Part III - Brokerage

In quite a few theories, social relations are considered as channels which transport information, services, or goods between people or organizations. In this perspective, social structure helps to explain how information, goods, or even attitudes and behavior diffuses within a social system. Network analysis reveals social structure and helps to trace the routes which goods and information may follow. Some social structures permit rapid diffusion of information whereas others contain sections which are difficult to reach.

This is a bird's-eye view of an entire social network. However, we can also focus on the position of specific people or organizations within the network. In general, being well-connected is advantageous. Contacts are necessary to have access to information and help. The number and intensity of a person's social relations is called his or her 'sociability' or 'social capital', which is known to correlate positively to age and education in Western societies. Some people occupy central or strategic positions within the system of channels and are crucial for the transmission process. Such positions may put pressure on their occupants, but they may also yield power and profit.

In this part of the book, we focus on social networks as structures which allow for the exchange of information. In this approach, the direction of relations is not very important, so we will only discuss undirected networks (with one exception). In Chapter 6, we present the concepts of centrality and centralization. In Chapter 7, we discuss the structure of the immediate network of actors, especially the pressure or power which is connected to particular structures of this ego-network. In Chapter 8, we take time into account as we study the role of network structure in the diffusion of innovations and diseases.

6 Center and periphery

6.1 Introduction

In this chapter, we present the concepts of centrality and centralization, which are two of the oldest concepts in network analysis. Most social networks contain people or organizations which are central. Because of their position, they have better access to information and better opportunities to spread information. This is known as the **ego-centered approach** to centrality. Viewed from a **socio-centered** perspective, the network as a whole is more or less centralized. Note that we use **centrality** to refer to positions of individual vertices within the network, whereas we use **centralization** to characterize an entire network. A network is highly centralized if there is a clear boundary between the center and the periphery. In a highly centralized network, information spreads easily but the center is indispensable for the transmission of information.

In this chapter, we will discuss several ways of measuring the centrality of vertices and the centralization of networks. We will confine our discussion of centrality to undirected networks because we assume that information may be exchanged in both ways between people or organizations which are linked by a tie. Concepts related to importance in directed networks, notably prestige, are discussed in Part IV of this book.

6.2 Example

Studies of organizations often focus on informal communication: who discusses work matters with whom, whom do people turn to for advice? Informal communication is important to the operation of the organization and it does not always coincide with the formal structure of the organization. For the diffusion and retrieval of information, it is crucial to know the people who occupy central positions in the communication network.

Our example is a communication network within a small enterprise: a sawmill. All employees were asked to indicate the frequency with which they discussed work matters with each of their colleagues on a five-point scale ranging from less than once a week to several times a day. Two employees were linked in the communication network if they rated their contact as three or more. It is not known whether both employees had to rate their relation in this way or that at least one employee had to indicate a strength of three or more. The network is stored in the file `Sawmill.net`.

In the sawmill, the employees are Spanish-speaking (H) or English-speaking (E), which, of course, is relevant to their communication. The sawmill contains two main sections: the mill (M), where tree trunks are sawn into logs, and the planer section (P), where logs are planed. Then there is a yard (Y) where two employees are working, and some managers and additional officials.

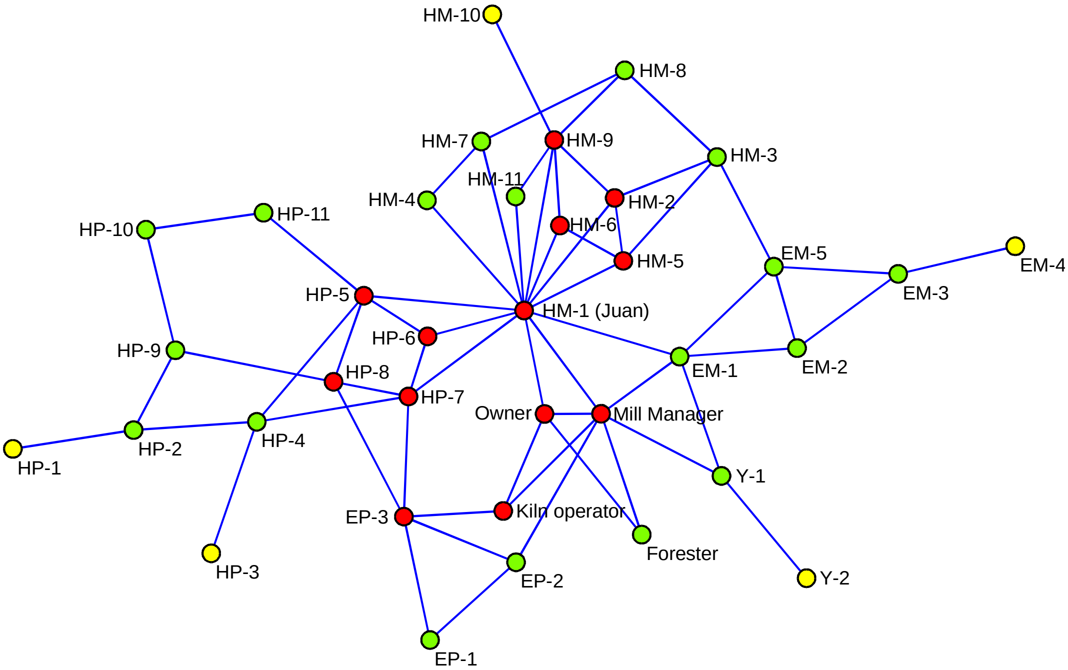


Figure 1 - Communication relations within a sawmill (colors indicate *k*-cores).

Figure 1 shows the communication network in the sawmill. Note that vertex labels indicate the ethnicity and the type of work of each employee, e.g. HP-10 is an Hispanic (H) working in the planer section (P). In this figure, vertex labels instead of and vertex colors identify the attributes of employees. It is quite easy to see that work-related communication is structured along work section (planers at the left, sawyers at the right) and ethnicity: Hispanics at the top and English-speakers at the bottom - assuming that management, forester, kiln operator, and employees in the yard are English-speakers.

Intuitively, centralization seems to be a special case of cohesion, which was presented in Part II of this book. If we compute the *k*-cores of the communication network, we find a tightly knit 3-core (red) in the center, surrounded by a 2-core (green), and some vertices which belong to a 1-core in the periphery. The 3-core seems to be the center of the network. All red vertices communicate with at least three others within the core, so they receive information from several sources and information will reach them easily. In this example, they are the ones that can pass information on from one part of the periphery to the other, e.g., information from planer HP-1 can only reach a sawyer or a person working in the yard via employees in the 3-core.

Clearly, HM-1 (‘Juan’) is very important to the flow of information because he connects Hispanic sawyers to English-speaking sawyers, to management, and to the Hispanic planers. Juan seems to be the center of the 3-core, which implies that a single person (vertex) can be the center of a network instead of a cohesive group. Centrality, therefore, is not a special kind of cohesion. It belongs to another family of structural concepts, which we label ‘brokerage’ in this book.

6.3 Distance

One approach to centrality and centralization is based on the simple idea that information may easily reach people who are central in a communication network. Or, to reverse the argument, people are central if information may easily reach them.

The larger the number of sources accessible to a person, the easier it is to obtain information, for instance, an elderly person will acquire information about where to look for help more easily if his or her social support network is larger. In this sense, social relations constitute a 'social capital' which may be used to mobilize social resources. Hence, the simplest indicator of centrality is the size of a person's network, which is his or her degree in a simple undirected network (see Chapter 3). The higher the degree of a vertex, the more sources of information it has at its disposal, the quicker information will reach the vertex, so the more central it is. In the sawmill network, Juan communicates with no less than 13 colleagues, whereas the manager of the mill has only seven communication relations (Figure 1). Juan is more central than the manager and information from the shop floor will reach him more easily than the manager.

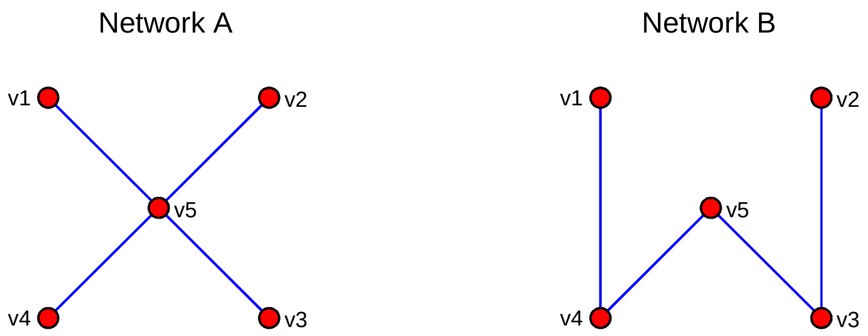


Figure 2 - Star and line-network.

If degree is the simplest measure of the centrality of a vertex, what is the associated measure of centralization for the entire network, which expresses the extent to which a network has a center? Let us first answer another, related question: Given a fixed number of lines, what is the most efficient structure to exchange information? First of all, we should note that this network must be connected, otherwise information can not reach all vertices. Then, the star-network is known to be the most efficient structure given a fixed number of lines. A star is a network in which one vertex is connected to all other vertices but these vertices are not connected among themselves, e.g., network A in Figure 2.

Compare the star-network in Figure 2 to the line-network (network B). It is much easier to identify the central vertex in the star-network than in the line-network because the difference between the central vertex (v5) and the peripheral vertices (v1, v2, v3, and v4) is much more apparent than in the line-network. This leads to an idea which may be counterintuitive, namely that a network is more centralized if the vertices vary more with respect to their centrality. The more variation in the centrality scores of vertices, the more centralized the network.

Degree centralization is the variation in the degrees of vertices divided by the maximum degree variation which is possible in a network of the same size.

Now we can define degree centralization as the variation in the degree of vertices divided by the maximum variation in degree which is possible given the number of vertices in the network. The division by maximum degree variation ensures that degree centralization ranges from zero (no variation) to one (maximum variation) in the case of a star network.

In network A (Figure 2), for instance, four vertices have minimum degree, which is one in a connected network. One vertex has maximum degree (four) because it is connected to all other vertices. The variance of the degrees, which is a standard statistical measure of variation, is 1.44 (consult your statistics text book if you want to check this number). In a connected network, the degree of vertices can not vary more than this, so 1.44 is the maximum variation and dividing 1.44 by itself, of course, yields a degree centralization of 1.00. In network B, two vertices have degree one (v1 and v2) and the other vertices have a degree of two; the variance of the degrees is approximately 0.24. To obtain the degree centralization of network B, we divide 0.24 by 1.44, which is the maximum variance, and we obtain 0.17. If we add a line between v1 and v2 , it becomes minimal (0.00) because all vertices have equal degree, so variation in degree is zero and degree-centralization is zero.

Degree-centrality is just the number of neighbors of a vertex. In some cases, this is all we know about the network position of people, for instance, when data are collected by means of a survey in which people are asked to indicate the size of their personal network. If we want to analyze the communication structure of the network, however, we need to know who is connected to whom in the entire network and we must pay attention to indirect relations because information can flow from one person to the next, and on to other people. In a communication network, information will reach a person more easily if it does not have to 'travel a long way'. This brings us to the concept of **distance** in networks, namely the number of steps or intermediaries needed to travel from one person to another. The shorter the distance between vertices, the easier it is to exchange information.

In Chapter 3, we defined paths as a sequence of lines in which no vertex in between the first and last vertices occurs more than once. Via a path, we can reach another person in the network: we can inform our neighbor, who passes the information on to his neighbor, who passes it on, until the information finally reaches its destination. We say that a person is **reachable** from another person if there is a path from the latter to the former. Note that two persons are mutually reachable if they are connected by a path in an undirected network, but that two paths (one in each direction) are needed in a directed network.

A **geodesic** is the shortest path between two vertices.
The **distance** from vertex u to vertex v is the length of the geodesic from u to v.

In an undirected network, the distance between two vertices is simply the number of lines or steps in the shortest path which connects the vertices. A shortest path is also called a **geodesic**. In a directed network, the geodesic from one person to another is different from the geodesic in the reverse direction, so the distances may be different. This sounds strange if you are used to geographic distances but think of a directed network as a system of one-way streets: it is easy to imagine that the route from A to B is different from the journey back. In this chapter, however, we will only use undirected networks, so you do not have to worry about this now.

With the concept of distance, we can define another index of centrality which is called closeness-centrality. The closeness-centrality of a vertex is based on the total distance between one vertex and all other vertices, where larger distances yield lower closeness-centrality scores. The closer a vertex is to all other vertices, the easier information may reach it, the higher its centrality. Note that the network must be (strongly) connected because it is impossible to compute the distance between vertices which are not reachable due to the simple fact that there is no path between them, hence no geodesic with a countable number of lines. If the network is not (strongly) connected, closeness-centrality can not be computed.

The **closeness-centrality** of a vertex is the number of other vertices divided by the sum of all distances between the vertex and all others.

In network A (Figure 2), vertex v5 has maximum closeness-centrality (1.00) because it is directly linked to all other vertices. The sum of distances to the other vertices is minimal, so closeness-centrality is maximal (1.00). The other vertices of network A have a closeness-centrality score which is considerably lower (0.57) because three vertices are two steps away from them.

In network B, v5 also has the highest closeness-centrality because it is in the middle, but now its closeness-centrality is not maximal (0.67) and it differs less from the other vertices, which have closeness-centrality 0.57 (vertices v3 and v4) and 0.40 (v1 and v2). Since the variation of closeness-centrality scores in network B is less than in network A, network B is less centralized. Its closeness-centralization is 0.42 against the maximum centralization of network A (1.00).

Application

Net>Partitions>Degree

In Chapter 3, we explained how to compute the degree of vertices in Pajek. The degree partition contains the simple degree-centrality scores of the vertices and the normalized degree vector contains the degree centrality of the vertices expressed as a proportion of the maximum degree, which is the number of other vertices in the network. If you inspect these scores, you will see that Juan has 13 neighbors, which we already noted in Figure 1, which is more than one third (0.37) of all 35 people with whom he can be connected. In addition to the degree partition, Pajek automatically writes the degree-centralization of the entire network to the Report screen. Here, it is 0.289 which is meaningful only in comparison to other networks.

Net>k-Neighbours

If you want to know the distance between one vertex, e.g. Juan, and all other vertices in the network, you can use the commands in the *Net>k-Neighbours* submenu, which create a partition of classes containing the distances between one vertex and all other vertices. In an undirected network, you may choose the *Input*, *Output*, or *All* command: they yield the same results. The *From Cluster* command is useful if you want to compute distances from a subset of vertices in a large network.

When you execute a *k-Neighbours* command, you must first specify the vertex number or the label of the vertex from which distances will be computed. In the case of Juan, enter 12 (his vertex number) or HM-1 (his vertex label). Next, you can set a limit to the maximum distance which will be computed. In very large networks, setting a limit may speed up computation considerably. In this dialog box, 0 means that you want all distances, which is usually the right choice in the case of a small network. The distances are stored in a partition and unreachable vertices or vertices further away than maximum distance are placed in class number 999998, which indicates that their distance is not known.

In Figure 3, colors indicate the distances of employees to Juan. Most employees are directly connected to Juan (yellow) or indirectly connected with one intermediary (green, distance two). Two employees are four steps away from Juan, namely HP-1 and EM-4.

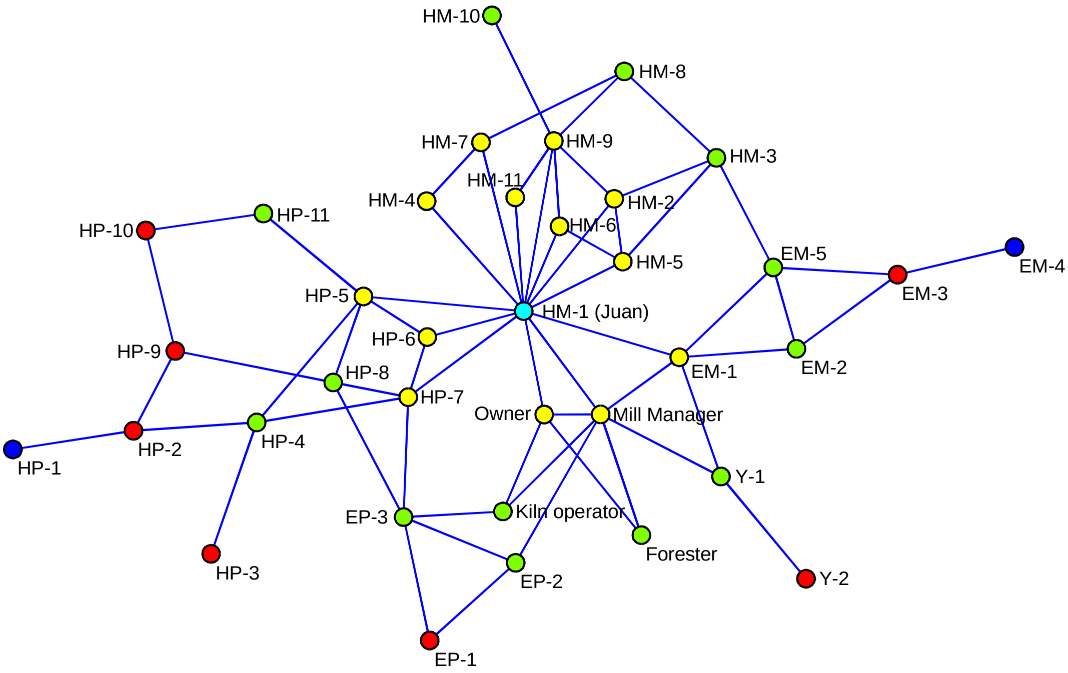


Figure 3 - Distances from Juan.

Net
>Paths between 2 Vertices
>All Shortest

Employees HP-1 and EM-4 seem to be furthest apart in the communication network because their distance to Juan is four. However, their geodesic does not necessarily include Juan, so they may be connected in less than eight steps. In Pajek, the geodesics between two vertices can be found with the command *Net>Paths between 2 Vertices>All Shortest*. In the first dialog box, enter the vertex number or label of HP-1. In the second dialog box, enter EM-4 and subsequently answer *Yes* to the question ‘Forget values of lines’ because you do not want to weight the lines by their values. This is the right thing to do unless line values indicate distances, for instance, geographical distances. Finally, a dialog box asks whether the paths must be identified in the source network. If you answer *Yes* to this question, Pajek will produce a partition for the original network which assigns vertices on the geodesics to class one and other vertices to class zero. Whatever your choice in this dialog box, Pajek creates a new network with the vertices and lines which constitute the geodesics (Figure 4). In addition, it prints the distance in the Report screen. In our example, all geodesics between HP-1 and EM-4 include Juan (see Figure 4), so the distance between HP-1 and EM-4 can not be less than eight, which is the sum of their distances to Juan.

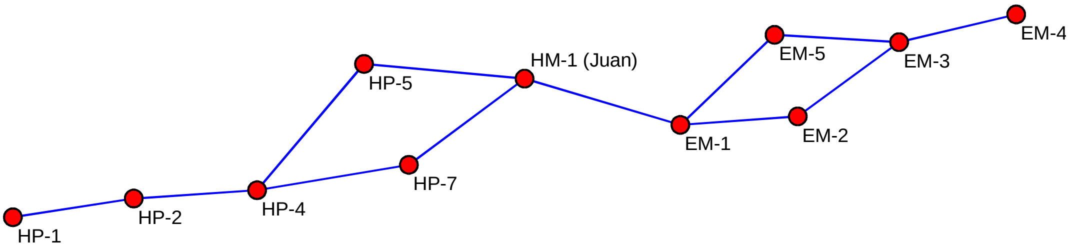


Figure 4 - Geodesics between HP-1 and EM-4.

Net>Vector>Centrality >Closeness

In Pajek, the computation of closeness-centrality is straightforward. The *Net>Vector>Centrality* submenu has commands to compute closeness-centrality for all vertices in the network. For undirected networks, you may choose the command *Input*, *Output*, or *All*, which yield the same results. Note that the network must be (strongly) connected, otherwise Pajek will issue an error message in the Report screen. If the network is not (strongly) connected, closeness-centrality may be computed for each (strong) component separately. In Chapter 2, you learned how to extract a subnetwork and Chapter 3 presented the methods to compute components, so you are able to extract a (strong) component from the network. For medium sized and large networks, closeness-centrality demands a lot of computing time so it should be applied with care.

If the network is (strongly) connected, Pajek creates a vector with the closeness-centrality scores of the vertices. You may inspect this vector or use it for computations in the ways explained in previous chapters. In our example, closeness-centrality scores range from 0.20 to 0.51 and Juan (0.51) turns out to be more central than the manager (0.42). In addition, Pajek computes the closeness-centralization of the network, which is printed in the Report screen. The sawmill communication network has a closeness-centralization score of 0.38, which, again, must be interpreted in comparison to other networks.

6.4 Betweenness

Degree and closeness-centrality are based on the reachability of a person within a network: How easily can information reach a person? A second approach to centrality and centralization rests on the idea that a person is more central if he or she is more important as an intermediary in the communication network. How crucial is a person to the transmission of information through a network? How many flows of information are disrupted or must make detours if a person stops passing on information or disappears from the network? To what extent may a person control the flow of information due to his or her position in the communication network?

This approach is based on the concept of betweenness. The centrality of a person depends on the extent to which s/he is needed as a link in the chains of contacts which facilitate the spread of information within the network. The more a person is a go-between, the more central his or her position in the network. If we consider the geodesics to be the most likely channels for transporting information between actors, an actor who is situated on the geodesics between many pairs of vertices, is very important to the flow of information within the network. This actor is more central.

Juan, for instance, is important to the communication between HP-1 and EM-4 in the sawmill, because all (four) geodesics include Juan (Figure 4). In contrast, HP-5 and HP-7 or EM-2 and EM-5 are less important because if one fails to pass on information, the other may fulfill this role and the communication chain between HP-1 and EM-4 is still intact.

Each pair of vertices may contribute to the betweenness-centrality of a vertex. HP-5 and EM-1, for example, contribute to the betweenness-centrality of Juan, because their geodesic includes Juan. In contrast, the pair HP-4 and HP-5 does not contribute to Juan's betweenness-centrality, because he is not included in their geodesic. In general, we may say that the betweenness-centrality of a vertex is the proportion of all geodesics between other vertices in the network which include this vertex.

The **betweenness-centrality** of a vertex is the proportion of all geodesics between pairs of other vertices which include this vertex.

It is easy to see that the center of a star network (vertex v5 in Figure 2) has maximum betweenness-centrality: all geodesics between pairs of other vertices include this vertex. In contrast, all other vertices have minimum betweenness-centrality (0) because they are not located in between other vertices. The centrality scores of vertices in a star have maximum variation, so the betweenness-centralization of the star is maximal: remove its central vertex and all communication relations are destroyed. In the line-network (B in Figure 2), removal of a vertex may also break the flow of information, but parts of the chain remain intact. Therefore, centrality indices are lower than in the star-network and betweenness-centralization is also lower.

Application

Net>Vector>Centrality>Betweenness

The *Betweenness* command in the *Net>Vector>Centrality* submenu creates a vector of betweenness-centrality scores for the vertices in the network. In addition, the betweenness-centralization of the network is printed in the Report screen. In directed networks, the procedure automatically searches for directed paths, so there are no separate commands for *Input*, *Output*, and *All*. Even in unconnected networks, betweenness-centrality can be computed. This is quite an advantage of betweenness-centrality over closeness-centrality.

The sawmill communication network has a betweenness-centralization of 0.55. Betweenness-centrality scores of employees range from 0.00 to 0.59. In Figure 5, vertex size indicates betweenness-centrality. Several vertices are invisible because they have zero betweenness-centrality: they do not mediate between other vertices. In this example, the betweenness-centrality of vertices varies more than their closeness-centrality because vertices at the outer margin of the network have zero betweenness whereas they are still close to part of the network. As a consequence, betweenness-centralization is higher than closeness-centralization.

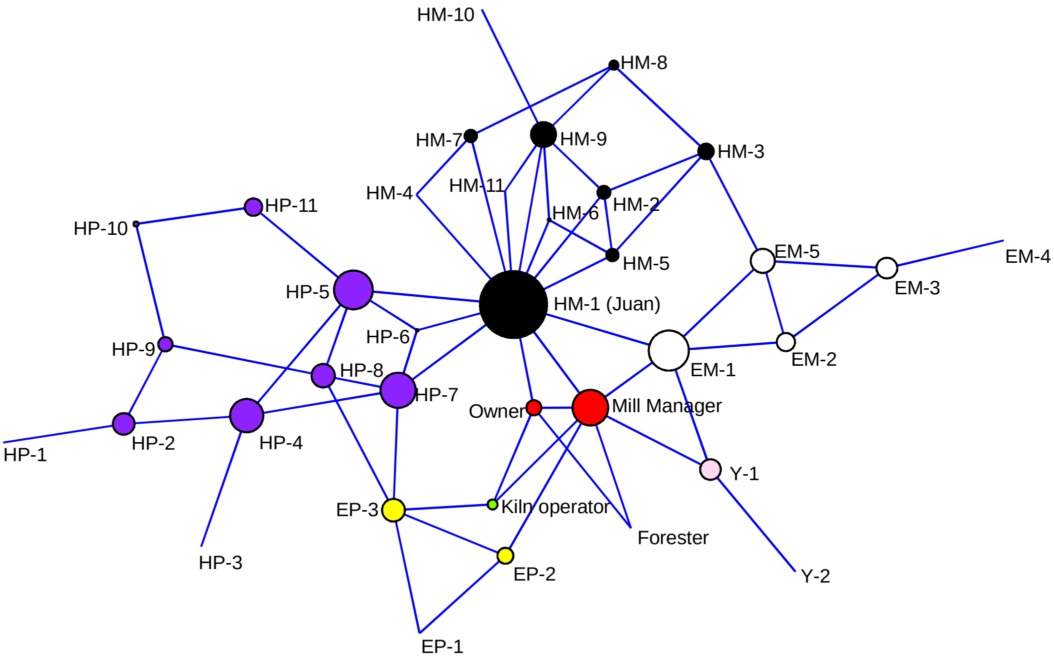


Figure 5 - Betweenness centrality in the sawmill.

It is interesting to note that Juan (0.59), EM-1 (0.20), and HP-5 (0.20) are more central than the manager of the mill (0.17). Each ethnic group within the mill's departments seems to have an informal spokesman who is taking care of the communication with other departments or ethnic groups. Juan, who is the spokesman of the Hispanic employees at the mill (black vertices), is clearly most central.

6.5 Summary

The concepts of vertex centrality and network centralization are best understood by considering undirected communication networks. If social relation are channels which transmit information between people, central people are people who either have quick access to information which circulates in the network or who may control the circulation of information.

The accessibility of information is linked to the concept of distance: if you are closer to the other people in the network, the paths that information has to follow in order to reach you are shorter, and it is easier for you to acquire information. If we take into account direct neighbors only, the number of neighbors (the degree of a vertex in a simple undirected network) is a simple measure of centrality. If we also want to consider indirect contacts, we use closeness-centrality, which measures our distance to all other vertices in the network. The closeness-centrality of a vertex is higher if the total distance to all other vertices is shorter.

The importance of a vertex to the circulation of information is captured by the concept of betweenness-centrality. In this perspective, a person is more central if s/he is a link in more information chains between other people in the network. High betweenness-centrality indicates that a person is an important intermediary in the communication network. Information chains are represented by geodesics and the betweenness-centrality of a vertex is simply the proportion of geodesics between pairs of other vertices which include the vertex. Betweenness-centrality can be computed for unconnected networks, whereas closeness-centrality requires a network to be strongly connected: all vertices must be reachable to all other vertices.

The centralization of a network is higher if it contains very central vertices as well as very peripheral vertices. Network centralization can be computed from the centrality scores of the vertices within the network: more variation in centrality scores means a more centralized network. There is an index of network centralization for each measure of centrality.

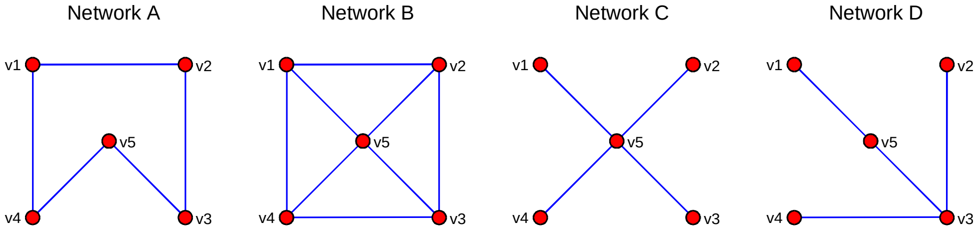
In this book, we apply centrality and centralization only to undirected networks. It is easy to devise centrality measures for directed networks. We could base degree-centrality on the outdegree of vertices, compute closeness-centrality from the distances from a vertex to all other vertices (and not in the reverse direction), and consider only shortest directed paths in the case of betweenness-centrality. In fact, other books on social network analysis advocate such an approach. We think, however, that it is conceptually more clear to restrict centrality and centralization to undirected networks and to apply other concepts, e.g., prestige, to directed networks.

6.6 Exercises

1 Which of the following statements is correct?

- a The centrality of a network may be computed from the degree of its vertices.
- b Centralization depends on the variation of centrality scores.
- c A single vertex is always the center of a network.
- d The center of a network is always a cohesive subgroup.

2 Put the four networks (below) in order of ascending centralization.



   a B, A, C, D.
   b A, D, B, C.
   c D, A, B, C.
   d A, B, D, C.

3 Manually compute the closeness-centrality of vertices v1 and v3 in network D of Exercise 2.

4 The betweenness-centrality of vertex v3 in network D of Exercise 2 is 0.83. List the geodesics which include v3 and the geodesics which do not.

5 The file `exercise5.net` contains the nominations made by 32 employees of an organization who were asked to name the colleagues with whom they discuss work matters. Note that not all nominations are reciprocated. Omit the unilateral nominations, which are less reliable, from the network and find out who is most central in this communication network.

6.7 Assignment

In the 1970s, Rogers and Kincaid studied the diffusion of family planning methods in 24 villages in the Republic of Korea. The files `Korea1.net` and `Korea2.net` contain the communication networks among women in two villages: a village with a successful family planning program (`Korea1.net`) and a village in which family planning was not adopted widely (`Korea2.net`). In both networks, a line indicates that two women discussed family planning. In addition, we know which women adopted family planning methods at least temporarily (class one in the partitions `Korea1_adopters.clu` and `Korea2_adopters.clu`) and which women were members of the local Mothers' Club, which played an important role in the diffusion of family planning methods (class one in the partitions `Korea1_members.clu` and `Korea2_members.clu`) in both networks. The project file `Korea.paj` contains all files.

Analyze the networks and find out whether centrality and centralization are associated with the success of the family planning program in one village and its relative failure in the other village. Try to explain the effects of centrality and centralization by discussing the role of communication in the adoption of family planning methods.

6.8 Further Reading

- The sawmill example is taken from J.H. Michael & J.G. Massey, 'Modeling the communication network in a sawmill' (*Forest Products Journal*, 47 (1997), 25-30).
- The data on the Korean villages stem from E.M. Rogers & D.L. Kincaid, *Communication Networks. Toward a New Paradigm for Research* (New York: The Free Press, 1981), which offers an overview over network analysis from the perspective of communication studies. Note that some of the methods and software packages discussed in the book are obsolete.
- Read more about centrality in J. Scott, *Social network analysis : A handbook* (London: Sage Publications, 1991 (2000)) Chapter 5 and S. Wasserman & K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge: Cambridge University Press, 1994) Chapter 5.

6.9 Answers

6 Answer b is correct: network centralization measures the variation of the centrality of the vertices within the network. The more variation, the easier it is to distinguish between the center and the periphery, the more centralized the network. Answer a is incorrect because centrality is a property of a vertex, not of a network. Answers c and d are incorrect, because the center of a network, if it has one, may either be a single vertex or a cohesive subgroup, e.g., a clique, consisting of a number of vertices which are equally central.

7 Answer d is correct. The star network is most central, so answer a is not correct. In the circle network (network A), all vertices have degree two, each vertex is equally distant from all other vertices, and each vertex is situated on one geodesic between pairs of other vertices, so there is no variation in centrality, hence minimum centralization. Network A is least centralized, so answer c is not correct. Network D is more centralized than network B, so answer d is correct.

8 The distances between vertex v1 and vertices v2, v3, v4, and v5 are 3, 2, 3, and 1 respectively. The sum distance is 9, so the closeness-centrality of vertex v1 is 4 (the number of other vertices) divided by 9, which is 0.44. The sum distance of vertex v3 to v1, v2, v4, and v5 is 2 + 1 + 1 + 1 = 5, so its closeness-centrality is 4 / 5 = 0.8.

9 The geodesics between vertex v1 and v2 (v1-v5-v3-v2), v1 and v4 (v1-v5-v3-v4), the geodesics between v5 and v2 (v5-v3-v2), v5 and v4 (v5-v3-v4), and the geodesic between v2 and v4 (v2-v3-v4) include vertex v3, whereas the geodesic between v1 and v5 (v1-v5) does not. Five out of six geodesics include v3, so its betweenness-centrality is 0.83.

10 The easiest way to omit all unilateral nominations is to change bi-directional arcs into edges (*Net>Transform>Arcs->Edges>Bidirected Only*) and remove the remaining arcs subsequently (*Net>Transform>Remove>all arcs*). Now you have an undirected network and it is easy to compute the three kinds of

centrality (degree, closeness, and betweenness) and display the vertices with highest scores with the *Info>Partition* and *Info>Vector* commands. Vertex v9 has highest degree (17), highest closeness-centrality (0.67), and highest betweenness-centrality (0.20). Person v9 consistently ranks highest on the three centrality indices.