes in the subgraph.

As with the definitions of connectivity, these are increasingly strict cohesive subgroup definitions.

Finding some kinds of n-cliques in directional dichotomous relations is straightforward. Finding weakly connected n-cliques and recursively connected n-cliques requires symmetrizing the relation using the appropriate rule, and then using a standard n-clique algorithm. Since weakly connected n-cliques require a semipath of length n or less between all members, the direction of the arcs in the semipath is irrelevant. Thus, we can construct a symmetric relation, $\mathscr{X}^{max}$, with values x_{ij}^{max}, in which a tie is present from i to j if either $i \rightarrow j$ or $j \rightarrow i$. The relation $\mathscr{X}^{max}$ is defined as:

$$x_{ij}^{max} = x_{ji}^{max} = \begin{cases} 1 & \text{if either } x_{ij} = 1 \text{ or } x_{ji} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The n-cliques in $\mathscr{X}^{max}$ are the weakly connected n-cliques in $\mathscr{X}$.

Recursively connected n-cliques require not only a path of length n or less between all pairs of members, but the paths must contain exactly the same nodes in the reverse order. Thus, one must only consider arcs in both directions. In order to find recursively connected n-cliques, one can construct a symmetric relation, $\mathscr{X}^{min}$ (as defined above), in which a tie in $\mathscr{X}^{min}$ is present only if both $x_{ij} = 1$ and $x_{ji} = 1$. The n-cliques in $\mathscr{X}^{min}$ are the recursively connected n-cliques in $\mathscr{X}$.

An Example. To illustrate n-cliques for a dichotomous directional relation we will use the friendship relation from Krackhardt's high-tech managers. We only present the recursively connected 2-cliques and the weakly connected 2-cliques, which, as we have discussed above, can be found by using an appropriately symmetrized sociomatrix, and a usual *n*-clique program (for example, in *GRADAP* or *UCINET IV*).

There are eight recursively connected 2-cliques:

- 1 2 4 12 16
- 1 2 4 12 17 21
- 1 2 18 21
- 1 4 8 12 17
- 3 5 11 15 19
- 4 5 6 11 12 17 21
- 5 11 13 15 17 19
- 11 14 15 19

Seventeen of the twenty-one managers belong to at least one of the recursively connected 2-cliques. Since managers 7 and 9 have outdegrees equal to 0 on this relation (they did not choose anyone on the friendship relation), they cannot belong to either recursively or strongly connected *n*-cliques.

There are four weakly connected 2-cliques:

- 1 2 3 4 5 8 9 10 11 12 15 17 19 20 21 16 6 14 7
- 1 2 3 4 5 8 9 10 11 12 15 17 19 20 21 16 6 18
- 1 2 3 4 5 8 9 10 11 12 15 17 19 20 21 13 18
- 1 2 3 4 5 8 9 10 11 12 15 17 19 20 21 13 14

All of the twenty-one managers belong to at least one of the four weakly connected 2-cliques, and the vast majority (fifteen of the twenty-one managers) belong to all of them. Clearly, these weakly connected 2-cliques are not very cohesive.

7.8 Valued Relations

Relations are often valued. Valued relations indicate the strength or intensity of ties between pairs of actors. For instance, social network data can be collected by having each person indicate their degree of "liking for" or "acquaintance with" each other person in a group using a five point rating scale. Or, one could record the number of social occasions at which each pair of actors were both present. Cohesive subgroups in valued relations focus on subsets of actors among whom ties are strong or frequent; thus, ties among subgroup members should have high values.

Valued relations are represented as valued graphs. A valued graph, $\mathscr{G}(\mathscr{N}, \mathscr{L}, \mathscr{V})$, consists of a set of nodes, $\mathscr{N}$, a set of lines, $\mathscr{L}$, and a set of values, $\mathscr{V}$, indicating the strength of each line. The value attached to a line codes the strength of the tie between the pair of actors. A valued relation can be represented as a sociomatrix where x_{ij} is the value of the tie from actor i to actor j. We will assume that measurements on the valued relation are at least ordinal, and take on C values, such that $0 \leq x_{ij} \leq C-1$, for all i and j. The highest possible value indicates the strongest tie between any pair of actors. Smaller values of x_{ij} indicate weaker ties. Thus, since the relation is assumed to be at least ordinal, if $x_{ij} < x_{kl}$, the tie from i to j is weaker than the tie from k to l. For simplicity we will limit our attention to nondirectional valued relations. In a nondirectional valued relation the strength of the tie from actor i to actor j is the same as the strength of the tie from actor j to actor i. If the relation is nondirectional $x_{ij} = x_{ji}$ for all i and j, and the sociomatrix is symmetric.

In general, a cohesive subgroup of actors in a valued network is a subset of actors among whom ties have high values. Thus, if we consider the values attached to the ties among subgroup members, these values should be relatively high. Since the values of the ties range from 0 (indicating the weakest possible tie) to $C-1$ (indicating the strongest tie), more cohesive subgroups will have ties with values close to $C-1$ whereas less cohesive subgroups will have ties with values lower than $C-1$. Thus, in a valued relation we can study cohesive subgroups that vary in the strength of ties among members.

In studying cohesive subgroups in valued relations we will consider a threshold value, c, for the value of ties within the subgroup. By increasing (or decreasing) the threshold value we can find more (or less) cohesive subgroups. Since the values of the ties range from 0 to $C-1$, the threshold value c can take on values between 0 and $C-1$.

We will now define a clique, n-clique, and k-plex for a valued relation. We then describe how valued relations can be analyzed to study these cohesive subgroups. Further discussion of cliques and related ideas for valued relations can be found in Doreian (1969) and Peay (1974, 1975a, 1980).

7.8.1 Cliques, n-cliques, and k-plexes

Let us first define a *clique at level c*. A clique at level c is a subgraph in which the ties between all pairs of actors have values of c or greater, and to the specified value, c. In this new dichotomous relation ties are present among all pairs of actors who have ties in $\mathscr{X}$ with values of c or greater. We denote this new, derived dichotomous relation as $\mathscr{X}^{(c)}$, with sociomatrix $\mathbf{X}^{(c)} = \{x_{ij}^{(c)}\}$, where

$$x_{ij}^{(c)} = \begin{cases} 1 & \text{if } x_{ij} \geq c \\ 0 & \text{otherwise.} \end{cases}$$

For any valued relation we can define an increasingly strict series of cutoff values, c, that spans the range of values $0 \leq c \leq C - 1$. Each value of c defines a dichotomous relation and its corresponding graph and sociomatrix. With larger values of c, ties are present in $\mathscr{X}^{(c)}$ only if there is a relatively strong tie between actors in $\mathscr{X}$. Thus for larger values of c, the relation $\mathscr{X}^{(c)}$ may be fairly sparse. For small values of c, a tie is present in $\mathscr{X}^{(c)}$ even if the strength of the tie in $\mathscr{X}$ is relatively low, and thus this relation can be fairly dense. In fact, a cutoff value of $c = 0$ results in a complete relation (and a complete graph) since all defined values of $\mathbf{X}^{(0)}$ will be equal to unity. Thus, in practice, there are $C - 1$ nontrivial graphs that can be defined from a valued graph with C levels. It is important to note that for two cutoff values, c and c', with $c' < c$, all of the ties present in $\mathscr{X}^{(c)}$ will also be present in $\mathscr{X}^{(c')}$; in other words, $\mathscr{X}^{(c')}$ "includes" $\mathscr{X}^{(c)}$.

We now illustrate cliques in a valued relation using a hypothetical valued graph and the dichotomous relations that can be derived from it.

An Example. Figure 7.4 presents the sociomatrix for a valued relation and the graphs that can be derived from this relation. The values of the relation range from 0 to $C - 1 = 5$. Thus, there are five possible nontrivial graphs that can be derived from this valued relation using increasingly strict cutoff values.

Consider the cliques that may be present in each derived graph, starting from the strictest cutoff value, $c = 5$. At the strictest cutoff, $c = 5$, there are no cliques. As c decreases there are more, and larger, cliques in the derived graphs. The results of a clique analysis of each of the five derived graphs are:

- $c = 5$: no cliques
- $c = 4$: $\{1, 2, 3\}$
- $c = 3$: $\{1, 2, 3\}$
- $c = 2$: $\{1, 2, 3\}$ and $\{3, 4, 5\}$
- $c = 1$: $\{1, 2, 3, 4\}$ and $\{3, 4, 5\}$

	1	2	3	4	5
1	-	5	4	1	0
2	5	-	4	1	0
3	4	4	-	3	2
4	1	1	3	-	4
5	0	0	2	4	-

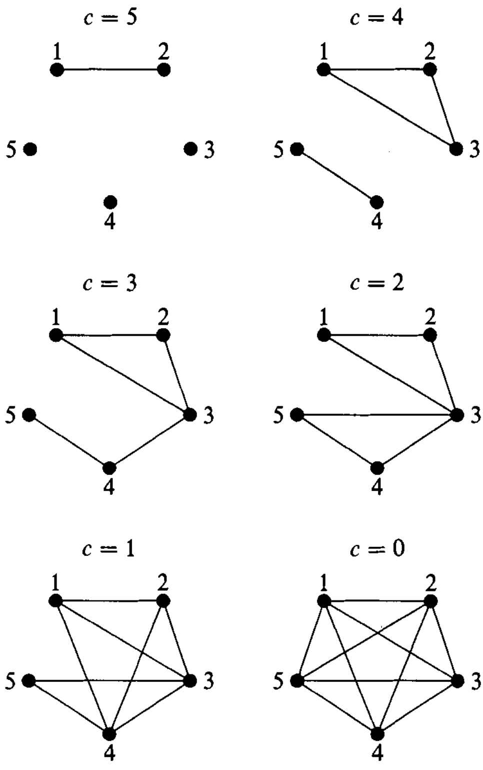


Fig. 7.4. A valued relation and derived graphs

Notice that the clique containing nodes 1, 2, and 3 that occurs at $c = 4$ continues to be a clique, or is subsumed within a larger clique, at all less stringent values of c.

In general, every derived dichotomous relation defines a graph that can be analyzed using methods for finding cohesive subgroups, described above. For example, one can analyze each of the graphs derived from a valued relation and study the cliques, k-plexes, or n-cliques that may exist in each of the graphs. Each of the $C - 1$ graphs may (or may not) contain cohesive subgroups.

Actors among whom ties have large values can appear in cohesive subgroups at strict cutoff levels, whereas actors among whom ties have small values can only appear in cohesive subgroups at less strict cutoff levels. As Doreian (1969) notes, analyzing the derived graphs for increasingly stringent values of c results in a hierarchical series of cohesive subgroups. This hierarchical series allows one to study the internal structure of cohesive subgroups.

The approaches presented in this section generalize cohesive subgroup ideas that were initially developed for dichotomous relations and apply them to valued relations. Thus, the definitions of clique, n-clique, and k-plex remain the same, but are applied to dichotomous relations derived from the valued relation. An alternative approach for studying cohesive subgroups in valued relations is to define cohesive subgroup ideas specifically for valued relations (Freeman 1992a).

7.8.2 Other Approaches for Valued Relations

In a recent paper, Freeman (1992a) reviews sociological approaches to the concept of social "group" and discusses formalizations of this concept using data on frequency of interactions among people in naturally occurring communities. Data on interaction frequencies give rise to valued relations. Freeman's argument, expanding on ideas presented by Winship (1977) and Granovetter (1973), is that membership in a "group" should be characterized by relatively frequent face-to-face interactions among members. Specifically, if actors i, j, and k are members of a "group," then if i and j interact frequently, and j and k interact frequently, then i and k should have at least some amount of interaction. This idea of cohesiveness of subgroups builds on Granovetter's (1973) ideas of strong and weak ties, and extends the ideas of transitivity and clusterability to valued relations. Advantages of this approach are that the resulting cohesive subgroups form a hierarchical series, and different subgroups do not overlap unless one subgroup is fully contained within another.

7.9 Interpretation of Cohesive Subgroups

The result of a cohesive subgroup analysis is a list of subsets of actors within the network who meet the specified subgroup definition. For example, the result of a clique analysis is a list of the cliques in the network and the actors who belong to each clique. For a given analysis it might be the case that no subsets of actors meet the specified subgroup definition (for example, it might be that there are no cliques in a given network), or it might be the case that there are numerous subsets of actors that meet the specified subgroup definition (for example, the *n*-clique analysis of the marriage relation among Padgett's Florentine families resulted in thirteen 2-cliques). In any case, the researcher must interpret the results of the analysis. In this section we discuss three levels at which one might interpret the results of a cohesive subgroup analysis. These levels are the:

(i) Individual actor
(ii) Subset of actors
(iii) Whole group

In terms of individual actors, the simplest distinction is between actors who belong to one or more cohesive subset(s), and actors who do not belong to any cohesive subset. Thus, we can make a distinction between "members" and "non-members." One can then relate this distinction to other actor characteristics, for example, by studying whether subgroup "members" differ from "non-members" in theoretically important ways. It could also be the case that "non-members" occupy critical locations between groups; for example, they might have high betweenness centrality. The network analysis program *NEGOPY* (Richards 1989a) uses a similar distinction to describe types of actors in a network.

The result of a cohesive subgroup analysis can also be interpreted in terms of the characteristics of the members of the subsets. If the network data set contains information on attributes of the actors, then one can use these attributes to describe the subsets. For example, it might be the case that members of the same subgroup are more similar to each other than they are to outsiders. This method of interpretation was used by Alba and Moore (1978) to describe the composition of subgroups of elite decision makers.

Finally, the result of a cohesive subgroup analysis can be used to describe the network as a whole. Consider two quite different ways that a network might be organized. On the one hand, a network could be a single cohesive set. On the other hand, the network could be "fragmented" into two or more subgroups. In the first case, cohesive subgroups within the network would be largely overlapping, and would contain most of the actors in the network. We saw this pattern for the *n*-clique analysis of Friendship among Kackhardt's high-tech managers. In the second case, fragmentation of the network would show up as two or more cohesive subgroups that did not share members in common. Hence, the numbers of actors in the subgroups and the degree to which these subgroups overlap can be used to describe the structure of the network as a whole.

7.10 Other Approaches

All of the cohesive subgroup ideas discussed in the previous sections define specific graph theoretic properties that should be satisfied in order to identify a subset of actors as a cohesive subgroup. For all of these approaches, the analytic problem is to examine a set of social network data to see whether any subsets of actors meet the specified subgroup definition. The result is the possible assignment of actors to one or more cohesive subgroups. An alternative, and more exploratory, approach to cohesion in social networks seeks to represent the group structure in a network as a whole. Collections of actors among whom there are relatively strong ties can become more visible by displaying functions or rearrangements of the graphs or sociomatrices. We now describe these approaches.

7.10.1 Matrix Permutation Approaches

The earliest contributions to cohesive subgroup analysis of social networks were concerned with systematic ways for ordering rows and columns of a sociomatrix to reveal the subgroup structure of a network (Forsyth and Katz 1946; Katz 1947). The subgroup structure is seen in the relative prevalence (or sparsity) of ties among some subsets of actors. An informative sociomatrix should make this subgroup structure readily apparent. If there are subgroups of actors in a network who tend to choose each other and tend not to choose actors outside their subgroup, then it is very useful to rearrange the rows and columns of the sociomatrix so that actors in the same subgroup occupy rows (and columns) that are close to one another in the sociomatrix. Thus, there might be some "preferred ordering" of the rows and columns of the sociomatrix that would best reveal the structure of the group (Katz 1947). If one had objective criteria for this ordering, then different researchers could construct the same preferred sociomatrix. One could then inspect the rearranged sociomatrix and identify subgroups of actors among whom there are prevalent or strong ties.

An important property of a good ordering of a sociomatrix is that subsets of actors who have strong ties to each other should occupy adjacent rows (and columns), or at least should occupy rows (and columns) that are close in the sociomatrix. If actors who "choose" each other occupy rows and columns that are close in the sociomatrix, then ties that are present will be concentrated on the main diagonal of the sociomatrix, and ties that are absent will be concentrated far from the main diagonal of the sociomatrix. For a dichotomous relation, 1's will be close to the main diagonal and 0's will be in the upper right and lower left of the sociomatrix. In analyzing a valued relation, ties with larger values will be concentrated along the main diagonal and ties with smaller values will be found in cells of the matrix that are off the main diagonal.

The goal is to permute the rows (and simultaneously the columns) of the sociomatrix to concentrate "choices" along the main diagonal (Katz 1947). Subgroups of actors who "choose" one another will then be close to each other in rows (columns) of the sociomatrix, and their choices will be close to the main diagonal of the sociomatrix.

Since the mid-1940's, numerous authors have proposed objective criteria for permuting rows and columns of a matrix to concentrate "choices" along the main diagonal of a matrix (Katz 1947; Beum and Brundage 1950; Coleman and MacRae 1960; Hubert 1985, 1987; Hubert and Arabie 1989; Hubert and Schultz 1976; Arabie, Hubert, and Schleutermann 1990). Some of these methods are applicable to matrices in general, and are thus not restricted to sociomatrices.

Figure 7.5 shows a small hypothetical sociomatrix, first in original order, and then with the rows and columns permuted so that actors who have ties to each other are close to one another in the sociomatrix.

Systematic procedures for permuting rows and columns of a sociomatrix seek to minimize a function that quantifies the extent to which ties with high values are far from the main diagonal (assuming that high values code strong ties). Recall that x_{ij} is the value of the tie from actor i to actor j. Furthermore, i and j index the rows/columns of the sociomatrix (for example, $i = 2$ refers to row 2 of the sociomatrix). Therefore, we would like to have large values of x_{ij} correspond to small differences between the indices i and j. Small differences between the indices can be

X

	n_1	n_2	n_3	n_4	n_5
n_1	-	0	1	0	1
n_2	0	-	0	1	0
n_3	1	0	-	0	1
n_4	0	1	0	-	0
n_5	1	0	1	0	-

X permuted

	n_5	n_1	n_3	n_2	n_4
n_5	-	1	1	0	0
n_1	1	-	1	0	0
n_3	1	1	-	0	0
n_2	0	0	0	-	1
n_4	0	0	0	1	-

Fig. 7.5. A hypothetical example showing a permuted sociomatrix

quantified either by small values of $|i - j|$ or by small values of $(i - j)^2$. The largest values of x_{ij} should occupy cells in which the indices i and j are close. The smallest values of x_{ij} should occupy cells in which the indices i and j are far apart.

For an entire matrix a summary measure of how close large values of x_{ij} are to the main diagonal is given by:

$$\sum_{i=1}^{g}\sum_{j=1}^{g} x_{ij}(i-j)^2 \text{ for } i \neq j. \tag{7.5}$$

The quantity in equation (7.5) is relatively small when large values of x_{ij} occupy cells of the sociomatrix with small differences between the indices i and j. This quantity is relatively large when large values of x_{ij} occupy cells of the sociomatrix with large differences between the indices i and j. If the value of equation (7.5) is small, then the ordering of rows and columns in the sociomatrix places actors among whom there are relatively strong ties close to each other, as is desired. On the other hand, if the value of this quantity is relatively large, then the ordering of rows and columns in the sociomatrix probably is not the best possible ordering for revealing cohesive subgroups of actors. Katz (1947) suggests permuting rows and simultaneously columns of the sociomatrix to minimize this quantity.

Beum and Brundage (1950), Coleman and MacRae (1960), and Arabie, Hubert, and Schleutermann (1990) suggest strategies for reordering rows and columns of the sociomatrix so that i and j corresponding to large values of x_{ij} are moved closer together. This problem of *sociomatrix permutation* to optimize a given quantity is an instance of the more general analysis problem of combinatorial optimization. Finding the single best ordering of rows and columns of a data array is computationally intensive, and, short of trying all possible permutations, there may be no guarantee that the optimum has been reached. Algorithms for permuting rows and columns to minimize a given objective function can be found in Arabie, Hubert, and Schleutermann (1990 and references therein) and a more general review of this data analytic approach can be found in Arabie and Hubert (1992).

The result of a matrix permutation analysis is a reordering of the rows and columns of the sociomatrix so that actors that are close in the sociomatrix tend to have relatively strong ties. However, a matrix permutation analysis does not indicate the boundaries between, or membership in, any subgroups that might exist in the network. Therefore, matrix permutation methods do not locate discrete subgroups. These methods do provide a preferred ordering in which to present a sociomatrix. Nevertheless, it can be quite informative to present the sociomatrix with rows and columns permuted to suggest the subgroup structure.

Other approaches to subgroup identification include methods for presenting the subgroup structure of a social network using standard data analytic methods to display proximities among actors. Approaches in this tradition use multidimensional scaling, hierarchical clustering, or factor analysis to represent the proximities among network actors. We will briefly describe multidimensional scaling and factor analysis for representing proximities among actors.

7.10.2 Multidimensional Scaling

Often the researcher is confronted with a set of network data and simply wishes to display the proximities among actors in the group. Such representations can be quite useful for understanding the internal structure of the group, for revealing which actors are "close" to each other, and for presenting possible cleavages between subgroups. Standard clustering and multidimensional scaling techniques can be used to represent proximities among actors when appropriate network measures are used as input.

Multidimensional scaling has been used by many network analysts to represent proximities among actors. Just a few of the many substantive examples include: studies of community elites (Laumann and Pappi 1973, 1976), naturally occurring communities (Freeman, Romney, and Freeman 1987; Freeman, Freeman, and Michaelson 1988; Arabie and Carroll 1989; Doreian and Albert 1989), organizational culture (Krackhardt and Kilduff n.d.), scientific communities (Arabie 1977), and state supreme court precedents (Caldeira 1988).

Multidimensional scaling is a very general data analysis technique, and there are numerous texts and articles describing multidimensional scaling (see for example Kruskal and Wish 1978; Schiffman, Reynolds, and Young 1981; and Coxon 1982). Multidimensional scaling seeks to represent proximities (similarities or dissimilarities) among a set of entities in low-dimensional space so that entities that are more proximate to each other in the input data are closer in the space, and entities that are less proximate to each other are farther apart in the space. The usual input to multidimensional scaling is a one-mode symmetric matrix consisting of measures of similarity, dissimilarity, or proximity between pairs of entities. To study cohesive subsets of actors in a network the input to multidimensional scaling should be some measure of pairwise proximity among actors, such as the geodesic distance between each pair of actors. The output of multidimensional scaling is a set of estimated distances among pairs of entities, which can be expressed as coordinates in one-, two-, or higher-dimensional space. Results are also displayed as a diagram in which the coordinates are used to locate the entities in the resulting one-, two-, or three-dimensional space. Using multidimensional scaling to study cohesive subgroups shows which subsets of actors are relatively close to each other in a graph theoretic sense.

An Example. To illustrate multidimensional scaling for studying cohesive subgroups we use the marriage relation for Padgett's Florentine families. Recall that this relation is dichotomous and nondirectional. Analyzing the sociomatrix directly using multidimensional scaling is unwise. Since there are only 0's and 1's in this matrix the multidimensional scaling solution would be very unstable. Instead, it is useful to compute a valued measure of proximity among pairs of actors. One such measure is the geodesic distance between pairs of actors. In our example we use the matrix of the geodesic distances among pairs of families as input to multidimensional scaling. We used *GRADAP* (Sprenger and Stokman 1989) to calculate the path distances, and *SYSTAT* (Wilkinson 1987) to do the multidimensional scaling. The Pucci family is an isolate on the marriage relation and thus was omitted from the multidimensional scaling. The final multidimensional scaling solution in two dimensions

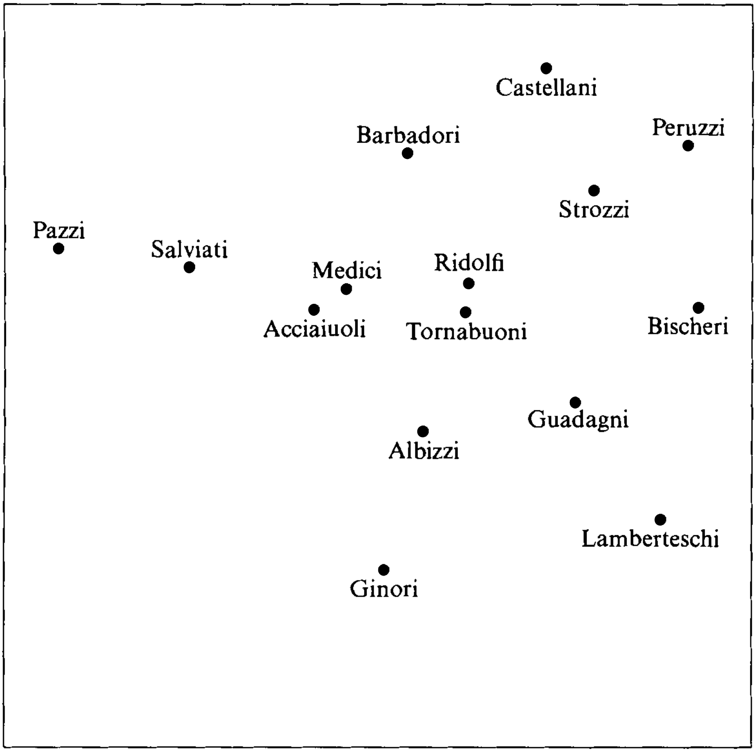


Fig. 7.6. Multidimensional scaling of path distances on the marriage relation for Padgett's Florentine families (Pucci family omitted)

has stress equal to 0.0198 (Kruskal, stress form 1). This result is presented in Figure 7.6.

Notice in Figure 7.6 that one of the most prominent families, Medici, is located in the center of the plot. It is also interesting to note that the six families (Bischeri, Castellani, Guadagni, Lamberteschi, Peruzzi, and Strozzi) identified by Kent (1978; also see Breiger and Pattison 1986) as being in the anti-Medici faction are, without exception, all on the right side of the plot.

7.10.3 ○Factor Analysis

Factor analysis of sociometric data was quite widespread and influential in the early history of network analysis (Bock and Husain 1952; MacRae 1960; Wright and Evitts 1961). Both direct factor analysis (in which a sociomatrix is input directly into a factor analytic program) and factor analysis of a correlation or covariance matrix derived from the rows (or columns) of a sociomatrix have been used to reveal aspects of network structure. In studying cohesive subgroups, Bonacich (1972b) shows that if a group contains non-overlapping subsets of actors in which actors within each subset are connected by either adjacency or paths, then a factor analysis of the sociomatrix will reveal this subgroup structure. However, one should be quite cautious about using factor analysis on dichotomous data, since results can be quite unstable.

Although factor analysis can be used to study cohesive subgroups in an exploratory way, the most influential and important cohesive subgroup ideas are those (such as cliques and related ideas) that express specific formal properties of cohesive subgroups and locate such subgroups that might exist within a network data set.

7.11 Summary

In this chapter we have presented methods for studying cohesive subgroups in social networks, for dichotomous nondirectional relations, directional relations, and valued relations. These methods are motivated by theoretically important properties of cohesive subgroups, and present alternative ways of quantifying the idea of social group using social networks. We also presented methods for assessing the cohesiveness of subgroups.