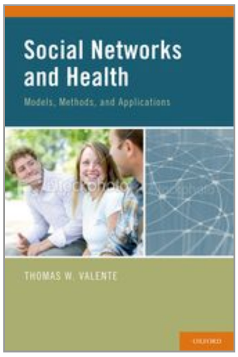

Social Networks and Health: Models, Methods, and Applications
Thomas W. Valente

Print publication date: 2010
Print ISBN-13: 9780195301014
Published to Oxford Scholarship Online: May 2010
DOI: 10.1093/acprof:oso/9780195301014.001.0001

Groups

Thomas W. Valente

DOI:10.1093/acprof:oso/9780195301014.003.0006

Abstract and Keywords

This chapter describes how network analysts define and measure groups. Components are the building blocks of group definitions and consist of all the nodes connected to each other through any number of steps in network. Nodes that cannot reach each other are in different components. More complex group definitions are then provided, namely, k-plexes and n-cliques which permit variation in the degree of connectivity among group members needed to be a group member. The Newman-Girvan (2004) algorithm is introduced which provides mutually exclusive groups and a measure of how well the group definitions characterize the data. The chapter closes with a discussion of how groups influence behavior.

Keywords: network groups, Newman-Girvan, cliques, norms

This chapter describes how network analysts define and measure groups. Components are the building blocks of group definitions and consist of all the nodes connected to each other through any number of steps in network. Nodes that cannot reach each other are in different components. More complex group definitions are then provided, namely, *k*-plexes and *n*-cliques, which permit variation in the degree of connectivity among group members needed to be a group member. The trouble with many group definitions, however, is that individuals can be members of many groups. While this reflects empirical reality, it can complicate statistical analysis. The Girvan-Newman (2002) algorithm is introduced, which provides mutually exclusive groups and a measure of how well the group definitions characterize the data. The chapter closes with a discussion of how groups influence behavior.

Most people enjoy being in groups. Groups help people define their identity and provide a sense of belonging. Belonging to a group signals to others our identity and how we relate to the rest of the world. Groups also provide protection. By belonging to a group a person does not have to defend his or her ideas, but rather can adopt those of the group and feel secure that others feel the same way he or she does.

Groups also provide the context for socializing, talking, and being with others. Humans are social animals; people like to talk and spend time with **(p.101)** others and groups can provide an organized, and sometimes not so organized, way to be with other people. But not just any other people: rather, people who have similar ideas, attitudes, opinions, and behaviors. In sum, people often join groups to be with others who are a lot like themselves or at least share some particular trait.

Scholars have conducted a lot of research on groups, how they form, the social pressure they induce, and the tendency for groups to reinforce existing beliefs (Moscovici, 1976). Network analysts are concerned primarily with providing network definitions for what it means to be in a group and how to define and analyze groups from a network perspective. The starting point is to take data collected on who is connected to whom, then define what it means to be in a group in the network, and then create these network groups. Whether the network groups match with other definitions of groups remains a somewhat open research question. For example, to what extent do network-defined groups among high schools students correspond to commonly labeled groups such as “jocks,” “nerds,” “geeks,” “skaters,” etc. (these are often referred to as identities)?

Network analysts define groups as any subset of a network. For example, the boys in a network of high school students (assuming it is coeducational) can be defined as a group within the network. Generally, though, a subset of a network is a group if it meets some network definition of being a group. The simplest such definition is a component.

Components and *K*-Cores

A component consists of the connected nodes in the network, all nonisolates. Everyone who can reach everyone else and be reached by everyone else is in the same component. One can also define weak and strong components. A weak component ignores the direction of the tie, while a strong component does not (Scott, 2000). In other words, strong components are nodes connected to one another in both directions along every step in the path connecting them. A component is a simple concept but can be useful if one has a large network that has several separate components. The components can be extracted and analyzed as separate networks.

It is also possible to reduce a network by reducing it to only reciprocated ties, and this reduction may create separate components. It is also useful to describe a network by the number of components. A network with one component is quite different than a network with many. Components then simply determine whether nodes are connected and define groups based whether the nodes can reach one another through any path of connections in the network. Once a component analysis is completed, the researcher can **(p.102)** progress to determining the groups, clusters, and cliques within the overall network.

There are several different definitions of a group in a network, one of which creates group definitions based on the number of connections the nodes have in the network. A component consists of all nodes that have at least one connection. This concept of the number of links defining a group can be generalized by creating *K*-cores. A *K*-core is a subset of the network in which each node within the *K*-core is connected to at least *K* other people. Thus, a 2*K*-core is the set of people connected to at least two other nodes. All the nodes with zero or one link are dropped from the network. Similarly, we can define a 3*K*-core as all the nodes with three or more links. As *K* is increased, successive pictures of who is left in the network will look increasingly dense. Once nodes are dropped from the network, the links from and to it are also dropped so the *K* criterion is calculated on the remaining nodes and links.

The pattern of node removal as *K* is increased can be used to describe the network structure. For example, some networks have a core-periphery structure, which is a network with a set of dense connections among a subset of nodes and another set residing on the periphery with fewer connections (see Chapter 8). Most of the connections in the periphery are to the core and not to each other. Figure 6–1 shows a network of friendship links

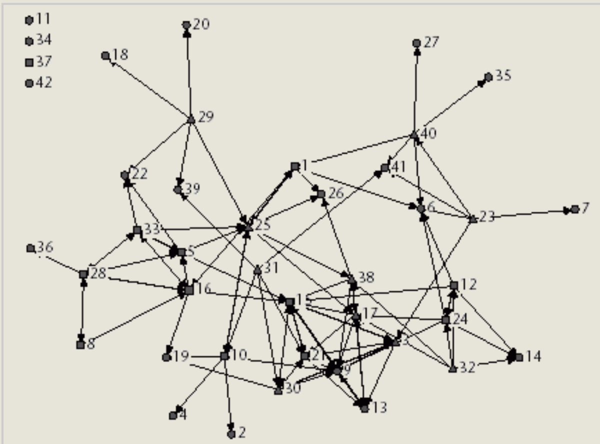

Figure 6–1. A network of friendship links between students in an alternate high school in southern California.

(p.103) between students in an alternate high school in southern California. We ran the *K*-core routine in UCINET on this friendship network and it indicated that the cluster of friends in the lower middle of the graph constituted the *K*-core. Seven students, numbers 38, 15, 17, 21, 9, 3, 13, and 30, all remain connected to each other as *K* is increased from one to four. The number of nodes dropped at each unit increase in *K* can be graphed as a histogram or bar graph at each increment of *K*. If the bars are all the same height, it indicates the same number of nodes were removed at each increase in *K*. If the bar graph is uneven or steeply increasing or decreasing, it indicates structural variations in which the number of nodes removed from the core changes abruptly. The pattern of this bar graph can provide an indication of whether the network has a core-periphery structure. If the bar graph is uniform across all bars, this indicates no core-periphery and little structure in the network. Conversely, if there is a sharp decrease in the bars, this indicates a core-periphery structure or at least considerable structure in the network. If there are a few nodes remaining after *K* is increased to a high number, then this indicates a *K*-core. If *K* is increased to seven and four nodes are left, these four nodes would be a 7*K*-core, indicating these four nodes have at least seven connections in the network. If the number of nodes removed is dramatic, such that when *K* equals three all the nodes are dropped, then the network does not have a strong core-periphery structure. In general, the greater the percent decrease in nodes left in the core, the less a core-periphery structure exists. This is not a strict rule, though, because it may be that many nodes drop early but the core retains to high levels of *K*. Suffice it to say, it is instructive to look at the *K*-core collapse, the pattern of dropped nodes as *k* increases.

Although finding the core group is important, it is also important to understand how groups are distributed in the network and who belongs to which group. For this purpose, network analysis uses the concept of a clique. Strictly speaking, a *clique* is the set of points all directly connected to one another. A clique is a dense pocket of interconnectivity with every person directly connected to everyone else. Yet this definition of clique can be somewhat restrictive as it is common to have a group of people one might think of as a clique, yet every member is not directly connected to every other member.

The strict clique definition can be relaxed to define an *n*-clique, which is the path length at which members of the clique are connected. For example, a 2*n*-clique is the set of people connected to each other within two steps. Two people are in the same clique if they are friends of the same friend. The *n*-clique definition allows people to be in the same group even if they are not directly connected to one another. One can also increase “*n*” to higher numbers such as 3*n*-cliques and 4*n*-cliques. Typically, though, values greater than two are rarely used because it seems counterintuitive for people to be in **(p. 104)**

Table 6–1. List of *N*-Cliques for the Network Given in Figure 6–1

Clique Members
1: 13 15 17 21 9
2: 13 15 17 38 9
3: 13 15 21 3 9
4: 15 21 3 30 9
5: 15 16 5
6: 12 14 24 32
7: 1 25 26
8: 1 40 6
9: 22 33 5
10: 23 40 41
11: 23 40 6
12: 17 24 32
13: 17 24 9
14: 12 24 6
15: 25 26 38
16: 17 25 38
17: 25 33 5
18: 16 28 33 5
19: 16 28 8
20: 21 30 31
21: 17 32 38

the same clique if they are three steps from one another. Table 6–1 reports the 2*n*-cliques for the network in Figure 6–1.

A second clique definition is a *k*-plex, which is defined as the set of points connected to all but *k* other nodes in the group. To find *k*-plexes, the researcher sets both “*k*” and “*n*” to the size of the groups. The minimum size for *n*, the size of the groups, is set to $k - 2$ (because values of *n* close to *k* return trivial groups) For example, 2*k*-plexes with $n = 7$ will find all groups of size seven in which each person is connected to at least five others in the group. If *k* is increased to three, 3*k*-

plexes, all groups of size seven in which each person is connected to at least four other members would be reported. So as *k* increases, the number of groups identified in the network increases. In practice, one sets *k* and finds all of the groups as *n* increases from *k* + 2 to *n* – 1. For example, with *k* = 2 and *n* = 4, the 2*k*-plexes are all of the groups size four and larger in which the members are connected to at least two others. Table 6–2 reports the *k*-plexes for our sample network.

The *K*-cores, *n*-cliques, and *k*-plexes provide good measures of network structure, and as the researcher studies a network, these group identification methods provide insight into the pattern of affiliations such that one can characterize the network according to who is in which groups with whom. UCINET returns a group co-membership matrix in which the numbers in the matrix indicate how many groups each person shares with everyone else. **(p.105)**

Table 6–2. List of *k*-Plexes for the Network Given in Figure 6–1

Group Members
k-plex
Value of K: 2 (each member of a *k*-plex of size *N* has *N*-*k* ties to other members)
Minimum Set Size = 5
Input dataset: C:\MISC\DIFFNET\snbh\tprc_4
WARNING: Directed graph. Direction of arcs ignored.
Six *k*-plexes found.
1: 13 15 17 21 3 9
2: 13 15 17 21 38 9
3: 13 15 21 3 30 9
4: 13 15 3 38 9
5: 15 17 21 30 9
6: 17 24 32 38 9

The diagonal of this matrix indicates how many groups (cliques or *k*-plexes) each person belongs to. The group co-membership matrix can be used in subsequent analysis, for example, to investigate whether people who share many groups have similar attitudes and behaviors. (This would be done, for example, by correlating the matrix of shared groups with the matrix of attitude similarity; see Chapter 8.)

One problem with these group identification methods, *n*-cliques and *k*-plexes, is that they return lists of groups in which many people are members of multiple groups. This may reflect reality and be advantageous in some research analysis. On the other hand, for researchers studying behavior it can be difficult to conduct analysis comparing group membership to behavior because people are not classified in mutually exclusive groups. One solution is the hierarchical clustering analysis provided as part of the group analysis in UCINET. The hierarchical clustering output indicates group assignments at various cutoff levels. There is no rule to indicate where that level should be set. Further, when studying multiple networks in the same study, one might use different thresholds for different networks, thus creating potentially biased analysis. Fortunately, researchers have developed a grouping technique that partitions a network into mutually exclusive groups. This way each person belongs to one and only one group. The method also provides a measure of how well the data partition into these groups.

Girvan-Newman Technique

Girvan and Newman (2002; Newman and Girvan, 2004) used a technique of deleting selected links from a network to identify components. If one link connects two components in a network, then deleting that link would **(p.106)** yield the two components, and the group structure of the network is nicely described by these two components. In this hypothetical example, the network structure is clear—one deleted link created two groups in which all the ties are contained within each group and there are no links between groups. Measuring how many links are within the groups and how many between the groups provides an indication of how well the group definitions characterize the network.

The problem now becomes which links to remove that are most likely to return separate components. Girvan and Newman (2002) suggested deleting those links that are most central in the network, the links with the highest centrality.

Recall in Chapter 5 that it was mentioned briefly that centrality can be calculated on the network of links. Girvan and Newman suggested calculating centrality on the network of links (not the nodes). Every link has two nodes, but the location of those nodes affects the centrality calculations of the links. By calculating centrality on the links, researchers can identify those links that are the best candidates to remove to partition the network into mutually exclusive groups.

Grivan-Newman (GN) subgroups are derived by calculating betweenness centrality on the links, deleting the link with the greatest betweenness centrality, and then determining if there are any components. Each remaining component constitutes a group. After the link with the greatest betweenness centrality is located and deleted and any groups are detected, the process is repeated until the desired number of groups is reached. One can predetermine the number of groups desired, or the analyses can run until no groups greater than a specified size are detected. For example, if a researcher wanted to partition a large number of networks into six groups, then the number of groups can be fixed to six. On the other hand, the researcher can set the minimum group size, say three, and allow the algorithm to find multiple group structures and various group sizes.

The GN technique provides a partition of the network into mutually exclusive groups and it also measures how well the group partition characterizes the network. The measure for how well the groups characterize the network is called modularity (Newman and Girvan, 2004). Modularity is calculated by creating a matrix in which each row and column is one of the newly created groups. Each element in this new reduced matrix is the number of links from the original network between these created groups. The diagonal elements of the network are the number of links within each group. To calculate how well the group partition fits the data, the percentage of within group ties to between group ties is calculated. Newman and Girvan (2004) provides the equation:

6–1

$$Q = \sum_i \left(e_{ii} - a^2\right) = Tre - \| e^2 \|$$

(p.107) where *Tre* is the trace of the matrix (the sum of the diagonal elements) and $||e^2||$ is the sum of all the elements in the matrix. Thus, this quantity, *Q,* indicates the percentage of ties in the network that occur within the groupings found by the algorithm. A *Q* of, say, 90% indicates almost all links are to members within the groups identified by NG, whereas one of 10% indicates that few are. The NG algorithm provides a way to partition a network into mutually exclusive groups and an index of how well that partition reflects the overall pattern of ties in the network. Thus, the researcher can choose among many competing partitions, selecting the one that best fits the data or selecting one with a less good fit but that is preferable for substantive reasons. A similar index, E-I, was suggested by Krackhardt and Stern, which takes the difference between the within-group and between-group links (Krackhardt & Stern, 1988). Figure 6–2 redraws the network in Figure 6–1 circling the groups identified by the GN algorithm in Netdraw. For example, 18, 20, 29 and 39 are in a group. From the picture the groups seem to make sense. The calculation of *Q* was 48.4%, indicating the 48.4% of the links are contained with the defined groups.

In some applications, the GN approach has clear advantages over *n*-clique and *k*-plex analysis because it provides a unique partition to the network and a measure, *Q,* that indicates how well the partition describes the network. One difficulty, of course is that networks with a high *Q* score will also

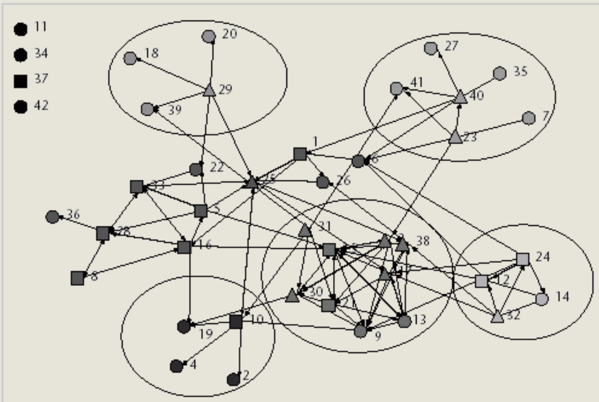

Figure 6–2. The network in Figure 6–1 with groups identified by the Newman-Girvan algorithm circled. Rates of smoking between groups varied, suggesting that group membership may affect individual smoking status.

(p.108) have very distinct clustering results and networks with a low *Q* score are not modular and so are not well described by network partitions or clique/cluster analysis. The network in Figure 6–2 has a *Q* of 48%, which is pretty high, and so clique/cluster analysis and NG are somewhat descriptive of this network. In this example, there is considerable modularity in the network but there is no rule currently for what constitutes an acceptable Q value and this may vary by applications and research questions.

Groups and Behavior

Logic and empirical evidence indicate that people who belong to the same group will engage in the same behaviors. Groups have norms and members of the group are expected to uphold those norms. For example, if a person joins a chess club it is expected that he or she will play chess, enjoy chess, and value it as a game. Chess club members also probably value other games of mental acuity such as “Go” or bridge. Conversely, joining a sports team, such as football, signals athleticism and competition, which values physical strength.

In addition to these norms, though, groups also exert social pressure, sometimes subtly and sometimes overtly. To be a member of a group could mean that some members will have to display commitment to the group's values and group members may sometimes communicate to other members that they expect everyone to uphold these values. There is pressure to conform to group norms for behavior, sometimes overtly. The influence of group membership occurs both through selection and influence. A person selects a group to belong to because he or she is interested in the things that the group stands for. Once becoming a member, group members may pressure or persuade him or her to adopt the groups’ norms.

Disentangling which comes first, selection or influence, is not always easy. Chance events may incline a person to join a particular group and then over time a person may feel pressured or compelled to adopt the group's norms. Moreover, once belonging to a group, the person may not want to risk the loss of camaraderie, companionship, solidarity, or belonging that would come from nonadoption of group behaviors. Both selection into a group and the influence of that group on individual attitudes and behaviors are processes that result in group members having similar attitudes and behaviors.

Groups also provide opportunities for information, resources, and support. Belonging to a group means that information available to the group will also be available to the individual and generally people will communicate more with group members than those outside the group. This increased communication means that members will be aware of and have access to whatever communication occurs within the group. It also means members may have access **(p.109)** to resources that nongroup members do not have access to. For example, many people join a country club so they can network with other members.

In this sense, group membership provides one form of social capital. People usually trust others who are members of their group. *Social capital* is the resources available from one's social networks. Members of the same group tend to trust one another more than nonmembers, and because they are connected, there is more opportunity for a person to access the resources from others within the group. Thus, groups can increase connectivity among members and increase social capital.

Groups provide pressure for behavior change in another way. Interdependent innovations such as fax machines, email, text messaging, Facebook, and so on are innovations in which both parties have to adopt them for them to be used. Most (perhaps all) communication technologies are interdependent since their value is in the ability to provide communication and connection to others. If most members of a group adopt an interdependent behavior, then the nonadopters feel pressure to adopt. For example, if a person's friends begin to text message each other, he or she may feel pressure to begin text messaging to stay connected to the group.

The graph in Figure 6–2 also indicates whether each student reported being a current smoker. The symbols for the nodes correspond to the following: circle, no data; square, nonsmoker; triangle, smoker. Inspection of Figure 6–2 shows some clustering. Specifically, in the largest group, only two of the seven students for whom we have data reported being a smoker, whereas five of the seven students in the second largest group reported being a smoker. The other four groups report rates in between. We collapsed the grouping into three mutually exclusive groups, retaining the largest two and combining the other four, and compared the mean rates of smoking statistically. The smoking rates differed statistically significantly between groups (mean = 14.3% [SD = 37.8%], 71.1% [SD = 48.8%], and 57.1% [SD = 53.4%], respectively; $F = 2.79$; $p = .09$).

There is some evidence that belonging to a group is associated with behavior. At the most basic level, isolates have often been the last to adopt new ideas and practices across numerous behaviors (Valente, 1995). Isolates have few or no interpersonal sources of information or influence. Isolates often lack emotional and social supports necessary to adapt to changing circumstances. Joining groups provides a means to avoid being isolated.

Group Membership and Disease

Although belonging to a group can be advantageous in terms of access to information and resources, for disease spread, belonging to a group can **(p.110)** increase risk. For example, belonging to a group of people who engage in risky sexual behavior may put one at increased risk of contracting a sexually transmitted disease. In general, group membership can protect one from disease as long as the incidence and prevalence of that disease in the group are low. Once prevalence reaches a critical threshold, however, being in the group puts one at increased risk since the disease will circulate rapidly in the closed group.

As long as the group is completely self-contained and has no outside contact, the group will remain immune and risk free. For example, communities that live in remote regions or on islands that are infrequently visited report being cold and flu free for some time. Eventually, however, a ship or airplane arrives bringing supplies, friends, families, and, of course, colds, the flu, and other illnesses (Gilmore, 1998). Group membership, therefore, can confer benefits, access to information, access to resources, and protection from disease if within-group prevalence is low. Conversely, group membership can be a liability if that group lacks resources or if disease prevalence is high.

Groups, Density, and Bridges

There is a tradeoff between the desire to form groups that are exclusive and consist of dense communications and commitments among its members versus having a group that maintains substantial ties outside the community. Earlier research has shown that community density was associated with more rapid diffusion of innovations (Valente, 1995). Dense networks provide more pathways than sparse ones, along which communication about new ideas and behaviors can flow. Sparse networks may not provide sufficient pathways for information to be circulated or for resources to be accessed, thus not conferring the benefits of the group. Density may also facilitate diffusion because dense networks may reflect a cohesive normative environment. A network with many links is more likely to have members who share common values or beliefs. Thus, a dense network may reflect a homogeneous community, and this homogeneity will facilitate information exchange and decision making.

Conversely, dense networks may not be efficient for several reasons. First, higher density may reflect more formal associations and these formal ties may not be as persuasive or trust-enhancing as informal ties (Krackhardt, 1992). Second, although there is a minimum density level needed for an organization or community to adopt innovations, once this level is reached, too much density may be a liability because it can limit connections to external information and resources. The lack of external ties then becomes a liability. **(p.111)** Organizational studies have shown that too much density can hurt performance (Oh et al., 2004; Uzzi, 1998). Finally, researchers have documented the tendency for groups to reinforce the opinions of members so that people fail to consider the perspectives of nonmembers. The relationship between density and performance, adoption behavior, or other outcomes is likely to be curvilinear as in Figure 6–3.

Figure 6–3 proposes that some basic level of density is necessary for organization, coalitions, and groups to function. The relationship between density and performance is likely to be nonlinear, the greater the density, the better the performance, or the more rapid the adoption of innovations. At some critical point, however, this relationship levels off and too much density begins to detract from performance. Exactly at which level density transitions from being an asset to a liability depends on the kind of performance or diffusion being studied and may depend on characteristics of the group members.

Support for this hypothesis was provided in a study of the effectiveness of satellite TV training of community coalitions to prevent substance use. Twenty-four community coalitions were randomly assigned to three conditions: control, satellite TV training, and satellite TV plus technical assistance. The coalitions were expected to adopt evidence-based programs for the prevention of substance use in their communities. Adoption was expected to be greater in dense coalitions and lesser in sparse ones. Surprisingly, however, coalitions that increased their density reported lower levels of program adoption (Valente et al., 2007). Figure 6–4 illustrates the

(p.112) main finding showing that an increase in density was associated with lower uptake of practices between year 1 and year 2. The critical observation here is that too much density may restrict the

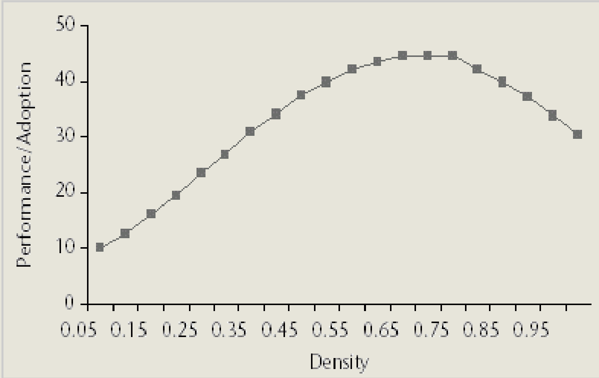

Figure 6–3. Proposed curvilinear association between performance or adoption and network density.

formation of bridges between groups or from a group to outside information and resources. Most groups (organizations, c the desire to create cohesion against the need to access new information and resources. Simply increasing the size of a coalition or hosting meetings designed to get everyone on the same page may not be the most effective means to

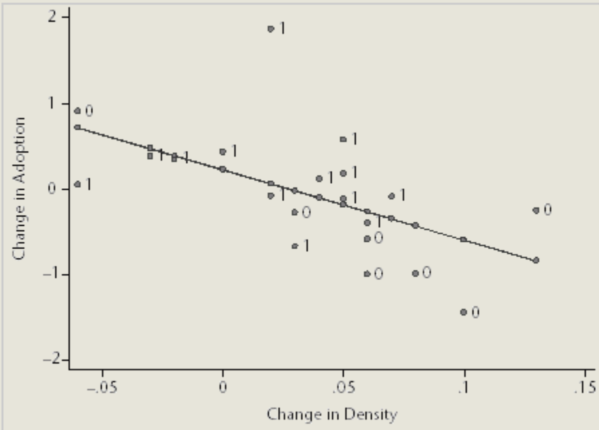

Figure 6–4. Decreased density was associated with increased adoption of evidenced based programs (or increased density was associated with less adoption). It is noteworthy that the communities labeled with “0” are the control ones and these control communities seem to pull down the association between density and adoption, indicating that they are responsible for this negative relationship.

create more effective organizations.

In follow-up analysis, we also showed that adoption of effective programs was also influenced by the centralization of the advice networks, after controlling for density (Fujimoto et al., in press). Decrease in centralization for the advice network and increased centralization in the discussion networks were associated with adoption. Networks based on friendships showed no consistent relationships to adoption. These analysis provide insight into the potential for optimizing networks for organizational and system performance. We return to this issue in Chapters 8 and 11 when discussing network-level indicators (Chapter 8) and network interventions (Chapter 11).

(p.113) Summary

This chapter reviewed procedures and approaches to understanding how to define a group in network analysis. There are many different definitions of a group, but most agree that a group is a set of at least three people who are more closely connected to each other than to other people in the larger network. The definition of a clique was defined as a group in which each member is no more than *n* steps from every other member, a so-called *n*-clique. *N* can be varied to calculate different types of groups. *k*-plexes were defined as groups based on the number of other people in the group each person is connected to.

One challenge noted in this chapter is that many group definitions do not necessarily provide a mutually exclusive partition of the network into separate groups. That is, *n*-clique and *k*-plex analysis return a listing of many overlapping groups. This group analysis reflects the data and real world experience that many groups are not mutually exclusive, yet it can hamper statistical analysis of the relationship between group membership and behavior. The GN algorithm provides a means to identify groups in a network that are mutually exclusive and simultaneously provides an index of how well the network conforms to a mutually exclusive grouping pattern. We closed the chapter with a discussion of how groups may affect the diffusion of behaviors.