4 COHESIVE SUBGROUPS

Embedded within a network there are often groups of actors who interact with each other to such an extent that they could be considered to be a separate entity. In a friendship network this could be a group of close friends who all socialise together, alternatively in a work environment a collection of colleagues who all support the same football team, or in a network of interacting organizations a collection of organizations which behave as a single unit (so called virtual organizations). We call any such group a cohesive subgroup. In these examples we have identified the underlying cohesive theme which unites the group, this would not necessarily be apparent from the network under study. In examining network data we would first try and detect the cohesive subgroups and then, by looking at common attributes, see if there was some underlying principle that could explain why they identify with each other.

At first sight it may appear easy to identify cohesive subgroups in a network by simply looking at it. Unfortunately it is very easy to miss group members or even whole groups when trying to find cohesive subgroups by simply looking at a network. The position of actors on the page and the preponderance of edges make this task almost impossible to do by hand and we need to resort to algorithms and computers to perform the task for us. This is particularly true if the data was collected either electronically or by questionnaire, but even with observational data it is recommended that a proper and full analysis be undertaken. It should be remembered that some cohesive subgroups are open and want to be identified, but for others there is a strong dis-benefit in identification (For example a cartel, or a drugs ring). It is therefore necessary to have some formal definitions that capture exactly what a cohesive subgroup is. Within the social sciences the notion of a social group is often used casually. It is assumed that the reader has an intuitive grasp of the concept involved and that it is not necessary to present an exact definition. Clearly such an approach cannot be used to analyse real data and we are forced to define precisely what is meant by a cohesive subgroup. There are a large number of possible realisations of the social group concept but we shall only concern ourselves with the more commonly used practical techniques.

4.1 Intuitive Notions

We start by considering the most extreme case of a cohesive subgroup. In this case we expect members of the group to have strong connections with every other member of the group. If we have binary symmetric data, that is data that is represented by a graph, then this would imply that every actor in the group is connected to every other group member. In addition the group has no connections to individuals on the outside. In graph theoretic terms this group would consist of a component of the network which was a complete graph. Clearly such a notion has no place in any practical data analysis as such structures are so rare as to render the concept useless. However this strong intuitive notion allows us to construct models of cohesive subgroups based upon different relaxations of this ideal.

One of the first considerations is the type of data we are dealing with. Our ideal situation involved us looking at symmetric binary data, how do we deal with directed or valued data? Let us first consider non-symmetric data. If the data is binary, then since each pair of actors within the group have a strong connection between them we would expect all ties to be reciprocated. It is therefore not necessary to consider directed networks and so we restrict our attention to undirected relations. If the

original data is directed then as a pre-processing stage we should symmetrize it. Since we expect strength within the group we should symmetrize by constructing a network of reciprocated ties only. If there are very few or no reciprocated ties then we could use the underlying graph and simply ignore the directions of the edges. But this clearly is second best and care should be exercised in interpreting any subgroups found under these circumstances. These ideas are still applicable when we consider valued data and consequently we only consider symmetric valued data and apply the same principles in our symmetrization as in the binary case. For our valued data we expect strong ties within the group and weak ties outside.

4.2 Cliques

If we do have undirected binary data then we can relax the condition on our extreme cohesive subgraph by removing the constraint that there are no external links. If in addition we insist that all possible members of the group have been included then we call the resultant structure a clique. A clique is therefore a subset of actors in which every actor is adjacent to every other actor in the subset and it is impossible to add any more actors to the clique without violating this condition. A subgraph in which every actor is connected to every other is called a complete or dense subgraph. We can therefore define a clique as a maximal dense subgraph. Here the term maximal means that we cannot increase its size and still have a dense subgraph. In applications we usually insist that any clique has at least 3 actors.

W can illustrate the idea of a clique by examining the network in Figure 4.1. We see that nodes 1,2,3 and 4 are all connected to each other, in addition we cannot increase this group and still retain this property. Node 5 is connected to 3 and 4 but not to 1 and 2. It follows that {1,2,3,4} is a clique. Other cliques are {3,4,5}, {7,9,10}

and $\{7,8,10\}$. Note that $\{1,2,3\}$ is not a clique because it is not maximal (we can add

INSERT FIGURE 4.1 ABOUT HERE

4 to it). Clearly cliques can overlap so that individual actors can be in more than one clique. In our example we see that nodes 3,4,7 and 10 are all in two cliques. Finally there can be actors who are not in any cliques. Again returning to our example we can see that node 6 is not in any cliques.

<u>An example.</u> Roethlisberger and Dickson (1939) observed data on 14 Western Electric employees working in a bank wiring room. The employees worked in a single room and included two inspectors (I1and I3), three solderers (S1, S2 and S4) and nine wireman (W1 to W9). One of the observations was participation in horseplay and the adjacency matrix for this is

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
		I1	I3	Wl	₩2	W3	₩4	₩5	W6	W7	W8	W9	S1	S2	S4
1	I1	0	0	1	1	1	1	0	0	0	0	0	0	0	0
2	I3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Wl	1	0	0	1	1	1	1	0	0	0	0	1	0	0
4	W2	1	0	1	0	1	1	0	0	0	0	0	1	0	0
5	W3	1	0	1	1	0	1	1	0	0	0	0	1	0	0
6	W4	1	0	1	1	1	0	1	0	0	0	0	1	0	0
7	W5	0	0	1	0	1	1	0	0	1	0	0	1	0	0
8	W6	0	0	0	0	0	0	0	0	1	1	1	0	0	0
9	W7	0	0	0	0	0	0	1	1	0	1	1	0	0	1
10	W8	0	0	0	0	0	0	0	1	1	0	1	0	0	1
11	W9	0	0	0	0	0	0	0	1	1	1	0	0	0	1
12	S1	0	0	1	1	1	1	1	0	0	0	0	0	0	0
13	S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	S4	0	0	0	0	0	0	0	0	1	1	1	0	0	0

The program UCINET was used to produce the following 5 cliques.

- I1 W1 W2 W3 W4
- W1 W2 W3 W4 S1
- W1 W3 W4 W5 S1
- W6 W7 W8 W9
- W7 W8 W9 S4

We note that although these cliques overlap there are two distinct groups, namely {11,W1,W2,W3,W4,W5,S1} and {W6,W7,W8,W9,S4} together with two outsiders I3 and S2. These two groupings are in exact agreement with the findings of Rothlisberger and Dickson who identified the groups as those at the front of the room and those at the back. In this instance a simple clique analysis has been successful in identifying important structural properties of the network. Unfortunately most analyses are not as straight forward as this one. Two problems can occur. The first is that there is a large number of overlapping cliques and it is difficult to deduce anything from the data. The second problem is that too few cliques are found and so that no important subgroups are identified. We shall return to the first of these problems later in the chapter. The second problem of too few cliques can be addressed by having less stringent conditions for a subgroup to be a cohesive subset.

4.3 k-plexes

The condition for being a clique can be quite demanding. The fact that every member must be connected to every other member without exception is a strong requirement particularly if the group is large. It also means that the data has no room for error, a missing connection, for whatever reason, immediately stops a particular group from being identified as cohesive. The idea of a k-plex is to relax the clique concept and allow for each actor to be connected to all but k of the actors in the group. It follows that in a 1-plex every actor is connected to all other actors except for one, but since we normally do not have self loops (if we do have self loops for any cohesive subgroup method they are ignored) this implies that every actor is connected to every other actor and we have a clique. In a 2-plex every actor is connected to all but one of the other actors. We do again insist that the cohesive subgroup is maximal. We can see that in Figure 4.1 {7,8,9,10} is a 2-plex since 7 and 10 are connected to all the other nodes and 8 and 9 are connected to all but one of the other nodes. Another 2plex is the clique $\{1,2,3,4\}$ note we cannot include the actor 5 in this 2-plex since it does not have connections to 1 and 2 (note that $\{1,2,3,4,5\}$ does form a 3-plex). Consider the subset of actors $\{5,6,7\}$. Is this a 2-plex? The answer is yes since each actor is connected to all but one (that is just one) of the other actors. Clearly this group does not fit our intuitive notion of a cohesive subgroup. The problem is that the subgroup is too small to allow us to make the relaxation and still retain the characteristics of a cohesive group. We should therefore increase the minimum size of our 2-plex from 3 to 4. This problem can become worse as we increase the value of k. Again in Figure 4.1 the subset {3,4,5,7,8,10} is a 4-plex since every node is adjacent to all but 4 others (that is each node is adjacent to two others). It can be shown that if for a particular value of k the subset is bigger than 2k-2 then the distance between the nodes in the cohesive subgraph is always 1 or 2. This formula works well for k bigger than 4 but does not help us for k=2 or 3. The reason for this is that when the subgroups are small the distance between them must be small, in our $\{5,6,7\}$ 2-plex all the nodes are a distance of 1 or 2. We can combine both of these findings to produce a recommended minimum size for a value of k as given below.

k	Minimum size
2	4
3	5
4	7
k	2k-1

We can now define a k-plex as a maximal subgraph containing at least the number of actors given in the table above with the property that every actor is connected to all but k of the other actors.

Returning to Figure 4.1, the 2-plexes are {1,2,3,4},{1,3,4,5},{2,3,4,5} and {7,8,9,10} and the 3-plexes are {1,2,3,4,5}.

If we now examine the wiring room matrix we obtain the following 2-plexes

- I1 W1 W2 W3 W4 S1
- I1 W1 W3 W4 W5
- W1 W2 W3 W4 W5 S1
- W6 W7 W8 W9 S4

We can again clearly see the two groups and we have managed to reduce the number of cohesive subgroups as the k-plexes have allowed us to merge together two overlapping cliques. If we again increase k and look at the 3-plexes then we obtain

- I1 W1 W2 W3 W4 W5 S1
- W6 W7 W8 W9 S4

Which are precisely the groups identified by Roethlisberger and Dickson.