Chapter 13

Observing and Measuring Social Interaction

This book to this point has been focused on what we know about social and economic networks, as well as how we model and analyze social networks and their impact. I close with a chapter about measurement of social networks as well as inference from them. This is not meant to be a primer on empirical work, but rather to provide background on some important issues that are particularly acute when doing empirical analyses involving social networks. Whether or not one performs such analyses, understanding these issues is important in interpreting and evaluating empirical and experimental studies of social and economic networks.

A challenge in working with social network data relates to the fact that social structure is generally endogenous and related to many characteristics of the agents involved. People who are neighbors share many characteristics with their neighbors, as we saw in our discussion of homophily in Section **??**, and they might be choosing their neighbors not just with those characteristics in mind but also with feedback from the outcomes or behaviors that we might be interested in understanding. If we want to establish that a certain behavior is influenced by social network structure, then we have to be sure to properly account for many other related characteristics which could be the driver of the behavior, and to properly account for the fact that there can be feedback between the social structure, behavior, and the background characteristics. This not only relates to properly specifying a model to be able to sort out various effects, but also being sure that these different characteristics are "identifiable" and can really be sorted out given the data. In addition, there is also a somewhat related question of when it is that we can deduce that social structure "causes" behavior, and

not the other way around. I discuss how these difficulties manifest themselves, as well as some techniques for overcoming them.

I also discuss how we can use social network data to discover the latent social structures that underlie many social networks. This is not so much a challenge of social network analysis, but rather an aspect of analysis of social structure that involves some methods that are special to network settings. This includes (stochastic) block modeling, identifying community structures, as well as latent space estimation. The idea is that the nodes of a network belong to groups or organizations that are not directly observed, or are part of an unobserved "spatial" structure. These unobserved social structures determine social networks, and are interesting to understand in their own right. There are various techniques that have been developed for uncovering latent social structures based on social network data, and I provide an overview of some of the techniques and the ideas behind them.

13.1 Specification and Identification

A basic challenge arises in isolating the impact that social networks have on various behaviors, and is related to a more general quest to identify "peer effects". Social relationships are present for a variety of reasons, and social neighbors might display similar behaviors because they are influenced by common traits or experiences. It can be difficult to sort out whether individuals are behaving in a certain way because of the influence of their neighbors, or because of some influences are common to them and their neighbors, or because of other factors that are related to their network position.

13.1.1 Specification and Omitted Variables

To get a feeling for this, let us reconsider the study of Coleman, Katz, and Menzel [154] discussed in Section **??**. Van den Bulte and Lilien [600] criticize the study saying that it did not properly account for the effects that marketing and advertizing efforts by various pharmeceutical firms had on the diffusion process. To the extent that the direct effects of marketing or advertising, or other market forces, are correlated with the degree that Coleman, Katz, and Menzel measured, then it could be that these other factors are responsible for the diffusion and the degree ends up appearing to be the driver of diffusion because it is correlated with the other factors. Further studies (e.g., see Bhatia, Manchanda and Nair [56]) have found evidence of such effects after

controlling for marketing and other characteristics. Nevertheless, the point that one needs to include all the relevant factors in the analysis is an important one.

As a very simple example, suppose that degree in a network is correlated with age, so that older people have more connections. Now suppose that one hypothesizes that people with higher degree behave differently from people with lower degree. So, one estimates an equation of the form

$$Y_i = \alpha_d + \beta_d d_i + \varepsilon_i, \tag{13.1}$$

where Y_i is a measure of the behavior that we are interested in explaining, d_i is individual i's degree, and ε_i is an error term that captures unobserved idiosyncratic factors. However, suppose, for instance, that degree varies with age, so that

$$d_i = \delta + \gamma age_i + \eta_i, \tag{13.2}$$

where α_d is a base degree and η_i is an idiosyncratic term. If the true relationship is in fact of the form,

$$Y_i = \alpha_a + \beta_a age_i + \nu_i, \tag{13.3}$$

then when we fit (13.1), given (13.2), we can end up seeing an estimate for the coefficient on degree, β_d, that looks significant, but is really just a proxy for the effect of age, which is filtered through degree.

Suppose instead that the true relationship is of the form

$$Y_i = \alpha_d + \beta_d d_i + \beta_a age_i + \varepsilon_i, \tag{13.4}$$

so that both degree and age affect the variable in question. For example, suppose that we are examining a model of social capital, and so Y_i is some measure of wealth or power accumulation, and we are testing for how this is influenced by a measure of social connectedness. If we then fit (13.1), including degree but omitting age, then given that degree and age are related according to (13.2), we will end up with a biased estimate of β_d as it proxies for some of the effect of age. For instance, with positive relationships between each of the variables, by omitting age we would end up with an inflated estimate of how degree affects wealth.

While such specification and omitted variable problems are not unique to analysis of social networks, they are particularly acute in social network settings because of homophily. Recalling the discussion from Section **??**, we know that people tend to associate with others who are similar when examined from a wide variety of perspectives.

So, if we see that behavior of individuals tends to match patterns of the social network, we cannot attribute that behavior directly to network influence without being sure to properly account for all of the other factors that might be related to whom interacts with whom.

Instrumental Variables

It is generally difficult to be sure of all of the potential factors that might covary with both social relationships and behavior, and so there can always be a lurking issue with omitted variables and other sorts of misspecifications in estimating the impact of social structure on behavior. One way to view the bias that emerges in estimation is that the omission or misspecification leads the variables that we wish to estimate effects of to be correlated with the error term, and that can result in the bias in estimation (and can lead the estimator to be inconsistent so that even large data sets will not overcome the problem).

A standard approach to dealing with such problems is to work with instrumental variables. To develop an intuitive understanding of instrumental variable methods, let us understand what goes wrong in the case of an omitted variable. Ideally, to estimate the effect of degree on wealth, we would like to have a set of observations where degree is varied, and other variables are held constant in order to isolate the marginal effect of degree on wealth. If we design controlled experiment, this is how we would set things up. The difficulty is that we do not have such control, and when degree is high, the omitted variable of age also tends to be high, and when degree is low, the omitted variable of age also tends to be low. The idea of an instrumental variable is that it is one that is correlated with degree, but uncorrelated with the error term, or in this case, uncorrelated with the omitted variable of age. If we find a variable that covaries with degree, but is uncorrelated with the error term (and in this case with the omitted variables), then that serves as a substitute for the controlled experiment that we would have liked to design. For example, we might find some sort of factor that leads to changes in degree, such as a club that opens branches in some communities but not others (for exogenous reasons), but in a way that we expect to be uncorrelated with things like age or anything else that might lead to correlation with the error or omitted variables. This variable becomes the instrument. Effectively, it provides the sort of experiment that we would have liked to have set up, but causing variation in degree in ways that are independent of other possible explanatory factors.

One can then derive estimates based on the instruments using standard techniques

developed for this purpose or we can use two-stage least squares estimation. That is, we first regress degree on the instrumental variable, to develop estimated values of degree based on the instrument. Then in the place of degree in the original regression we use these estimated values of degree as predicted by the instrument, which are then independent of the error terms as they are conditional on the instrument. In this way we obtain a consistent estimator.

Endogeneity

Another problem that arises in working with estimation based on social structure is that social structure is often endogenous to the things that it impacts. For example, in estimating the impact of social connectedness on wealth, it could also be that wealth impacts connectedness; e.g., by providing additional access or opportunities.[1] While this is a distinct problem from the specification and omitted variables problem discussed above, it could also lead degree to be correlated with the error term, and can also lead to biased and inconsistent estimation.

Instrumental variables can also be helpful in sorting out endogeneity issues. Again, we want an instrumental variable that is related to degree, but will not be related to the error terms. Thus, it is a variable that avoids being endogenous, but is related to degree, and so when it leads to high or low degree we are sure that the high or low degree is coming about for exogenous reasons. For example, if a government came in and randomly chooses some villages to provide subsidized communication to (in the form of internet or telephone, etc.) which might enhance degree in some villages and not others, then we could examine how these changes in degree for exogenous reasons influence wealth. More specifically, we could measure the induced degree differences that we attribute to the government program and then see the extent to which it influences wealth, again using two-stage least squares, or other methods of working with instrumental variables.

While instrumental variables are a very useful tool for dealing with the omitted variable and endogeneity problems, they are not a panacea. First, finding "good" instruments, that are related to the problematic variable(s), but are not endogenous and not related to omitted variables and error terms, can be problematic. In fact, we can never really be sure of this as we never directly observe the "true" error terms. Second, even if we find instruments that we are reasonably confident in, it can still

[1] See Durlauf [**?**] for more discussion of some of the endogeneity issues associated with studying social capital.

be difficult to find ones that have a strong enough relationship with the problematic variables to be very useful in producing powerful estimates.

The example above of a government program that randomly selects certain villages for improvements that could lead to changes in social capital is a form of a "natural experiment." That is, it sets up nice control situations that provides us with just the sort of variation that is useful, so that we see what happens with and without the program, and thus with and without some exogenously determined increase in social capital.[2] Often ideal instrumental variables are related to such exogenous variation, when we are lucky enough to find them. We saw examples of such natural experiments in the discussion in Section **??** of the empirical analyses of how network connections shape labor market outcomes, where various exogenous variations in social structure, such as immigration due to rainfall in the home country, or random assignment to military units, or assignment to a city by an agency for exogenous reasons.

13.1.2 The Reflection Problem and Identification

Another problem in specifying a model social influence on behavior has to do with making sure that the specified relationship is properly "identified," meaning that the parameters of the model are uniquely determined by a data set. To caricature the identification problem in social influence: it arises from the fact that behavior is being determined by behavior, and so there is a sort of circularity, and under some specifications it can be difficult to sort out the influence of behavior on behavior. A well-known illustration of this is due to Manski **??**, [425].

Manski **??**, [425] describes an identification problem that is faced when examining social influences on individual behavior and provides an useful paradigm for understanding identification problems in such settings.

To see the identification problem that Manski [424] refers to as the reflection problem, let us consider it in its starkest form just to get a clear understanding of it. Consider a situation where the behavior of an agent is a linear function of the average level of the behavior taken by other members of his or her cohort. To be specific, let us suppose that an agent i has some characteristics x_i and that the agent's cohort is other people who have the same attributes. So, what determines i's behavior is what he or she expects his or her peers to do. Let Y_i denote i's behavior. His or her expectation

[2]For an interesting example of an analysis of social capital, based on the presence of television (which depends on some exogenous programs and geography) in rural Indonesian villages, see Olken [488].

of what people with the same attributes x_i will do is simply $E[Y_i|x_i]$ and so we write the relationship as

$$Y_i = a + bE[Y_i|x_i] + \varepsilon_i \tag{13.5}$$

where ε_i is a zero-mean random variable (conditional on x_i) which is a noise term capturing some unmeasured idiosyncracies. What will happen if we try to estimate this relationship? Note that if we take the expectation of both sides conditional on i's attributes x_i, then we end up with

$$E[Y_i|x_i] = a + bE[Y_i|x_i].$$

This has a unique solution (provided $E[Y_i|x_i]$ is nonzero) which is $a = 0$ and $b = 1$. Thus, when we fit (13.5) we end up with a tautological relationship of

$$Y_i = E[Y_i|x_i] + \varepsilon_i$$

which simply says that Y_i is its expectation plus noise, which tells us nothing about endogenous social interaction.

This is what Manski refers to as the "reflection problem:" i's behavior is a function of the expectation of the peers' behaviors which is just the expectation of i's behavior, which reflects i's behavior. Effectively, we have not specified a system that has any real bite in terms of tying down the relationship, and so it is not properly identified.[3]

Even if we enrich the specification, to allow Y_i to also depend on i's attributes x_i directly, we still have an identification problem. That is, suppose that we specify that

$$Y_i = a + b_1E[Y_i|x_i] + b_2x_i + \varepsilon_i. \tag{13.6}$$

Again, taking expectations,

$$E[Y_i|x_i] = a + b_1E[Y_i|x_i] + b_2x_i.$$

If $b_1 \neq 1$ (so we avoid having this parameter predetermined and tautological), then we can rewrite this as

$$E[Y_i|x_i] = \frac{a}{1-b_1} + \frac{b_2}{1-b_1}x_i.$$

[3]The identification problem is different from a multiple equilibrium problem, which can also result in settings with interdependencies in behaviors. As we have seen in graphical games with strategic complementarities, there can be multiple equilibria. The multiple equilibrium problem its own challenges.

Substituting this back into (13.6), we obtain

$$Y_i = \frac{a}{1-b_1} + \frac{b_2}{1-b_1} x_i + \varepsilon_i \tag{13.7}$$

Presuming that x_i is not constant, then we can estimate the two composite parameters $\frac{a}{1-b_1}$ and $\frac{b_2}{1-b_1}$. However, this does not identify the parameters a, b_1 and b_2, as there are many different values of these which lead to the same composites. We could set b_1 to any value (other than 1) and find values of a and b_2 that are consistent with the composite parameters. Adding in extra variables, does not help, as although there are more variables, there also more parameters to identify, as discussed in Exercise 13.1.

As Manski points out, one remedy is to use instrumental variables to sort out the identification, since part of the difficulty stems from the fact that behavior is endogenous, entering on both sides of the equations in the model. There are a variety of other ways around identification problems in social network settings. Let us examine several of them.

Social Structure and Identification

Note that the reflection problem stated above stems from the fact that i's peers are not identified directly, but just assumed to be similar to i. We are ignoring any real social structure and information that might be available about i's neighbors. If we explicitly track i's neighbors, then that information can identify a model.[4] To see one way that this works, let a possibly weighted and directed network matrix g govern interaction. Also, to make the technique as transparent as possible, ignore constant terms as well as any node-specific characteristics, so that we are just working directly with the interaction of behaviors, and let us stay with a linear relationship. In particular, suppose that

$$Y_i = b \sum_j g_{ij} Y_j + \varepsilon_i. \tag{13.8}$$

Thus, each individual's behavior is a weighted average of his or her neighbors' behavior. If the g_{ij}'s are not degenerate,[5] so that $(I - bg)$ is invertible, then we can express this as

$$Y = (I\!I - bg)^{-1} \varepsilon, \tag{13.9}$$

[4] The techniques here are common to the spatial econometrics literature, as outlined in Ansel [17]. The specifics of the derivation that follows in this section are due to Marcel Fafchamps, who showed me how this approach adapts to network settings and pointed me to the related references.

[5] This is relative to b, but generically in g the matrix will be invertible for any given b.

where Y and ε are the corresponding vectors, and $I\!I$ is the identity matrix. Letting ^{T} denote transpose, it follows that

$$E\left[YY^T\right] = (I\!I - bg)^{-1} E\left[\varepsilon\varepsilon^T\right] \left((I\!I - bg)^T\right)^{-1}. \tag{13.10}$$

If we have knowledge of the social network matrix g and of the covariance matrix of the error terms $E\left[\varepsilon\varepsilon^T\right]$, for instance, using a standard assumption that errors are independently and identically distributed with a finite variance, then b will be identified by equation (13.10). ident22).

This technique relies on some knowledge of the error terms. This can be a problem, especially in situations where we worry that omitted variables could influence behavior; for instance, resulting in positive correlation in the errors. This does not close the door on identification, as the above specification does not take advantage of other factors that influencing behavior.

In particular, in the reflection problem i's peers' behavior were not identified other than through i's characteristics. If we have explicit information about i's peers, and they differ from i in characteristics and the social relationships are not entirely symmetric, then this can overcome the "reflection" problem. As an extreme example, just to illustrate the point, suppose that there are different types of agents, say young and old. Suppose that older agents do not pay attention to the younger agents, but the younger agents pay attention to the older agents. In that case, a straightforward regression can estimate older agents' behavior, and then this can be used to estimate the younger agents' behavioral relationship. While this is an extreme example, where there are some agents who are not influenced by social interaction, more generally some asymmetries in the ways in which agents are influenced by each other is enough to provide identification. The reflection problem noted above was an extreme case where an agent was only influenced by agents whose actions were forecasted by the agent's own characteristics, and so variation in background characteristics were not helpful in identifying the relative behaviors and their impacts on each other.

Nonlinearities in Social Interaction

Beyond the fact that the reflection problem can be overcome with richer information in terms of interaction, it also is important to note that even without such information it derives from the linear-in-means specification. To see how linearity matters, let us modify the specification in (13.5) so that instead of having an agent's behavior be

proportion to the mean of his or her peers, where $Y_i = a + bE[Y_i|x_i] + \varepsilon_i$, let i's behavior be influenced by the maximal action by his or her peers. For example, let

$$Y_i = a + b_1 E[\max_{j \in N_i} Y_j | x_i] + b_2 x_i + \varepsilon_i, \tag{13.11}$$

where N_i are the neighbors of i, whose behavior might still be estimated based on x_i or observed more directly as a social network. Now, if we take the expectation of each side, we no longer have a tautology, but instead have

$$E[Y_i|x_i] = a + b_1 E[\max_{j \in N_i} Y_j | x_i] + b_2 x_i,$$

which is identified as long as we can deduce $E[\max_{j \in N_i} Y_j | x_i]$ in a way such that it and x_i are not constant nor linearly dependent.

Note that is was not essential that the specification be based on the max. There are a wide variety of specifications that avoid the identification problems (see Brock and Durlauf [100] for more discussion and specifications). In fact, the identification problems only arise for very particular specifications, such as the simple linear/average behavior specification.[6] This emphasizes the importance of building a model that properly captures the social interaction structure and incentives that are at the heart of the particular application. In fact, chapters **??** and **??** provide a base for modeling social interaction in ways that will generally be nonlinear and identifiable.

Timing

Another aspect of specification is timing. In the reflection problem, part of the difficulty stems from the contemporaneous interaction of the decisions. If decisions are repleatedly taken over time, then it could be that an agent's decision depends on the observations of *past* decisions of his or her peers. This can help identify the relationship, and can also help sort out causation. One can see if the current decisions of a given individual vary with the past decisions of his or her peers. For example, Conley and Udry [**?**] examine the timing of past neighbors' success in experimenting with fertilizer in pineapple production in Ghana to test for social learning. This is still quite challenging as there can still be omitted variables that are at the root of adoption. However, careful attention to the timing of behavior allows Conley and Udry to check whether increases in use by one agent are triggered by past successes of that agent's neighbor(s), after properly sorting out other potential causes.

[6] One does need to be careful however. Nonlinear models can also face identification problems as discussed in Exercise [**?**].

It is important to note that taking advantage of timing has been used not only to uncover whether an agent's behavior is influenced by his or her peers and social neighbors, but also to examine the endogeneity of social structure. For example is it that an agent acts in a certain way because his or her friends do, or is it that he or she selects friends that act in similar ways? This sort of endogeneity question is fundamental to social network analysis, especially when examining peer effects, and timing can be used to help sort out these effects. For example, Kandel [**?**] uses time series data on high school friendships together with data on individual drug use, delinquency, political opinion, and educational aspirations, to examine whether friendships tend to form among agents with similar behaviors, or whether agents' behaviors are influenced by that of their friends.[7] By examining the formation of new friendships and the deterioration of old friendships together with behavior over time, Kandel is able to identify whether either or both of these effects are present. The timing helps her show that both effects are present. Agents are relatively more likely to form new friendships with others whose past behavior matches their own, and to sever past friendships with others whose past behavior differs from theirs. In addition, when agents' behaviors change, it tends to be influenced by the past behavior of the agents who remain their friends. Moreover, the magnitude of the effects differs across behaviors. Such an analysis is demanding in terms of the richness of the data that is necessary, especially if one also wants to control for other variables which might be adjusting over time and co-varying with behavior and or social structure. Nevertheless, longitudinal (time series) data is a powerful tool for sorting out both social influence and endogeneity.

13.1.3 Laboratory and Field Experiments

As we have seen in the previous sections, many of the challenges in empirical analysis of social structure and social interaction are in separating out effects to see which factors and behaviors are related to which others, and to begin to understand which ones might cause which others. Various forms of experiments on behavior, where the particular values of some variables are carefully controlled and varied, allow one to track which conditions have been altered and thus uncover their impact. Given the power of experiments, they are a rapidly emerging tool of network analysis.[8]

Experiments can be used in various ways. First, an experimenter can examine

[7] For alternative methods of studying the co-evolution of networks and behavior see Snijders, Steglich, and Schweinberger [575].

[8] For a survey of parts of the literature see Kosfeld [388].

how social structure affects behavior, and how an agent's position within a network influences his or her behavior. For example, one can put agents in specific network contexts, such as in an exchange or bargaining network in the experiments of Cook and Emerson [161] and Charness, Corominas-Bosch, and Frechette [135]. Then comparing behavior across network position, as well as tracking changes in behavior as one changes overall network structure, allows one to explicitly track the impact of social network structure on behavior. There the idea is that the only variable that is altered is the social structure, which helps show whether or not agents behave differently when embedded in different positions within a network or in different networks. One can also examine something like how well agents are able to communicate and learn in social contexts, and how that depends on the network structure (e.g., see Bavelas [44], Leavitt [405], Choi, Gale and Kariv [141]), or how well agents are able to coordinate their actions in a social network (e.g., see Kearns, Suri, Montfort [365]). One can also investigate whether or not agents will behave differently if the network that they are part of is exogenously imposed or chosen by them (see Corbae and Duffy [165] and Riedl and Ule [**?**]). The control present in the experimental setting has proven to be a very useful tool for identifying network effects.

Second, one can test alternative theories of behavior in network contexts. Here one does not necessarily need to use variations in treatments within an experiment; instead one can simply find a setting where the predictions of different theories lead to different outcomes. For example, one can test models of network formation such as "myopic" ones like pairwise stability against a "farsighted" notion, as in Pantz and Ziegelmeyer [497]. There, the payoffs to network formation can be fixed, but are chosen by the experimenter so that one should observe different networks form if agents are farsighted in their link formation rather than myopic. Alternatively, one does not necessarily have to run a horse race of different theories against each other, but one can simply test whether a given theory's predictions hold. For example, one can examine variations on undirected or directed connections model and then see whether pairwise stable or nash stable networks form (e.g., see Callander and Plott [112], Falk and Kosfeld [224], Deck and Johnson [178], Vonin [**?**], and Goeree, Riedl, and Ule [276]).

While the first approach above was useful because it allows one to adjust some aspect of the social setting and then isolate its affect on outcomes, the advantage of the second approach is different. Under the second approach one only needs a single treatment, but one that distinguishes between different theories or hypothesized behaviors. Here the power of the experiment comes from being able to impose a certain

structure, and be sure that one has more or less complete knowledge of the network of interactions or of the relevant payoffs, etc. These can be difficult to observe from field data.

Beyond laboratory experiments, field experiments can be a very useful tool as well. One loses some of the control of a laboratory environment, but gains access to a social network in its natural setting. Milgram's small-world experiments, and those that have followed, are examples of this technique. There, the objective was to discover things about a real social network (and how people navigate it), and the controlled aspect was giving subjects a particular task to perform in a way to elicit information about network structure, such as the distance between two agents. Such field experiments can be useful not only for discovering social network structure, but also for seeing how it influences behavior. For instance, Goeree et al [277] examine how players behave in a dictator game (where a player has a choice of how much of a sum of money to keep and how much to give to another player) as a function of the social network distance between the players. One can also sort through various theories of social capital, altruism, and reciprocity, by examining specific behaviors as a function of social network (e.g., see Leider et al [407]). In such field experiments the control of a task, game, or information seeding, and so forth, placed in the context of a (measured) real social network helps uncover how behavior is related to social structure.

13.2 Community Structures, Block Models, and Latent Spaces

Another aspect of empirical investigation that is special to social network analysis is uncovering the latent social structure that underlies a network and led to its formation. Such structure is often not fully observable, and so needs to be constructed from what is observed. Uncovering such social structure can be useful for a variety of reasons. For example, coupling such implicit structure with other attributes can help to investigate whether there are specific biases in a society, such as in hiring or publishing. It can also help in classifying and categorizing political and other ideologies, as well as economic patterns of behavior.

The models that we have seen in earlier chapters for how networks are formed, when fitted to network data, can provide some insight into underlying structure, especially in terms of fitting some of the strategic models to data. But beyond these, there are also

various algorithmic and statistical methods which are specifically designed to discover underlying social structures or groupings. Let us examine those.

13.2.1 Communities and Blocks

There are various ways in which the structure underlying a social network can be modeled. One basic and standard way is to presume that the nodes of the network belong to different blocks or communities. The network that emerges depends on the underlying blocks or communities, which in turn can be recovered by examining the network. The early literature provided some definitions that can be used to capture underlying structures and the subsequent literature has provided a number of algorithms for recovering them.

One notion of a block or community is that it consists of nodes that are somehow comparable or equivalent. An early concept of this sort is structural equivalence, where two nodes are said to be structurally equivalent if their relationships to all other nodes are identical. This was first discussed by Lorrain and White [**?**] and further by White, Boorman and Breiger [628]. In particular, two nodes i and j are *structurally equivalent* relative to a network g if $g_{ik} = g_{jk}$ for all $k \neq i$, $k \neq j$ (and the same for g_{ki} and g_{kj}. This equivalence notion leads equivalence classes of nodes, so that we can partition the set of nodes into sets of equivalent nodes.

It is clearly rare to find networks where many pairs of nodes are structurally equivalent as many factors affect network formation, resulting in substantial noise both in actual and observed relationships. In view of this, the literature has moved beyond such a strict definition to develop a variety of methods of grouping nodes into equivalence classes and defining how nodes relate to each other, and also to develop methods of uncovering some basic blocks which underly the society but are not directly observed.[9]

In order to discuss such methods, I begin with a formal definition of such a grouping of nodes into equivalence classes, called a community structure.

A community structure is a partition of the set of nodes N, Π.[10] Thus, it groups the set of nodes into separate communities.

For example, consider the nodes in the network pictured in Figure 13.2.1.

One community structure of the example in Figure 13.2.1 is pictured in Figure 13.2.1, where the ovals collect the nodes considered to be in the same community.

[9] For an overview of blockmodeling see Doreian, Batagelj, and Ferligoj [190].

[10] Recall that a partition, π, is a collection of disjoint subsets of N whose union is N.

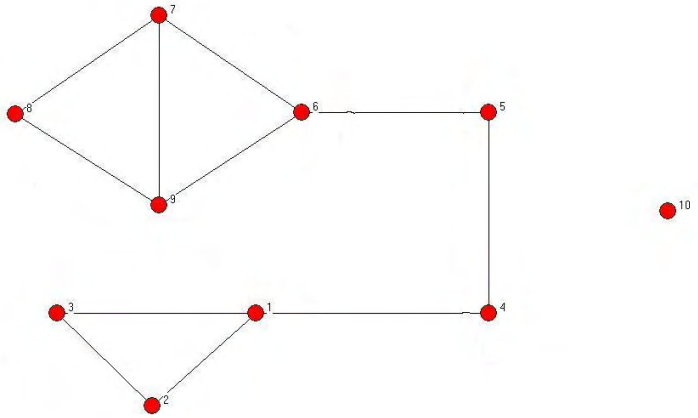

Figure 13.2.1. *A Sample Network.*

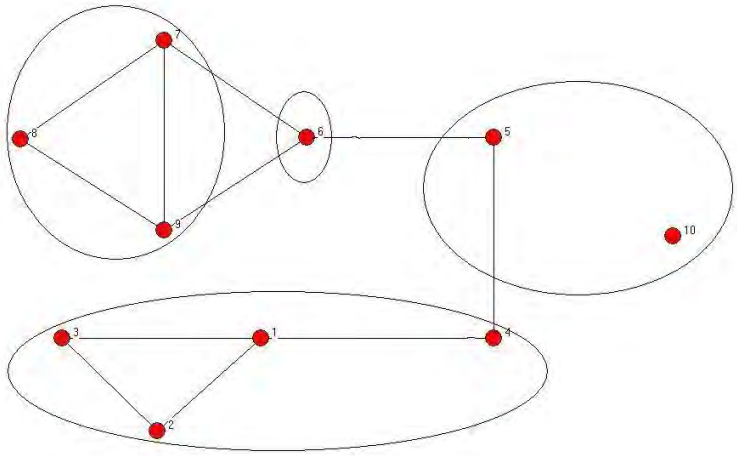

Figure 13.2.1. *A community structure for the network in Figure 13.2.1.*

There are many different ways in which the nodes of any network might be partitioned, and which community structure seems most appropriate can depend on the context and also on what we imagine a community to represent. Let us examine some of the more prominent methods for identifying community structures.

13.2.2 Methods for Identifying Community Structures

Just as we saw the proliferation of measurements of centrality and power, there are many different ways to think about what it means to say that two nodes are equivalent or belong in the same equivalence class or "community." I discuss a few prominent approaches of classifying nodes into equivalence classes, chosen to give a feel for the spectrum of potential viewpoints.[11]

Let me begin with a criticism of some of the literature, which is important to keep in mind when examining the techniques presented below. In order to make sense of a community structure, we should have a well-specified notion of what a community represents. For example is it based on some common but unobserved traits of nodes? Is it defined by some factors that influence nodes' behaviors? Is it defined by some affinity that agents feel for each other? Is it meant to capture some natural complementarity in association that is not directly observed but which favors link formation? As we vary what a community represents, the optimal method for identifying communities will correspondingly change. In particular, it is important to have an idea of how community structure will affect network formation. As if we do not have an idea of how community structure influences the observed network, then it is difficult to know how to recover the community structure from the observed network. Unfortunately, much of the literature by taking a "I will know a community when I see it" approach. That is, the literature has generally started with a simple algorithm for partitioning nodes of a network, based on some heuristic, without a firm foundation in terms of defining what communities are, how they influence network formation, or why this algorithm is a natural way for uncovering them. Thus, communities have tended to be defined to be whatever the algorithms find rather than deriving the algorithms as a technique for identifying a well-defined notion of community. That is not to say that we could not develop a well-defined notion of community corresponding to each technique, but rather that the literature has generally failed to be careful about this

[11] Wasserman and Faust [615], Snijders and Nowicki [574], and Newman [**?**], [**?**] provide additional background on various parts of this literature.

point and has not proceeded in as scientific a manner as one might like. At the end of this chapter I return to discuss some approaches that are built from the ground up. Hopefully, such techniques will be proliferate, and there will be a re-examination of the many techniques for identifying community structures with more attention paid to what communities represent and how they relate to network structure.[12]

CONCOR

An early and widely used method for partitioning nodes into communities is called CONCOR (for "convergence of iterated correlations"), as developed by Breiger, Boorman, and Arabie [87].[13] The idea is as follows.

Start with an observed social network described by an adjacency matrix g, which can be either directed or undirected. Two nodes are thought of as being similar if they have a similar pattern of relationships with other nodes. One way to gauge how similar node i is to node j in terms of the network of relationships is to examine how similar row $(g_{i1}, \ldots, g_{in})$ is to row $(g_{j1}, \ldots, g_{jn})$.[14] A measure of how similar these rows are to each other is to examine the correlation between row i and row j.[15]

The CONCOR algorithm does not stop here. This first step leads to a correlation matrix C, where c_{ij} is the correlation between row g_i and row g_j of the adjacency matrix. Next, let us measure how similar row i of the correlation matrix C is to row j, via the same method. If nodes i and j are similar, then they should have similar rows of correlations with other nodes. So, the algorithm then measures the correlation between rows c_i and c_j of the correlation matrix. In this way, we form a new correlation matrix $C^{(2)}$. Iterating, the algorithm generates a matrix of correlations of correlations

[12] To varying extents, the same criticism can be made of many measures in social network analysis. For example, there is much still to be learned about what different measures of centrality, prestige, clustering, ..., really capture, beyond some of the heuristics discussed in Chapter **??**.

[13] Breiger, Boorman, and Arabie provide references to earlier versions of CONCOR that were used in some specific studies, and CONCOR was further elaborated upon by White, Boorman and Breiger [628] and Boorman and White [86].

[14] In the case of directed networks, this misses some of the action as it focuses on relationships that are directed outward, while in the undirected case it captures all of a node's relationships. One can also perform the described algorithm on columns. Using columns would not change the analysis in undirected networks, but could lead to a very different analysis in the case of a directed network as it focuses on inward relationships, which can differ dramatically from outward ones.

[15] That is, view the row g_i as a random variable that takes on value g_{ik} if state k is realized. Placing equal likelihood on the n states, the correlation described above is simply the correlation between g_i and g_j.

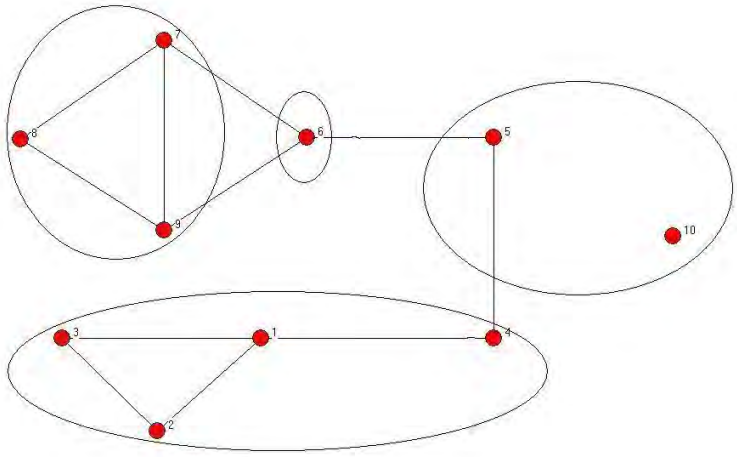

Figure 13.2.2. *The community structure found via CONCOR on the network in Figure 13.2.1 when looking for four communities.*

of..., denoted $C^{(t)}$, after t iterations. Generally, this converges as t grows, so that the entries of $C^{(t)}$ approach some limit. In fact, except in exceptional cases, this process converges to a matrix of (at most) two blocks where the entries are 1 and -1.[16] That is, two blocks (which partition the nodes into two groups) emerge such that $C_{ij}^{(t)}$ converges to 1 when i and j are in the same block and $C_{ij}^{(t)}$ converges to -1 when i and j are in separate blocks.

As this only produces two blocks, one can repeat the procedure on the network starting with each block separately to further subdivide the blocks. By choosing when to stop subdividing, one ends up with a community structure. To see how this works, let us pre-specify that we wish to find a community structure with four communities. Then we operate CONCOR once to find two communities. Then we operate it again on the resulting communities to find four communities. For example, in this way CONCOR finds the structure in Figure 13.2.2.

While CONCOR and its variations are included in many programs for network

[16] See Schwartz [**?**] for background on convergence properties of CONCOR and related methods of iterated covariance.

analysis, what one actually obtains through iterating on the correlation is not so obvious.[17] That is, what does it really mean for two nodes to be in the same block under this procedure? Why is this a reasonable way to group nodes? Another difficulty is that it is not obvious how many times to split the blocks.

Repeated Bisection

CONCOR is a method that repeatedly bisects the set of nodes based on a specific technique of grouping nodes. There are other methods, which developed out of the computer science literature, that also work by repeated bisection. For example, a quite simple principle is to simply start by bisecting the set of nodes into two groups such that there is a minimal number of links between the two groups, with some rule based on what to do in the case where more than one bisection leads to the same minimum. Then one can repeat the procedure on the emergent groups. Again, using some stopping rule, one ends up with a community structure.

There are many variations on this basic idea. Instead of minimizing the number of links between the two groups, one might try to maximize some measure of how many links exist within each group, less how many go across the groups (e.g., see Kernighan and Lin [368]). Or, one might rely on more sophisticated methods where one examines the eigenvectors of the adjacency matrix or an associated Laplacian matrix (e.g., see Fiedler [230] and Pothen, Simon and Liou [520]).[18]

Beyond the question of the how one measures how good a bisection is, there is also a question of whether one searches over all possible bisections or limits attention to certain types of bisections. For example one can pre-specify the sizes of the two sets of nodes. For example, one might only consider bisections into equal (or nearly equal) sized sets of nodes. This is not a minor technicality, as it can lead to a sizeable difference in the number of potential bisections that one needs to consider. Also, many of these methods would tend to bisect the network very asymmetrically if no constraints are imposed, since, for instance, minimizing the links between groups might be found by identifying a single node with low degree and separating that from the rest of the network, and it is not clear that this would be sensible. This relates back to the earlier criticism that we are not quite sure what we are looking for, and so the answer to this question is not tied down.

In addition to measuring the optimality of a bisection, and which bisections to

[17] See Schwartz [552] for a criticism along these lines.

[18] See Newman [483] for an overview of some of those techniques.

consider, there is the stopping decision which we saw can be somewhat arbitrary. Finally, one can also move beyond bisections. For example, one can start with the problem of splitting the set of nodes into some number of (nearly) equal sized groups in a way that minimizes the links across groups and/or maximizes the number of links within groups.

Edge Removal

One method that avoids the issue of predetermining the sizes of the groups to bisect a network into, instead works by repeatedly removing edges of the network and keeping track of the component structure of the resulting graph to determine a community structure. Such a method, developed by Girvan and Newman [271], iteratively removes edges by calculating the betweenness of each link and then removing the link that has maximum betweenness. The logic is that if a link has a very high betweenness score, then it is connecting (at least) two groups of nodes that otherwise are quite separate. These groups would then be natural candidates to be separate communities of nodes. This process can be done in various ways, depending on the notion of betweenness used.

Let us demonstrate this technique on the example from Figure 13.2.1, using the easy-to-calculate measure of the betweenness of a link that Girvan and Newman use (a variation on Freeman's notion discussed in Section [**?**]), which is the number shortest paths between pairs of nodes in the network that involve the link in question. That is, for each link we count the total number of geodesics in the network that include that link. The link with the highest count is the one we remove. (It could be that a pair of nodes has more than one shortest path that involves the same link. For example, nodes 8 and 5 have two shortest paths between them that involve the link 56.) Once a link is removed, one repeats the process on the resulting subgraph, calculating new betweenness scores at each step.

Figure 13.2.2 shows the resulting community structure under this algorithm if one stops when the betweenness of an edge reaches 2 (so that the highest betweenness measure of any link remaining in the network is 2; see Exercise 13.3 for the precise steps).

The calculations involved in order to run the Girvan and Newman algorithm can become extensive as the number of nodes grows, as the betweenness measures involve finding all shortest paths through a given link. In view of this, there are various alternative measures and algorithms that have been developed (see Newman [483] for

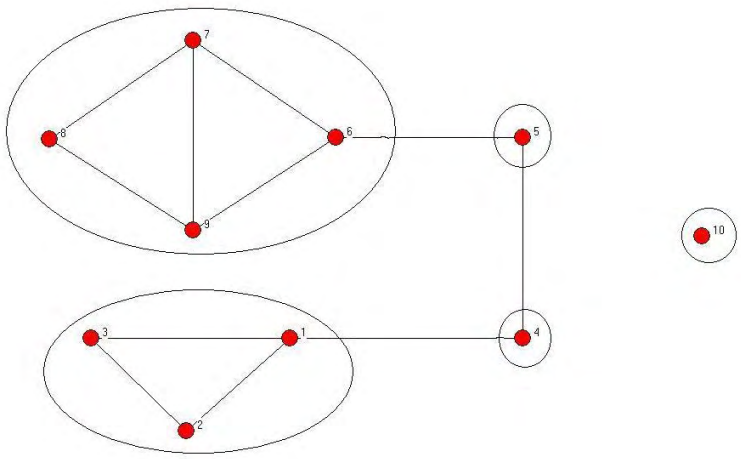

Figure 13.2.2. *The community structure found via the Girvan-Newman Algorithm on the network in Figure 13.2.1.*

discussion and references).

This algorithm runs into the same issue that we saw with bisection methods: when to stop. Newman and Girvan [484] propose a method for determining when to stop this (or another) algorithm. For any community structure Π, one can calculate the following measure that Newman and Girvan call *modularity*. For two communities $\pi \in \Pi$ and $\pi' \in \Pi$ let $e_{\pi\pi'}(g)$ denote the fraction of all edges in the network that connect nodes in π to nodes in π'. Then the modularity of the community structure Π is

$$M(\Pi, g) = \sum_{\pi \in \Pi} e_{\pi\pi}(g) - \sum_{\pi \in \Pi, \pi' \in \Pi \pi'' \in \Pi} e_{\pi\pi'}(g) e_{\pi'\pi''}(g).$$

Modularity measures the proportion of edges that lie within communities minus the expected value of the same quantity in a graph such that all nodes have the same degrees but links are generated uniformly at random (ignoring community structure). When this measure is 0, then the communities are not capturing much of anything. If the measure is positive, then the communities are capturing more of a fraction of links internally than one would expect at random. When the measure is negative, then the community structure is cutting against the link pattern in that there are more links

across communities and fewer within than one would see at random. So, the idea is to maximize this modularity measure, and a rule for stopping an algorithm is that one stops if the modularity decreases by further edge removal. One has to be careful, as there can be local maximizers that are not global maximizers, and so one technique is to continue operating the algorithm in question to exhaustion, and then to go back and pick the community structure which leads to the highest modularity measure.

Hierarchical Clustering

A approach that differs from CONCOR, repeated bisection techniques, and edge removal builds up communities by adding new nodes to groups successively by examining how similar pairs of nodes are, rather than starting from one large community and breaking it apart. This methodology underlies a whole class of algorithms, called "hierarchical clustering" methods.

The ideas are as follows. The foundation is to have some measure of how similar two nodes are based on a given network. There are many such measures. One could use the correlation coefficients between the rows of the adjacency matrix, $C^{(1)}$, which was the first step of the CONCOR process above. Instead, we could calculate the distance between the vectors g_i and g_j (using Euclidean distance; or city-block distance - the number of entries that differ across the two vectors). We could also use various measures built on path distances or other indications of how the roles of two nodes within the network compare. The important point is that there is some measure that the analyst believes captures the appropriate notion of how similar two nodes are in the given application.

For the purpose of illustration, let us measure similarity between nodes i and j, via a slight variation[19] on the city block distance between rows, or in particular set the distance between nodes i and j to be

$$b_{ij}(g) = \#\{k | g_{ik} \neq g_{jk}, k \neq i, k \neq j\}.$$

Thus, if $b_{ij}(g) = 0$, then i and j are structurally equivalent. More generally, $b_{ij}(g)$ indicates the number of other nodes k with whom i and j differ in their relationships. So it directly counts the differences between the neighborhoods of nodes i and j.

In order to use the distance information to construct a community structure, let us start with a threshold of 0. We form a graph (that is purely for algorithmic purposes

[19] The variation is that we do not examine the cases where $k = i$ or $k = j$, as those are not relevant for our purposes.

and might look quite different from the original g), denoted $g^{(0)}$, as follows. Link any two nodes together that have $b_{ij}(g) = 0$. Thus, a link between two nodes indicates that we think they are "similar" in that they are at a low distance from each other. If we stop here, then we end up with a community structure that is the partition induced by the components of this network, $g^{(0)}$.[20] As there may not be many pairs of nodes that are structurally equivalent, $g^(0)$ will tend to have very few links, and so we usually end up with a very sparse community structure with many small communities if we stop with a threshold of 0. This suggests raising the threshold for how distant nodes can be from each other and be linked - or considered "similar". If we set a threshold t, then we end up with a graph

$$g^{(t)} = \{ij | b_{ij}(g) \leq t\}.$$

As we raise the threshold, the components of the induced graph continue to grow as more edges are added. The community structure at some threshold t, denoted $\Pi^{(t)}$, is the partition of nodes induced by the components of the network $g^{(t)}$ at some threshold t, so it is $\Pi(N, g^(t))$ (recalling notation from Section **??**). Eventually, as we raise t to $n-1$, we end up with a completely connected graph and just one community.

To get a feeling for how this works, let us apply it to the network in Figure 13.2.1. First, note that nodes 2 and 3 are structurally equivalent, as are nodes 7 and 9, so that $b_{23}(g) = 0 = b_{79}$. As these are the only such pairs, if we set the threshold to be 0 then the resulting community structure is

$$\Pi^{(0)} = \{\{1\}, \{2,3\}, \{4\}, \{5\}, \{6\}, \{7,9\}, \{8\}, \{10\}\}.$$

If we raise the threshold to 1, then we find new pairs of nodes that are "similar" to each other: nodes 7 and 8 only differ by one relationship, and in fact $1 = b_{78}(g) = b_{89}(g) = b_{12}(g) = b_{13}(g)$. We end up with

$$\Pi^{(1)} = \{\{1,2,3\}, \{4\}, \{5\}, \{6\}, \{7,8,9\}, \{10\}\}.$$

Next, given that $b_{3,10}(g) = 2$, $b_{4,10}(g) = 2$, $b_{5,10}(g) = 2$, $b_{8,10}(g) = 2$, $b_{6,7}(g) = 2$, we end up with just a single component under $g^{(2)}$ and so the resulting community structure at a threshold of 2 is already

$$\Pi^{(2)} = \{\{1,2,3,4,5,6,7,8,9,10\}\}.$$

[20] Note that structural equivalence is not a transitive relationship. That is, it is possible to have $b_{ij}(g) = 0 = b_{jk}(g)$, while $b_{ik}(g) = 1$. This happens when $g_{ij} \neq g_{kj}$. Thus, it is possible to have components of $g^{(0)}$ that are not cliques.

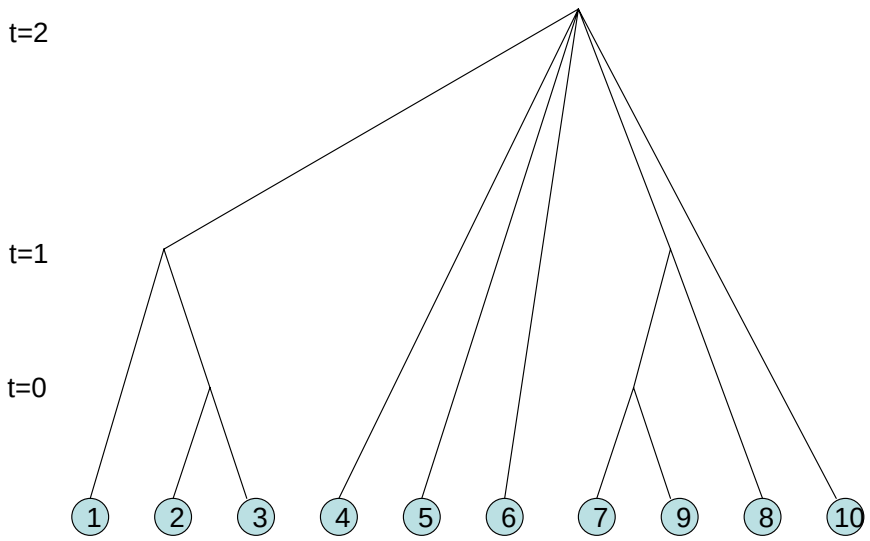

Figure 13.2.2. *The dendrogram or hierarchical tree generated by running a hierarchical clustering algorithm based on a simple (dis)similarity measure on the example from Figure 13.2.1.*

The term hierarchical clustering refers to the fact that as we raise the distance threshold (or lower the similarity threshold) for when we consider two nodes to be similar, groups of nodes continue to merge and we end up with a sort of hierarchy. The hierarchical tree, known as a *dendrogram*, generated for this example is pictured in Figure 13.2.2.

There are several challenges with such methods. First, we need to know when it is reasonable to stop the process. This can often be more of an art than a science, and so researchers will report the full hierarchy indicating at which thresholds different components coalesce. This report is often in the form of a hierarchical tree like the one pictured in Figure 13.2.2. One can also use modularity, or some other method, to choose among potential community structures. Second, the way in which communities coalesce is a bit questionable. For example, once we set the threshold distance to be 2 in the above example, the similarities between node 10 and several other nodes is largely responsible for bringing all of the nodes together into one community. However, "similarity" in measured in this way is not necessarily transitive. For example, nodes in the group $\{1, 2, 3\}$ are now in the same community with the nodes in the group

$\{7, 8, 9\}$ because they are each sufficienty "similar" to node 10, and yet any node from the first group is at a distance of at least 4 from any node in the second group, and in some cases are at a distance of 6 (e.g, $b_{1,7}(g) = 6$). This is generally a challenge with such methods, as a single node can cause many nodes to be grouped together that are not very similar to each other.

13.2.3 Stochastic Block Models and Communities

As mentioned at the outset of this section, a central difficulty behind all of the algorithmic methods discussed above is that we are not sure what "communities" they are designed to uncover. If we iteratively remove edges from a network, what are we uncovering and why is that a sensible way to proceed? What sorts of communities do we end up with when we examine a similarity measure and identify communities via hierarchical clustering? The problem is that the methods described above are defined by the algorithms and not by having a theory or model of what a community structure is. Communities simply happen to be what we end up with. This is not to say that such methods cannot be built up from some foundation, but rather that such foundations do not yet exist.

A different approach is to start with an explicit idea of what a community is and how networks are generated as a function of the underlying community structure. Then we can work from the foundation: given the network structure, we can try to deduce which community structure is most likely to be present, as we know the likelihood with which different community structures lead to various networks. This is a standard statistical approach to the problem - one presumes a model of how data (here a network of relationships) are generated from underlying parameters (here a community structure) and then one examines the data to statistically infer the parameters of the model. When applied to social networks this is sometimes referred to as *a posteriori block modeling*.

A natural version of this is a variation of a model of Holland, Laskey, and Leinhardt [316], as analyzed by Snijders and Nowicki [574].[21] Each node belongs to a group, which can be thought of as a block or community. To stick with the definitions above, let us call it a community and work with community structures, so that each node belongs to exactly one community.

[21] This is also related to other models, such as those of Holland and Leinhardt [**?**] and Fienberg and Wasserman [231]. It has experienced a recent resurgence (e.g., see Newman and Leicht [**?**]) in a portion of the literature unaware of its origin in the older literature.

The model is that the probability of a link between two nodes depends on which communities the given nodes lie in. The probability, for instance, could be higher within a community than across communities. There might also be some pattern of link probabilities depending on specific communities. In particular, the general form of the model is to have the probability of a link between a node in community π and a node in community π' be designated by a parameter $\eta_{\pi\pi'}$. The formation of each link is independent of the formation of every other link. The restriction of the model is that any two nodes in a given community are "equivalent" in the sense that the probability that either of them forms a link with any other node is the same. So, one can think of this as a form of probabilistic structural equivalence.

If we work with undirected networks, then $\eta_{\pi\pi'} = \eta_{\pi'\pi}$, while we can allow the probabilities to differ in a directed network.

A community structure with m communities, together with a list of the m^2 parameters $\eta_{\pi\pi'}$, leads to a well-defined probability that any given network will form. The probability that link ij forms is $\eta\pi_i\pi_j$ where π_i is the community to which i belongs. So, the probability (or "likelihood") of a network g is

$$L(g|\pi,\eta) = \left(\prod_{ij\in g} \eta_{\pi_i\pi_j}\right)\left(\prod_{ij\notin g}(1-\eta_{\pi_i\pi_j})\right). \tag{13.12}$$

One of the difficulties with the model in its fullest generality is that it allows for many parameters and might not really tie much down. For example, by setting the community structure to have each node in its own community, and then setting link probabilities across communities to be 1 when a link exists and 0 otherwise, we would end up predicting that our observed network should have been the one that formed. There are so many free parameters that we can explain any possible network exactly. So, in order for this approach to be meaningful, we need to place additional restrictions on the parameters of the model. These will generally be governed by the specific application and thus some additional information of what communities represent.

A special case of this (studied by Copic, Jackson, and Kirman [164][22]) is one where there is one probability for links within a community and another probability for links across communities. That is, there are just two probabilities, $1 \geq p_{in} > p_{out} \geq 0$, such that $\eta_{\pi\pi} = p_{in}$ for any $\pi \in \Pi$, and $\eta_{\pi\pi'} = p_{out}$ when $\pi \neq \pi'$. So, here communities are

[22] They consider a more general variant which is special in the η's but also allow for multiple links and varying capacities across links.

groups of nodes that are more likely to interact with each other, and interaction across communities is less likely.

To see how this model can be used to uncover a community structure given network data, let us explore this two-probability case in more detail. As mentioned above, specifying Π, p_{in}, and p_{out} leads to a well-defined probability of observing any particular network g for the unweighted and undirected version of the model (and for a weighted and directed version, see Exercise [**?**]).

Given a community structure Π, let $In(\Pi)$ be the set of all pairs of nodes that lie within the same community under Π and $Out(\Pi)$ is the set of all pairs of nodes that are in different communities under Π. That is,

$$In(\Pi) = \{ij \mid \text{ there exists } \pi \in \Pi \text{ such that } \{i, j\} \subset \pi\},$$

and $Out(\Pi)$ is the set of all pairs of nodes that are not in $In(\Pi)$. Let

$$T_{in}(g, \Pi) = |g \cap In(\Pi)|$$

be the number of links that are in the network g and that lie within communities under Π. Similarly, let $T_{out}(g, \Pi) = |g \cap Out(\Pi)|$ be the number of links that are in the network g and that lie across communities under Π.

The probability of observing network g if Π is the community structure is described by the likelihood $L(g|\Pi, p_{in}, p_{out})$, where

$$L(g|\Pi, p_{in}, p_{out}) = p_{in}^{T_{in}(g,\Pi)} (1 - p_{in})^{|In(\Pi)| - T_{in}(g,\Pi)} p_{out}^{T_{out}(g,\Pi)} (1 - p_{out})^{|Out(\Pi)| - T_{out}(g,\Pi)}. \tag{13.13}$$

Thus, given a community structure Π, we have a well-defined probability of each possible network.

13.2.4 Maximum-Likelihood Estimation of Communities

Given the probabilities of seeing each possible network as a function of the community structure, we can then invert the problem and ask which community structure leads to a highest likelihood of generating the network that we have actually observed. Maximizing this likelihood is equivalent to maximizing the log of this likelihood, which is often easier to work with. By taking the log of the likelihood function, $\ell(g|\Pi, p_{in}, p_{out}) = log\,(L(g|\Pi, p_{in}, p_{out}))$, we end up with a very manageable expression for the relative likelihood of observing a particular network g if the community

structure is Π:

$$\ell(g|\Pi, p_{in}, p_{out}) = \log(L(g|\Pi, p_{in}, p_{out})) = k_1|In(\Pi)|+k_2T_{in}(g,\Pi)+k_3|Out(\Pi)|+k_4T_{out}(g,\Pi), \tag{13.14}$$

where the k's depend only on p_{in} and p_{out}.[23]

Noting that $|Out(\Pi)| = \frac{n(n-1)}{2} - |In(\Pi)|$ and $T_{out}(\Pi) = |g| - T_{in}(\Pi)$, we can rewrite (13.14) as

$$\ell(g|\Pi, p_{in}, p_{out}) = (k_1 - k_3)\,|In(\Pi)| + (k_2 - k_4)\,T_{in}(g,\Pi) + r, \tag{13.15}$$

for a r that depends only on $|g|$, p_{in} and p_{out}, and not on Π.

So, the community structure that is identified via this method is the Π that maximizes $(k_1 - k_3)\,|In(\Pi)| + (k_2 - k_4)\,T_{in}(g,\Pi)$. We have boiled down the problem of finding a community structure to a calculation that just involves examining a weighted difference of the number of pairs of nodes lie within the same communities and the number of links lie within communities. Noting that $k_1 - k_3 < 0$ and $k_2 - k_4 > 0$, as we group more nodes together we see two competing effects in terms of how the likelihood changes: the second expression in (13.15) increases, while the first one decreases. The relative rates at which this occurs depends on the *relative* density of how many links are within a community compared to how many pairs of nodes lie within communities.

Algorithms for Maximum Likelihood Estimation

There are two challenges to implementing this method. One is that p_{in} and p_{out} need to be estimated along with the community structure, as they determine the k parameters. Dealing with this is relatively straightforward and involves an iterative procedure: one begins with an initial estimate of these parameters, which then leads to a initial estimate of a community structure. Next, based on a the first estimation of the community structure one can estimate p_{in} and p_{out} directly by examining the fraction of links within communities compared to the potential number, and similarly for across-community links. One can then iterate on this process.[24]

The second challenge is a bit more difficult to deal with. It is that the number of potential community structures grows exponentially in the number of nodes. With

[23] In particular, $k_1 = \log(1 - p_{in})$, $k_2 = \log\left(\frac{p_{in}}{1-p_{in}}\right)$, $k_3 = \log(1 - p_{out})$, and $k_4 = \log\left(\frac{p_{out}}{1-p_{out}}\right)$.

[24] As Copic, Jackson and Kirman [164] show, as the number of potential relationships is increased, there is a unique consistent pair of a community structure Π and p_{in} and p_{out} which lead to estimates of each other.

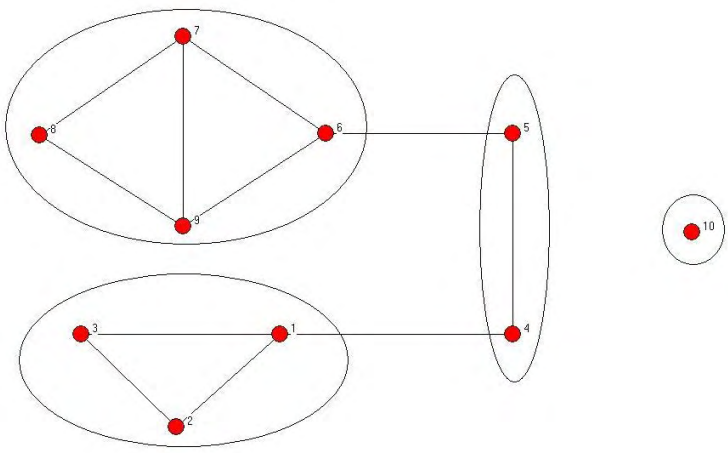

Figure 13.2.4. *The community structure found via the Maximum Likelihood algorithm on the network in Figure 13.2.1.*

even moderate numbers of nodes, this makes it impossible to calculate the likelihoods of the given network for all possible partitions. With a small number of nodes, one can do the calculations directly, as for the network in Figure 13.2.1, where one ends up with the community structure pictured in Figure 13.2.4.

However, for larger numbers of nodes, one has to employ some sort of approximation technique. Various techniques can be found in Snijders and Nowicki [574], Copic, Jackson, and Kirman [164], and Newman and Leicht [**?**].

13.2.5 Latent Space Estimation

The model of community structures in Section 13.2.3 is relatively simple in that it simply posits that link probabilities are completely governed by community membership. In some settings, there are more complex relationships that underlie the relationships between nodes. There might be many attributes, including socio-economic, geographic, and a variety of status attributes like profession, religion, gender, race, membership in organizations, and so forth that influence the relationships. There might also be hierarchies of nodes according to various measures, and these all might come together

in different ways to influence the chance that two nodes are linked to each other. A community structure is an extreme model of this. More generally, one can model richer underlying structures that help determine relationships.

A straightforward generalization of the maximum likelihood approach outlined above for community structures is to posit some model of structures and attributes and how a given network is to emerge under various parameters of the model. Take some general structure S to be the primitive, which might include all sorts of information about nodes and layers of groupings. Each S then leads to a likelihood of observing a given network g, denoted $L(g|S)$. Given that we observe g, we can look across potential S's to find the one which maximizes the likelihood of having seen g – that is the S that maximizes $L(g|S)$.[25]

An example of this, beyond the basic community membership model, is known as "latent space estimation." The idea is that nodes are located in a space and the probability that they are linked is dependent on their spatial locations (e.g., see Hoff, Raftery, and Handcock [**?**] and Hoff [312]), generally with the probability of a link increasing as nodes are closer together.[26] The space can take many forms, but provides an explicit model of how networks emerge and what we wish to uncover. The critical element in such estimation is to ensure that the model has some limits in terms of the number of parameters to be estimated, so that one does "overfit" the data.

13.3 Exercises

EXERCISE **13.1** *Identification with Contextual Effects.*

Consider the following generalization from Manski [424] of the model in (13.6).

$$Y_i = a + b_1 E[Y_i|x_i] + b_2 x_i + b_3 z_i + b_4 E[z_i|x_i] + \varepsilon_i. \tag{13.16}$$

Here z_i is a contextual effect, and i's expectation of z_i can matter too. Show that the parameters a, b_1, b_2, and b_4 are not identified.

[25] One can alternatively do a Bayesian analysis, where one has a prior probability distribution, P, over possible S's, and then given the likelihoods $L(g|S)$, one applies Bayes' rule to derive a posterior probability that S is actually in place. One chooses S to maximize $\frac{L(g|S)P(S)}{\sum_{S'} L(g|S')P(S')}$. This is equivalent to maximizing $L(g|S)P(S)$ and is the same as maximum likelihood estimation if we set the prior, $P(S)$, to be equal across structures.

[26] One can think of the community structure model as a special case where the space is a hypercube with n vertices, all nodes in the same community are located at the same vertex.

EXERCISE **13.2** *Identification Problems with Nonlinear Models.**

Consider the following two alternative models of behavior.

The first model is as follows. In each period a parent is replaced by its child, who then becomes a parent in the next period. The child makes a 0 or 1 decision (e.g., to attend university) and dependent upon the parent's choice. If the parent took decision 1, then the child takes decision 1 with probability q_1, while if the parent took decision 0, then the child takes decision 1 with probability q_0, where $1 \geq q_1 \geq q_0 \geq 0$.

In the second model, there are two such "families". In each period, one of the two families is selected by the toss of a fair coin. That family (and only that family) has its member ("the parent") die and be replaced by a child. In this model, the child then makes a decision. The child looks to the other family (its neighbor), and if that neighbor has taken decision 1, then the child takes decision 1 with probability p_1, while if the neighbor took decision 0, then the child takes decision 1 with probability p_0, where $1 \geq p_1 \geq p_0 \geq 0$.

Show the following result from Calvo-Armengol and Jackson [**?**]. For any such specification of $1 \geq q_1 \geq q_0 \geq 0$ in the first model, there is a specification of $1 \geq p_1 \geq p_0 \geq 0$ in the second model that leads to an observed probability of a child taking action 1 conditional on the parent's choice which is exactly the same as that of the first model. Do this by calculating these conditional probabilities for any given p_1 and p_0, and show that the range is the set of (q_1, q_0) such that $1 \geq q_1 \geq q_0 \geq 0$.

EXERCISE **13.3** *Applying the Girvan-Newman Algorithm to the Network in Figure 13.2.1.*

Which link has the highest betweenness score at the first step in the Girvan-Newman algorithm applied to the network in Figure 13.2.1 and what is that score? Indicate the next two edges to be removed. Which edges would be removed next if we continued?

EXERCISE **13.4** *Multiple Community Memberships.*

Consider the following variation on a model of community structures. There are a set of "clubs" that agents can be members of.[27] In the special case where each agent must be in one club and only one club, then this reduces to a community structure (ignoring any empty clubs), but more generally we might allow agents to belong to

[27] One can also think of these as activities that an agent can undertake, such as a sport; or as an attribute that an agent might have, such as their ethnicity, profession, etc.

more than one club, or even no clubs. An agent's likelihood to be linked to another agent depends on the number of clubs they are in together. The probability that two agents are linked is p_k, where k is the number of clubs that they are both members of and p_k is increasing in k.

Describe a likelihood method for recovering club membership for a fixed number of clubs.

EXERCISE **13.5** *Hierarchies and Communities.*

Consider augmenting a community structure by a hierarchy. That is, starting with a community structure Π, let h be a "hierarchy" function such that $h(\pi) \in \{1, 2, \ldots, K\}$ indicates which level of the hierarchy pi lies in. Let the probability that a node links to another node in the same community be p_{in} and the probability that a node in a community in level k links to a node in a different community in level k' be $p_{kk'}$.

Describe a likelihood method for recovering the community structure and hierarchy function given a presumption that there are at most K levels to the hierarchy and fixing some starting estimates of the $p_{kk'}$'s.

Allowing for a directed network, describe a method for recovering the community structure, hierarchy function, and the $p_{kk'}$'s; under a constraint that $p_{kk'} > p_{k'k}$ when $k > k'$.