4

Network Structure and Social Outcomes

*Network Analysis for Social Science**

Betsy Sinclair†
Washington University in St Louis

1 WHY SHOULD SOCIAL SCIENTISTS STUDY NETWORKS?

Human behavior is characterized by connections to others. We define ourselves by these connections: our families, our friends, our neighbors, our co-workers all form a social geography. Social scientists who study networks serve as cartographers for these social plains, identifying actors who influence others. In their overview of the study of political networks, McClurg and Young (2011) state, "We would probably all agree that one primary tie among political scientists is our emphasis on power, and understanding how and why power is used. We are all inherently interested in the exercise of power between and among individuals and groups and the implications that this exercise holds for social outcomes. We contend that this unifying concept is, at its very core, relational." Social scientists have an interest in relational social science, with roles as either researchers directly focusing on relationships between actors or else as scholars accounting for interdependence among actors and institutions in their analyses.

Additionally, we have seen an explosion in the availability of networked data. With the rise of social media, the relationships between ordinary citizens and political elites, among ordinary citizens, and even among political elites is more easily quantified. When once scholars of Congress had to "soak and poke" to understand a legislator's relationship with her constituents (Fenno 1978), now it is possible to directly observe the connections that legislators establish with their constituents over Twitter, as well as the connections between the constituents themselves (Barbera 2015), the donations made to legislators and

* I thank participants from the University of Iowa and University of Southern California Master Class sessions on political networks for their comments and suggestions on this material.
† Department of Political Science, 1 Brookings Drive, St Louis, MO 63130; bsinclai@wustl.edu.

from one legislator to another (Bonica 2014), and the relationships between friends that lead one friend to influence another to cast a ballot (Bond et al. 2012). The availability of these new data resources demands new empirical techniques in part because of these resources' sheer magnitude and in part because the relational structure of the data allows different analyses than ever before. This chapter focuses not on the computational innovation but rather on the empirical strategies.

A range of scholars from political science to sociology to economics to even physics have innovated to develop empirical techniques to account for networked relationships.[1] Interestingly these methods are applied to a range of substantive questions. Existing scholarship has documented that any two Americans are connected by six intermediaries (Milgram 1967), for example, and that any two unrelated Web pages are separated by only 19 links (Albert, Jeong, and Barabasi 1999). For another example, all of the 19 hijackers responsible for the 9/11 disaster could be tied together using shared and available data (such as addresses, telephone numbers, and frequent-flier numbers) and a disproportionate number of network metrics converged on the leader Mohamad Atta (Krebs 2001). Across substantive questions it is consistently clear that citizens are connected to other citizens and that our political and economic institutions are connected to each other. The structure of those connections has the potential to yield meaningful insight into the kinds of outcomes that are possible or expected. The goal of this chapter is to introduce the reader to a range of tools of network analysis that provide insights into a variety of social science problems. First the chapter focuses on those tools best designed for exploratory data analyses – visualizing a network and producing simple summary statistics of the network's characteristics. It then turns to a range of effective strategies to draw causal inferences.

Visualization of networks can provide answers – and quick ones – to all kinds of questions. If you want to be a law school professor, where should you get your JD? By simply visualizing the network of "sending" and "receiving" schools – by drawing a network where two universities are linked if one hired a JD who graduated from another – it is pretty clear that the best bet is to have gone to Yale or Harvard. A quick comparison to a similar network for political science faculty demonstrates that this network is much more diverse.

Quantification of networks can similarly provide surprising answers. Why did the Medici rise to power in Renaissance Florence? They were neither the most powerful nor wealthiest family. By visualizing the Medici family's marriages with other powerful families, it is possible to see how they are central to the network of Florentine families. Quantifying the strength of the Medici position in the network and noting that this family dominates all standard network quantification metrics in terms of the marriage network suggest that it is not that surprising they were the most powerful family at the end of the Renaissance in Florence. Other quantification tools can also provide surprising insights (such as those related to community detection).

Neither of these tools (visualization and quantification), however, allows us to test hypotheses, although they are useful for exploratory data analyses. Randomized field experiments have made great forays into providing designs for explicitly testing the role of network ties in influencing outcomes of interest for social scientists. These experimental designs have been pushed into the field because of serious concerns about the endogeneity between the outcome of interest and the variables that generate a network tie. Much of the observational work in this area has suffered from this criticism (and in fact some current statistical work addresses this problem). Even these experimental designs, however, are subject to a set of assumptions about the data-generating process. Consequently, a separate line of researchers have suggested that the best tools that social scientists can offer to test networks are those that explicitly test models. The final section of this chapter explores the tension between testing specific models and designing randomized trials.

Social scientists are increasingly interested in relationships. Whether those relationships are between countries, between individuals, or between institutions, the field of social networks has offered a number of useful tools to gain insight into how the structure of these relationships influences outcomes of interest. This chapter aims to advance the appropriate use of these tools in the social sciences.

2 WHAT IS AND IS NOT A NETWORK?

Before proceeding with a discussion of the vast array of empirical tools available for social scientists, it is important to first define what is and is not a network. The field of network analysis has been quite generous, so a network is typically defined as a set of actors and their relationships where the range of potential actors and relationships is defined quite broadly. Nodes typically represent actors or institutions, whereas edges represent connections between such entities. Edges can be either directed, to represent connections that flow from one node to another, or undirected. They can simply signify the existence of a connection, taking on a discrete value of 0 or 1, or they can be weighted to reflect the strength of the connection between two nodes.

Key to this definition of a network is that relationships describe how the actors are connected. These networks can be cities connected by interstate highways, scholars connected by collaborations (Berardo 2009), or even bills introduced in the House of Representatives that are connected by their committee of reference (Gailmard and Patty 2013). It is the responsibility of the researcher to posit a network where the relationships are something meaningful. What is not a network? An instance of data that is not a network is one where the edges have no social or relational meaning – suppose, for example, you drew a network based on numbers from a phone book that had the same last four digits. While these actors (the phone numbers) do have something in common (the last four digits), the edges have no role in defining a relationship

between them.[2] Yet edges (relationships) can represent a wide class of potential connections, including social ties, the flow of goods between firms of countries, or the linkages between various actors or institutions. The type of network under study is frequently defined by the relationships that connect the actors. For example, a *social network* is a social structure made up of a set of social actors (such as individuals or organizations) and of a complex set of the dyadic ties between these actors (Wasserman and Faust 1994), whereas a *political network* consists of the social network structure that focuses on politics, elections, or government (Sinclair 2012).

After a network is defined, it is then possible to conduct an analysis. This next section reviews several common tools used to quantify characteristics of a network. Although these are summary statistics and do not easily lend themselves to statistical tests, they do often provide clear insights into the association between the structure of a network and its outcomes, suggesting the possibility that it is through networked influence that these outcomes occur.

3 NETWORK ESTIMATION

What is important to focus on when considering the importance of networks? Each person or institution is connected to a myriad of potential other actors. The first step in network estimation is simply to properly define the network. The definition, paired with a visualization of the network's structure, can sometimes yield great insights (for example, defining the academic hiring market for law and political science as a network between institutions that grant terminal degrees and those that hire faculty). Yet not every network can be easily visualized to capture the key insight. Quantification of the network characteristics, then, can either support the insight gained by visualization or more frequently helps provide evidence that the structure corresponds with a pattern of influence posited by the researcher. This section reviews some of the most common quantification tools using a handful of canonical examples in the network literature . Additionally, there are a range of more sophisticated tools, such as those designed to detect communities, that can be employed when studying social networks. These tools look for the presence of key structures and can handle very large networks where visualizing the structure of the network is simply not feasible.

3.1 Visualizing Structure

The hiring of graduates from law schools and graduate schools for academic positions is a good example of how simply visualizing a network is can yield insights into the structure of an institution, in this case the academic hiring market. Here, the actors are the academic institutions that grant the degrees and the academic institutions that might hire the graduates for professorial positions. Edges represent a hire and are directed, with the sending

institution granting the degree and the receiving institution hiring the graduate. It is possible to draw these networks for both law schools and for political science departments and then to compare whether these two different fields have very different hiring structures (Katz et al. 2011). Interestingly, they do. Law schools are primarily populated with individuals who received degrees from Yale or Harvard, whereas political science departments are populated with individuals who received their degrees from a much more diverse array of institutions. Merely by drawing the figure that characterizes this network we learn something about the hiring norms in these two respective fields. Networks have the power to yield insights when they are simply visualized.

During the 1400s, Florence, Italy, was ruled by an oligarchy of elite families. The Medici family rose to power in Florence, despite having less wealth and less political power than other ruling families in Florence at that point in time. A question that has long puzzled historians was how to understand the Medici rise to power. By simply visualizing the structure of the network it is easy to grasp the insight that Florentine family marriages allowed them to secure influence in the early 15th century (Padgett 1993). Indeed, Cosimo de Medici, "the godfather of the Renaissance," orchestrated strategic marriages to ensure a central position within the Florentine social network. This data is frequently used in the network community to illustrate the importance of network quantification tools.

To understand the Medici rise to power, it is possible to plot the marriage network and then capture the essence of how the strategic choice of marriages resulted in the Medici family becoming central figures in the Florentine social network. In this network, each node represents one of the elite families. Each link represents a marriage between members of two families. Figure 4.1 presents this network.[3] Again, by doing nothing other than plotting the network figure, it is possible to gain insight into the Florentine power structure.

In terms of the social-political structure of the Florentine families, the Strozzi had the most wealth at the start of the 15th century. Why did they not rise to power and instead the Medici did? Looking at this network it is clear that the Medici are better positioned. They look more central to the network. How much more central are they? The next section quantifies this characteristic.

3.2 Network Quantification

How can we quantify network position? What types of characteristics, for example, make the Medici more central to the network than the Strozzi? To illustrate the differences between the Strozzi and the Medici, for example, we can use two common network quantification tools. First, *degree centrality* counts the number of edges for each node. In this case, that means that we count the number of families to which a given family is linked through marriages. Second, *betweenness centrality* calculates the fraction of the total number of shortest paths between any two nodes on which a particular node

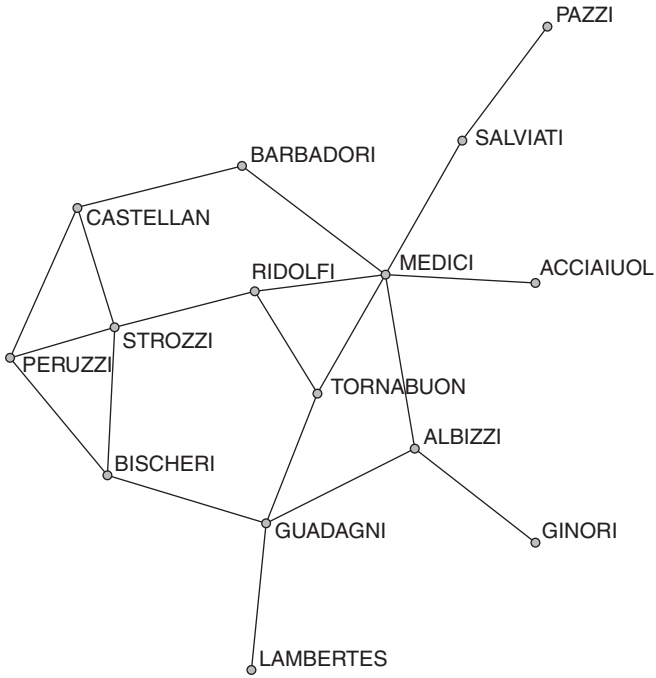

FIGURE 4.1. Florentine family marriages in the early 15th century.

lies. In this case, we consider paths to be marriages and then consider the fraction of the total number of shortest paths between any two families on which a particular family lies. Returning to the Florentine marriage network shown in Figure 4.1, it is easy to add up the number of ties to discover that the degree centrality of the Medici is 6, whereas the degree centrality of the Strozzi is 4. This is only one way in which the Medici are better positioned. If instead we were to calculate their respective betweenness centrality, the Medici have a betweenness centrality of .522, whereas the Strozzi have a betweenness centrality of .103. The intuition behind betweenness centrality is that if these edges were roads, and if you were committed to driving the shortest distance (where distance is calculated by the number of nodes you pass through) between any two nodes, you would want to know which actor was on the greatest fraction of shortest paths for any possible set of paths. For example, consider the shortest paths between the Barbadori and the Guadagni families. There are two such paths: Barbadori-Medici-Albizzi-Guadagni and Barbadori-Medici-Tornabuon-Guadagni. The Medici lie on both paths, and the Strozzi lie on neither.

Networks are frequently quantified using four types of values (1) degree centrality, which calculates how connected a node is to his neighbors; (2) closeness,

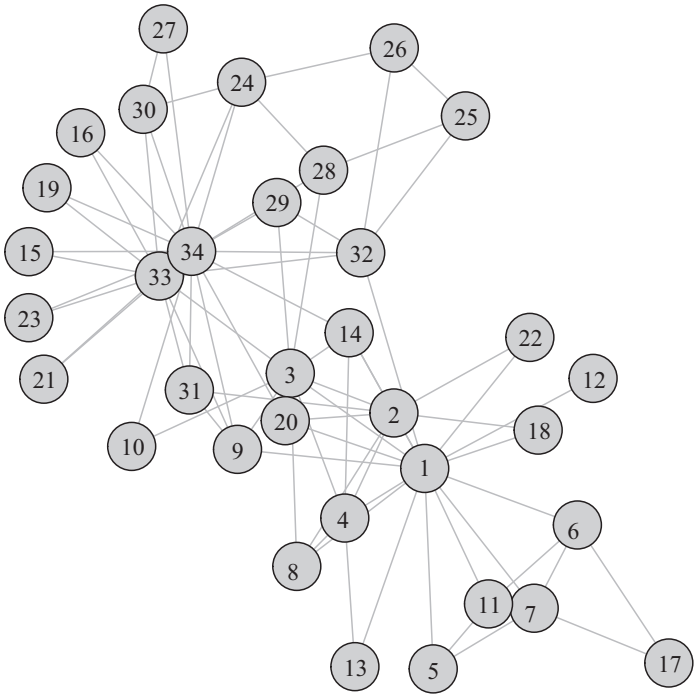

FIGURE 4.2. The karate club.

which calculates how easily a node can reach other nodes; (3) betweenness centrality, which calculates how important a node is in terms of connecting other nodes, and (4) eigenvector centrality, which estimates how important, central, or influential a node's neighbors are.

To illustrate all four of these tools, we turn to another canonical data set – the "karate club" data. A researcher documented observations from a karate club at a Midwestern university for three years, from 1970–1972 (Zachary 1977). The club's activities included both social activities (parties, dances, and banquets, for example) and regularly scheduled karate lessons. The club president and the part-time karate instructor disagreed over the price of karate lessons and more specifically under whose authority it fell to raise prices. Over time, the club became divided over this issue. After a series of confrontations, the supporters of the karate instructor resigned and formed a new organization. The structure of friendships within the club – that is, those individuals who consistently interacted outside of the karate lessons – provides a high degree of prediction into who would split into which group after the subsequent confrontation.[4]

To visualize this network, the nodes are the individual members, and the links are their social interactions. The network can then be seen in Figure 4.2.

TABLE 4.1. *Quantifying Network Position: The Karate Club*

Node	Degree	Closeness	Betweenness	Eigenvector
1	16	0.017	231.07	0.952
2	9	0.014	28.47	0.712
3	10	0.016	75.85	0.849
...	...	...	...	...
12	1	0.011	0.00	0.141

It is easily possible to calculate each of these four centrality quantifications for each node, which I do in Table 4.1. To illustrate their differences, focus on nodes 1 and 2 versus node 12. As is clear from the visualized Figure 4.2, nodes 1 and 2 are much more central to the karate network than node 12: they have a higher degree (16 and 9 versus 1), closeness (.017 and .014 versus .011), and eigenvector centrality (.952 and .712 versus .141). When focusing on their betweenness, however, it is clear that node 1 – which is quite critical in terms of connecting other parts of the network – has the most influence. Node 1's betweenness centrality is 231.07, whereas node 2's betweenness centrality is 28.47 (less even than node 3 with a betweenness centrality of 75.85). Node 12 has a betweenness centrality of 0.

Since the karate club network is small, it is easy to compare our intuitions with the quantified values in Table 4.1. These kinds of quantification tools are now frequently used across the social sciences to gain additional insight into how the structure of a network can relate to the influence of particular individual nodes. For example, in a pioneering study on legislative cosponsorship, Fowler (2006) focuses on the most and least connected legislators in the 108th U.S. House of Representatives, where legislators are nodes and are connected by their cosponsorship decisions. Looking at the 20 most central legislators, we see a preponderance of legislative leadership and other representatives reknowned for their legislative prowess. None of the 20 least central legislators are household names. We can sometimes explain changes in these centrality measures as well. In understanding which legislators are most central in the California State Assembly, for example, before and after the California blanket primary, it is clear that those legislators elected under a blanket primary system are more likely to be influential to both parties (Alvarez and Sinclair 2012). The quantification of social network measurements can then associate network characteristics with outcomes of interest.

3.3 Community Detection

Another example of the quantification of networks that can play a key role in linking the geography of a network to social science outcomes is that of community detection. In this case, we theorize that there exists a latent social

structure of a network that is often not fully observable and thus needs to be constructed from the set of observed ties between nodes. We formally define a *network community* as a subset of a social network graph that is more internally connected than externally connected. Community detection allows us to uncover the structures that split a network into communities. Uncovering these kinds of structures allows us to address questions such as the following. Are there specific biases in a society, such as in hiring or publishing? Are there systematic ways to classify and categorize political ideologies or economic patterns of behavior?

One main advantage of using community detection algorithms is that they do not presuppose knowledge of the number and size of the underlying groups. That is, it is possible to apply a community detection algorithm to a network and to find a result that indeed there are no specific communities within that network. Alternatively, the algorithm could unearth many separate communities. One limitation of these algorithms is that they typically can only find non-overlapping communities.

These algorithms have been used in a number of different applications that appear sensible and where the communities that are detected correspond with our understanding of communities that are present within that particular data. For example, Newman (2004) uses a community detection algorithm to detect the pages on a website using the hyperlinks between them, and Porter et al. (2005) use a related algorithm to detect congressional committees using roll-call votes. Some of these analyses require enormous computational power, such as those that analyze the purchasing decisions by Amazon customers with respect to books on American politics, political blogs, or coauthor networks (Girvan and Newman 2002).

In Figure 4.3 using the canonical "karate club" data, I rely on modularity maximization, formalized by Girvan and Newman (2002). There are many possible alternatives, but this algorithm seems particularly robust when handling large networks of data. I first define modularity. Let $\mathbf{e}$ be a symmetric $n \times n$ matrix where each element e_{ij} represents the fraction of all edges that link community i to vertices in community j. Modularity is defined, then, as

$$Q = \sum_i \left(e_{ii} - a_i^2\right) = \text{Tr } \mathbf{e} - \|\mathbf{e}^2\|,$$

where a_i is defined as the row sums, $a_i = \sum_j e_{ij}$. Tr $\mathbf{e}$, the trace of $\mathbf{e}$, gives the sum of edges connecting vertices in the same community, and $\|\mathbf{e}^2\|$ represents the sum of the elements of the matrix $\mathbf{e}^2$. We can think of $\|\mathbf{e}^2\|$ as measuring, holding the detected community structure fixed, the expected number of edges connecting communities if connections between vertices were random. Modularity, Q, can then take on values between 0 and 1. A value of 0 implies no more community edges than would be expected. To calculate modularity in this context, we first calculate betweenness (the number of shortest paths that run along

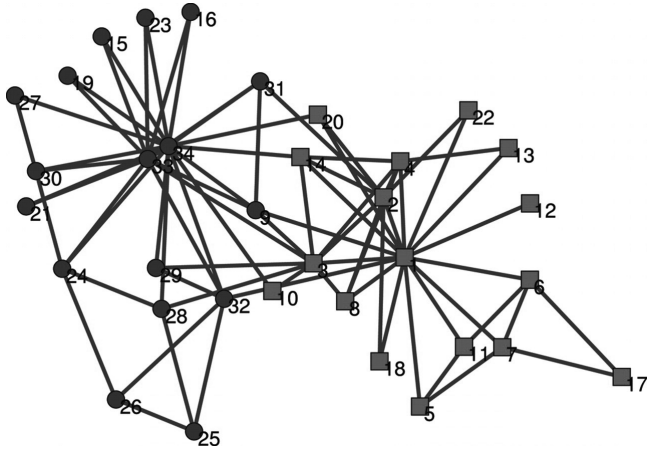

FIGURE 4.3. The karate club communities.

a given edge) for every edge. We then remove edge with the highest betweenness score because this edge is most likely to connect communities as opposed to lie within communities. We then estimate modularity, which measures the number of within-community edges relative to a null model of a random graph with the same degree distribution. After this phase, the betweenness scores are recalculated within the new graph and the process repeats, continuing to remove edges until modularity is maximized.

Here I apply modularity maximization to the karate club data, where edges represent shared social activities and nodes represent members. The algorithm detects two communities, which are shown in Figure 4.3.

Like the previous examples, these two communities that are discovered correspond to what we know about this data. Recall in the previous section we described how the karate club split into two separate factions after the disagreement over the price of karate lessons. This example illustrates that the topography for the split – that the people who interacted socially were divided into two membership groups – was already set before the disagreement took place.

Although in this case the community detection algorithm was successful, the field of network science has done little to develop serious hypotheses tests. We have no notion of a test statistic in the case of modularity maximization, but rather are simply drawing a comparison to a null model (a random network) without any formalized statistics. In the field of network analysis, there are many possible algorithms to detect communities within the observed networks.[5] Much of the literature takes an "inter-occular test" approach, looking for whether the communities that are detected fit some kind of "common sense" for what is "known" about the network ex ante. Thus the difficulty for these methods is that they are primarily designed to provide tools for exploratory

social science. They are interesting and allow us to develop hypotheses about social geography, but for networks to advance in the social sciences, we need to improve our ability to draw inferences about networks.

4 NETWORK INFERENCE

The key goals for a successful marriage between network science and social science are to generate two key empirical tools: first, a test statistic that allows a comparison to a null model to evaluate how likely we are to construct the network measurements at random, and second, a way to allow network traits to explain outcomes in a causal framework. Although little has been done to advance to the appropriate null model comparisons, because network measurements are frequently used as explanatory variables in regression-type model, then threats to inference come primarily from measurement error. This chapter focuses on causal inference.

Scholars have made key advances in terms of causal inference and networks. The question of causal inference is a fundamental one to all social science: we are interested in questions of cause and effect. In the field of network science, we frequently observe associations or correlations between network parameters (centrality, for example) and other outcomes, but these associations may not be causal. In particular the field has been plagued with criticisms regarding selection biases. Friends become friends because of shared interests or characteristics, for example, and these shared traits may lead to shared outcomes regardless of the friendship. In the following subsections, I discuss some of the strategies that can be employed to allow serious causal inference with networks.

4.1 Observational Approaches and Endogeneity

Several of the most well-known social science studies that involve networks are penned by Nicholas Christakis and James Fowler (2009). In these studies, they frequently test whether health outcomes spread through networks (obesity, smoking, loneliness, etc.). The key concern they have faced in their analyses is that people select their social networks (their families and friends) because of shared traits (proclivities toward particular health outcomes or geography, for example) that are causally associated with these health outcomes. This produces an omitted variable bias problem with the additional complication that the people in the network are not independent of each other.

For purpose of illustration I focus on the Christakis and Fowler study (2007) on the spread of obesity through a social network. Their data includes 12,067 people (5,125 "egos" and the remaining "alters") from the Framingham Heart Study. The data was collected from 1971 to 2003. Egos took part in physical exams that measured their weight in regular intervals over this 32-year time frame. Egos also identified social ties – they were asked to provide the names of

individuals who could help locate them for repeat health evaluations. Because of the geographic nature of the study, many of the social ties who were named (the "alters") were also in the study, and thus this data provides both longitudinal network data and longitudinal health data.

The study then explains obesity (at time $t + 1$) as a function of obesity (at time t), the alter's obesity (at time t and $t + 1$), the geographic distance between the alter and the ego, and finally a set of ego attributes (age, sex, and education). Its principal finding is that the alter's obesity is a significant predictor of the ego's obesity. The study gives three possible explanations for its findings: homophily (that the egos associate with like alters), confounding (that the egos and alters share unobserved attributes or experiences), or social influence. By explicitly controlling for the alter's weight at time t, the study argues that homophily is not the most likely explanation. By explicitly controlling for geography, the study finds that decreasing geographic distance does not increase the probability of obesity, potentially ruling out confounding (moreover, because not every ego-alter pair names each other as friends, the effect is only observed in the direction of the named friendship). Thus, the study concludes, the most likely explanation is social influence.

Although this study has been criticized by a number of scholars, the most compelling critique has come from Shalizi and Thomas (2011). They argue that unfortunately the limitations in the statistical modeling of observational network data in general result in very limited capabilities to draw causal inferences without exceptionally strong assumptions. In particular they argue that additional explanations cannot be ruled out in the Christakis and Fowler approach. They motivate their work asking, "If your friend Joey jumped off a bridge, would you jump too?" and then go on to say that if the answer is "yes," it could be because (a) Joey inspires you (social influence), (b) your friendship is founded on a shared love of jumping off bridges (homophily), or (c) your friendship is founded on a common risk-seeking propensity (confounding). Unless the Christakis and Fowler study could eliminate the potential variables in category (c), for example, then it would be impossible to distinguish confounding from social influence. In general, these three explanations (social influence, homophily, and confounding) are (almost) impossible to distinguish in observational network studies. How, then, are the social sciences to advance knowledge in this area?

Fortunately there have been a handful of recent advances in the realm of sensitivity analysis that allow researchers to evaluate the potential role of unmeasured confounders in changing the perceived empirical results of their work when they are handling network data. These techniques, long employed in the service of observational data, were some of the first to demonstrate the causal link between smoking and lung cancer, for example. In particular, VanderWeele (2011) proposed a sensitivity analysis technique to assess the extent to which an unmeasured factor responsible for homophily or confounding would have to be related to both the ego's and the alter's state to

substantially alter qualitative and quantitative conclusions. This work was extended by VanderWeele, Ogburn, and Tchetgen (2012) and indicates that there are reasonable conditions to use the kinds of models that Fowler and Christakis advance. Substantively, it appears that the Fowler and Christakis results are fairly robust to the skepticism about selection biases. In general network tools can be employed while accounting for this kind of endogeneity. Sensitivity analyses are also possible when network characteristics are used as explanatory variables in regression contexts.

Although the substantive findings of the obesity study emerge basically unscathed from the critiques of selection bias, there remained a significant push to study the effects of networks experimentally. I explore this possibility in the next subsection.

4.2 Experimental Approaches

A common suggestion when observational data is faced with some threat of selection bias is for the researcher to conduct a randomized field experiment. This is incredibly complicated in the context of social networks because most of the experimental subjects are understood to be socially connected to others. As a consequence, if the experimenter wishes to randomly assign some of the population to receive a treatment and the rest of the population to be considered a control condition, it becomes difficult to isolate a control group that is unconnected to a treatment group. Two primary experimental options exist. First, it is possible to randomly assign some members of the network to receive a stimulus and see how that percolates through the network. In this case, the researcher needs to carefully consider who is randomly assigned to the control condition and in particular to ensure that those assigned to the control condition are socially isolated from those assigned to treatment. Second, it is possible to randomly assign a new network structure – to assign new social ties, for example. This kind of study is best illustrated by the "dorm room" experiments, where undergraduate freshmen are randomly assigned to roommates. Yet, this type of experiment is complicated by homophily: if the subjects fail to comply with the new network structure – that is, to make new friends who are very different from themselves – then it is impossible to evaluate whether or not there are network effects. This is the paradox of homophily – it is the presence of homophily that generates such concerns about observational studies, but homophily also makes it very difficult to randomly assign a new network structure. The fundamental component of any experiment is to find a way to stimulate something truly exogenous into the network.

Field experimentalists have long been concerned about spillover from within social networks. Randomized field experiments frequently take place in concentrated geographic regions. Many subjects in those experiments are then exposed to multiple other individuals in both the treatment and control groups, so if the treatment is something that is contagious – information, for example – subjects

can be exposed to the treatment indirectly. If this happens, then estimates of the direct effect of the experiment (the local average treatment effect) may be biased. Yet, subject to certain restrictions or assumptions, it is possible to estimate the direct effect without bias. These include making the heroic assumption of the stable unit treatment value (Holland 1986), designing the experiment to estimate and account for spillover within specified hierarchical groups (Sinclair et al. 2012; Hudges and Halloran 2008), or limiting the exposure possibilities (Aronow and Samii 2013). In this chapter we explore the second option in greater depth.

In a classic randomized experiment, individual subjects are assigned to treatment and control. The central problem with the classic experiment is indirect exposure: either a treated subject could communicate with an untreated subject (and thus indirectly treat someone in the control condition), or else a treated subject could communicate with another treated subject (and thus amplify the treatment effect). The solution to this kind of design is simply to have a control group that is sufficiently isolated from the treatment group via a multilevel randomization. For example, Sinclair et al. (2012) conducted a voter mobilization campaign where registered voters were randomly assigned to treatment and control conditions based on their households and zipcodes. In this framework, then, it is possible to compare the average outcome of three types of individuals: one who is treated but lives in a household and a zip code where no one else is treated, one who is not treated but lives in a household where someone else is treated and in a zip code where no one else is treated, and finally one who is not treated and lives in a household and zip code where no one else is treated. If we assume that the social network of transmission is based on geography and is limited by zip code, then we can legitimately compare these three individuals: by comparing the turnout decision of the treated individual to the person who is not treated (and has no one around him treated), we can estimate the direct effect of the experiment, and by comparing the turnout decision of the untreated individual who lives in a treated household to the person who is not treated (and has no one around her treated), we can then also estimate the indirect effect of network transmission. Other experimentalists have employed similar approaches to try to isolate individuals who are eligible to be both directly and indirectly treated based on a model of network transmission (Nickerson 2008; Bond et al. 2012). This approach requires some assumptions about the network structure, as well as the ability to know something about the network itself, but it also implicitly allows the researcher to test whether this is a good network model if the treatment is something that is known to have a modest transmission rate.

4.3 Modeling Approaches

Network analyses tend to be ones that impose serious limitations on what can be estimated. This is because network data quickly becomes "big data"

once you consider the potential number of relationships that may influence outcomes. For a study with N individuals, for example, they may be connected to 2^N others where those relationships create additional parameters to be estimated in any kind of networked model. This quickly increases the computational complexity of the conventional analytical models. In the field of networked data, there are two primary types of modeling approaches, and I discuss only one briefly in this subsection. First, spatial statistical models assume that links are more likely between nodes closer in latent space. These spatial models incorporate not only a latent trait but also include spatial parameters in the error term. These are fundamentally additional modeling approaches that draw causal inferences to respond to the kinds of critiques made about the Christakis and Fowler work (2007), but are particularly useful in more limited data frameworks. Second, exponential random graph models (ERGM) allow researchers to specifically model the network (Cranmer and Desmarais 2010). It is this later topic that I explore in more depth in this subsection, because it employs a slightly different approach to drawing different kinds of inferences.

Imagine that the outcome variables of interest are alliance-ties between county-pairs. The primary motivation for ERGM is to provide a model whereby we can explain those ties as a function of the characteristics of the network. In this sense, ERGM is very similar to regression where the network ties are the outcome variable, but it does not require assumptions of independence. ERGM, unlike regression, does not require the researcher to assume that each state decides its alliance portfolio independent of all other decisions made by all other states and moreover that decisions even within a state's alliance portfolio are independent – assumptions that seem heroic by most standards. Instead, ERGM models the probability of observing that network of alliance relationships over the other networks we could have observed. It does make the assumptions that there is no omitted variable bias and that the particular realization of the outcome variable is drawn from a multivariate distribution based on a particular structure of the network. A reasonable intuition is to think of ERGM as analogous to a logistic regression that accounts for the network dependencies properly. ERGM models are appropriate in a limited range of circumstances – for example, they will fail to converge if there are too many covariates in the model and moreover will converge only in the context of thin networks, where nodes are more frequent than edges. They are also not equipped to handle nonbinary edges, networks that change over time, or missing data. Finally, they fundamentally require a very strong assumption about the presence of the appropriate set of covariates: the researcher must include those covariates that are theorized to affect the formation of a network tie (but not excess covariates of the model that may fail to converge) or else the model is subject to omitted variable bias. That said, this is an exceptionally practical tool when the researcher wishes to explain the formation of a network based on the presence of covariates.

TABLE 4.2. *ERGM on Florentine Marriages*

	Estimate	Std. Error	*p*-value
Edges	−2.30	0.40	≤1e-04***
Abs Diff Wealth	0.015	0.006	0.0131*

To illustrate the capacity of an ERGM, let us return to the example of the Florentine marriages. Here suppose that we believed that the Florentine marriages were to be explained by both the degree centrality of a family (the total number of other families they were already connected to by marriage) and by a difference in relative wealth between families. That is, this would provide some evidence of strategic marriages – marriages that took place between families could be explained by a strategic incentive to "marry up" in terms of wealth or in terms of connection. Table 4.2 illustrates the coefficients from this estimation procedure.[6]

These results indicate that there is little to suggest that the network in general is formed as a consequence of the total number of edges for each family: the degree centrality is in fact a negative and statistically significant predictor for the formulation of the network. Yet the absolute difference in wealth is a weakly positive and statistically significant predictor. This would suggest that in fact the entire network is not based on the kinds of strategic marriages the Medici are theorized to have orchestrated – which in part explains why the Medici were so successful with their innovation.

5 CONCLUSION

What tools from network analysis are most useful to social scientists? Networks emerge as both areas of interest and empirical complications, and in data as small as the ruling families of Florence to the enormity of Facebook. Visualization of networks, quantification of networks, and community detection have both strengths and limitations in partnering with traditional social science hypothesis testing. These tools are best employed in conjunction with randomized field experiments and modeling approaches.

Some of the most standard tools that are employed to quantify networks illustrate what we can see when conceptualizing individuals as part of a network and subsequently quantifying some of those network's characteristics. By evaluating the centrality of the Medici in terms of marriages, for example, we can understand their rise to power during the Renaissance. By appropriately defining the network, these network parameters provide summary statistics that can be quite revealing about the distribution of power and influence within a network. As demonstrated with the example of the karate club, they can also be predictive of vulnerabilities in an organization.

Yet, although these network tools are interesting, they are only truly useful to social sciences when paired with some kind of statistical test. Network inference

is best conducted with either sensitivity analyses, experimental approaches, or direct modeling approaches (which have the disadvantage of requiring some very strong assumptions). These three areas are being actively developed by methodologists and statisticians.

Substantively we observe consistent evidence that networks do influence a panoply of outcomes. A growing body of research documents the influence of context on voter turnout – the extent to which citizens' experiences are affected by their experiences of race and poverty, for example (Hersh and Nall 2015), or the participation choices of their friends (Bond et al. 2012). With the rise of big data, we can now observe this kind of association, but it requires extensive computing: simply to illustrate what is needed, Hersh and Nall (2015) analyze 73 million geocoded registration records to determine the relationship between racial context and turnout and Bond et al. (2012) analyze 61 million Facebook users' participation choices. The most exciting frontier in this line of research is to pair many of these network tools with these large data sets to allow for new kinds of hypothesis testing. Community detection algorithms appear prime candidates for such a marriage.

Human beings are part of a social environment. We typically characterize people as atomistic actors, yet the most basic actions (health choices, the decision to turn out to vote) are decisions that are associated with each person's social network. Thus social networks lead to particular outcomes of interest. Studying networks allows us to deepen our understanding of individual behavior and attitudes and to formalize the relationships across political institutions. In the end, we are all connected to each other, and the tools of network analysis help us understand our susceptibilities to each other, as well as our capabilities to effect real change in our own environment.

Notes

1. Indeed there is rising interest in the field of political networks, resulting in a new American Political Science Association section founded in 2009. Yearly membership dues are $8.00.
2. This example is drawn from Seth Masket's lecture "But Is It a Network?" presented during the IGNITE talks at Boulder PolNet 2012. The lecture is available online: http://www.youtube.com/watch?v=hv6wfR5WU4I.
3. In terms of visualizing a network, it is easiest to use the igraph `R` package. It enables both static and dynamic network graphs, and input data can be either a two-column edge list or an n × n adjacency matrix. Plotting your data allows you to make sense of the network so long as it is not too big, particularly in the case where the network changes over time. A wonderful example of a dynamic network is one where re-tweets are plotted over time, available through http://youtu.be/XX9he5lkNSo.
4. Social scientists who study networks use this data so frequently that they have developed an award, called the "Karate Club Award," for the first person at a conference to use this data: http://networkkarate.tumblr.com/.

5. Much of what is analyzed in this chapter is conducted using the igraph package in R. There are very good alternative options as well, including UCINET (https://sites.google.com/site/ucinetsoftware/home), Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/), and Siena (http://www.stats.ox.ac.uk/~snijders/siena/). Data is also available from Mark Newman's Network Data (http://www-personal.umich.edu/~mejn/netdata/) and the Stanford Large Network Data set Collection (http://snap.stanford.edu/data/).
6. Practically speaking this is the result of merely 20 iterations.

References

Albert, Reka, Hawoong Jeong, and Albert-Laszlo Barabasi. (1999) "The Diameter of the World Wide Web." *Nature* 401: 130–135.

Alvarez, R. Michael and Betsy Sinclair. (2012) "Electoral Institutions and Legislative Behavior: The Effects of the Primary Processes." *Political Research Quarterly* 65(2).

Aronow, Peter M. and Cyrus Samii. (2013) "Estimating Average Causal Effects under Interference between Units." http://arxiv.org/abs/1305.6156.

Barbera, Pablo. (2015) "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1): 76–91.

Berardo, Ramiro. (2009) "Processing Complexity in Networks: A Study of Informal Collaboration and its Effect on Organizational Success." *Policy Studies Journal* 37(3): 521–539.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jamie E. Settle, and James H. Fowler. (2012) "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489: 295–298.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2): 367–387.

Christakis, Nicholas A. and James H. Fowler. (2007) "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine* 357: 370–379.

Christakis, Nicholas A. and James H. Fowler. (2009) *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little, Brown.

Cranmer, Skyler and Bruce A. Desmarais. (2010) "Inferential Network Analysis with Exponential Graph Models." *Political Analysis* 19: 66–86.

Fenno, Richard. 1978. *Home Style: House Members in their Districts*. Little, Brown.

Fowler, James H. (2006) "Legislative Cosponsorship Networks in the U.S House and Senate." *Social Networks* 28(4): 454–465.

Franzese, Robert and Jude Hayes. (2007) "Spatial-Econometric Models of Cross-Sectional Interdependence in Political-Science Panel and Time-Series-Cross-Section Data." *Political Analysis* 15(2): 140–64.

Gailmard, S. and J. W. Patty. (2013) Learning while Governing: Expertise and Accountability in the Executive Branch. Chicago, University of Chicago Press.

Girvan M. and Newman M. E. J. (2002) "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99: 7821–7826.

Hersh, Eitan and Clayton Nall. (2015) "The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records." *American Journal of Political Science*. Forthcoming.

Holland, P. (1986) "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945–970.
Hudges, Michael G. and M. Elizabeth Halloran. (2008) "Towards Causal Inference with Interference." *Journal of the American Statistical Association* 103(482): 832–843.
Katz, Daniel Martin, Joshua R. Gubler, Jon Zelner, Michael J. Bommarito II, Eric Provins, and Eitan Ingall. (2011) "Reproduction of Hierarchy? A Social Network Analysis of the American Law Professoriate." *Journal of Legal Education* 61(1): 1–28.
Krebs, Valdis E. (2002) "Uncloacking Terrorist Networks." *First Monday* 7(4): 1–15.
McClurg, Scott and J. K. Young. (2011) "Political Networks" *PS: Political Science and Politics* 44(01): 39–43.
Milgram, Stanley. (1967) "The Small World Problem." *Psychology Today* 60–67.
Newman, M. E. J. (2004) "Modularity and Community Structure in Networks." *Proceedings of the National Academy of Sciences* 103(23): 8577–8582.
Nickerson, David W. (2008) "Is Voting Contagious? Evidence from Two Field Experiments." *American Political Science Review* 102: 49–57.
Padgett, John F. (1993) "Robust Action and the Rise of the Medici, 1400–1434." *American Journal of Sociology* 98: 1259–1319.
Porter, Mason A., Peter J. Mucha, M. E. J. Newman, and Casey M. Warmbrand. (2005) "A Network Analysis of Committees in the U.S. House of Representatives." *Proceedings of the National Academy of Sciences* 102(20): 7057–7062.
Shalizi, Cosma and Andrew C. Thomas. (2011) "Homophily and Contagion Are Generically Confounded in Observational Social Network Studies." *Sociological Methods and Research* 40: 211–239.
Sinclair, Betsy. (2012) *The Social Citizen: Peer Networks and Political Behavior*. University of Chicago Press.
VanderWeele, T. J. (2011) "Sensitivity Analysis for Contagion Effects in Social Networks." *Sociological Methods and Research*, 40: 240–255.
VanderWeele, T J., E. L. Ogburn, and E. J. Tchetgen. (2012) "Why and When "Flawed" Network Analyses Still Yield Valid Tests of No Contagion." *Statistics, Politics, and Policy* 3: 1–11.
Wasserman, Stanley and Katherine Faust. (1994) *Social Network Analysis: Methods and Applications*. Cambridge University Press.
Zachary, Wayne W. (1977) "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research* 32(4): 452–473.