11
Networks

Elizabeth L. Ogburn and Alexander Volfovsky

CONTENTS
11.1 Introduction .......................................................... 171
11.2 Network Models .......................................................... 173
  11.2.1 Erdos–Renyi–Gilbert Model ............................................. 173
  11.2.2 Stochastic Blockmodel .................................................. 174
  11.2.3 General Latent Space Model ............................................ 176
  11.2.4 Testing for Network Structure ........................................ 177
    11.2.4.1 Bickel and Sarkar (2013) Main Result ......................... 177
    11.2.4.2 Volfovsky and Hoff (2014) Main Result ..................... 177
11.3 Observations Sampled from Network Nodes ....................... 178
  11.3.1 Terminology and Notation .......................................... 179
  11.3.2 Experiments in Networks .............................................. 179
    11.3.2.1 Fisherian Hypothesis Testing .................................. 180
    11.3.2.2 Estimation of Causal Effects in Randomized Experiments on Networks ........................................ 181
  11.3.3 Observational Studies ................................................. 182
    11.3.3.1 Peer Effects ..................................................... 182
    11.3.3.2 Instrumental Variable Methods .................................. 183
    11.3.3.3 Dependence ..................................................... 184
    11.3.3.4 Targeted Maximum Loss-Based Estimation for Causal Effects in Networks ........................................ 185
11.4 Discussion and Future Directions .................................... 185
References ................................................................. 186

11.1 Introduction
Networks are collections of objects (nodes or vertices) and pairwise relations (ties or edges) between them. Formally, a graph $G$ is a mathematical object composed of two sets: the vertex set $V = \{1, \ldots, n\}$ lists the nodes in the graph and the edge set $E = \{(i, j) : i \sim j\}$ lists all of the pairwise connections among the nodes. Here $\sim$ defines the relationship between nodes. The set $E$ can encode binary or weighted relationships and directed or undirected relationships. A common and more concise representation of a network is given by the $n \times n$ adjacency matrix $A$, where entry $a_{ij}$ represents the directed relationship from object $i$ to object $j$. Most often in statistics, networks are assumed to be unweighted and undirected, resulting in adjacency matrices that are symmetric and binary: $a_{ij} = a_{ji}$ is an indicator of whether $i$ and $j$ share an edge. A pair of nodes is known as a dyad; a network with $n$ nodes
has $\binom{n}{2}$ distinct dyads, and in an undirected graph, this is also the total number of possible edges. The degree of a node is its number of neighbors, or nodes with which it shares an edge. In a directed network, each node has an in-degree and an out-degree; in an undirected network, these are by definition the same. Some types of networks, such as family trees and street maps, have been used for centuries to efficiently represent relationships among objects (i.e., people and locations, respectively), but the genesis of the mathematical study of networks and their topology (graph theory) is usually attributed to Euler’s 1741 Seven Bridges of Königsberg (Euler, 1741).

Beginning with Euler’s seminal paper and continuing through the middle of the twentieth century, the formal study of networks or graphs was the exclusive domain of deterministic sciences such as mathematics, chemistry, and physics; its primary objectives were the description of properties of a given, fixed graph, for example, the number of edges, paths, or loops of a graph or taxonomies of various kinds of subgraphs. Random graph theory was first introduced by the mathematicians Erdos and Renyi (1959). A random graph is simply a random variable whose sample space is a collection of graphs. It can be characterized by a probability distribution over the sample space of graphs or by the graph-generating mechanism that produces said probability distribution. Random graph theory has become a vibrant area of research in statistics: random graph models have been used to describe and analyze gene networks, brain networks, social networks, economic interactions, the formation of international treaties and alliances, and many other phenomena across myriad disciplines. Common to all of these disparate applications is a focus on quantifying similarities and differences among local and global topological features of different networks. A random graph model indexes a probability distribution over graphs with parameters, often having topological interpretations; the parameters can be estimated using an observed network as data. Parameter estimates and model fit statistics are then used to characterize the topological features of the graph. We describe some such models and estimating procedures in Section 11.2.

Over the past 5–10 years, interest has grown in a complementary but quite different area of network research, namely the study of causal effects in social networks. Here, the network itself is not causal, but edges in the network represent the opportunity for one person to influence another. Learning about the causal effects that people may have on their social contacts concerns outcomes and covariates sampled from network nodes—outcomes superimposed over an underlying network topology—rather than features of the network topology. A small but growing body of literature attempts to learn about peer effects (also called induction or contagion) using network data (e.g., Christakis and Fowler, 2007, 2008, 2010): these are the causal effects that one individual’s outcome can have on the outcomes of his or her social contacts. A canonical example is infectious disease outcomes, where one individual’s disease status effects his or her contacts’ disease statuses. Interference or spillover effects are related but distinct causal effects that are also of interest in network settings; these are the causal effects that one individual’s treatment or exposure can have on his or her contacts’ outcomes. For example, vaccinating an individual against an infectious disease is likely to have a protective effect on his or her contacts’ disease statuses.

Simple randomized experiments that facilitate causal inference in many settings cannot be applied to the study of contagion or interference in social networks. This is because the individual subjects (nodes) who would be independently randomized in classical settings do not furnish independent outcomes in the network setting. In Section 11.3, we describe recent methodological advances toward causal inference using network data. Before that, in Section 11.2, we describe some current work on probabilistic network generating models. While it is possible to be relatively complete in our survey of the literature on causal inference for outcomes sampled from social network nodes, the literature on network
generating models is vast, and we limit our focus to models that we believe to be appropriate for modeling social networks and that we see as potential tools for furthering the project of causal inference using social network data.

Networks are inarguably examples of big data, but just how big they are is an open question. Big data often points to large sample sizes and/or high dimensionality. Networks can manifest both kinds of bigness, with a tradeoff between them. On the one hand, a network can be seen as a single observation of a complex, high-dimensional object, in which case sample size is small but dimensionality is high. On the other hand, a network can be considered as comprising a sample of size on the order of the number of nodes or the number of edges. In this case, sample size is large but complexity and dimensionality are less than they would be if the entire network were considered to be a single observation. In reality, the effective sample size for any given network is likely to lie somewhere between 1 and the number of nodes or edges. We are aware of only one published paper that directly tackles the question of sample size for network models: Kolaczyk and Krivitsky (2011) relate sample size to asymptotic rates of convergence of maximum likelihood estimates under certain model assumptions. A notion of sample size undergirds any statistical inference procedure, and most of the models we describe below inherently treat the individual edges or nodes as units of observation rather than the entire network. In some cases, this approach ignores key structure and complexity in the network and results in inferences that are likely to be invalid. We do not explicitly focus on issues of sample size and complexity in this chapter, but note that the tradeoff between network-as-single-complex-object and network-as-large-sample is a crucial and understudied component of statistics for network data.

11.2 Network Models

Different network models are designed to capture different levels of structure and variability in a network. We discuss three models in increasing order of complexity: the Erdos–Renyi–Gilbert model, the stochastic blockmodel and the latent space model. For each of these three models, we describe the parameters and properties of the model and, where appropriate, propose estimation and testing procedures to fit the model to observed network data. For more extensive surveys of the literature on random graph models, see Goldenberg et al. (2010) and Kolaczyk (2009).

11.2.1 Erdos–Renyi–Gilbert Model

The first random graph model, developed simultaneously by Paul Erdos and Alfred Renyi and by Edgar Gilbert, considers a random graph $G$ with a fixed number of nodes $n = |V|$ and a fixed number of undirected edges $e = |E|$ that are selected at random from the pool of $\binom{n}{2}$ possible edges. This induces a uniform distribution over the space of graphs with $n$ nodes and $e$ edges (Erdos and Renyi, 1959; Gilbert, 1959). A slight variation on this model fixes $n$ but only specifies $e$ as the expected number of edges in an independent sample—that is, the probability of any particular edge is given by $p = e/\binom{n}{2}$, so that $e$ is the expected but not necessarily exact number of realized edges. Under both of these formulations, the primary objects of interest are functions of $p$ and $n$; therefore, these models collapse all of the possible complexity in a network into two parameters and provide only a high-level overview of the network.

Much of the early work on the Erdos–Renyi–Gilbert model concentrated on its asymptotic behavior. One of the most celebrated results describes a phase change in the structure of Erdos–Renyi–Gilbert random graphs as a function of expected degree $\lambda = pn$, \ldots
namely the almost sure emergence of a giant component as $n \to \infty$ when $\lambda$ converges to a constant greater than 1. A giant component is a connected component (a subgraph in which all nodes are connected to one another by paths) that contains a strictly positive fraction of the nodes. According to the phase change results, all other components are small in the sense that none of them contain more than $\mathcal{O}(\log n)$ nodes. If $\lambda$ converges to a constant smaller than 1, then almost surely all components are small in this sense. Finally, for $\lambda = 1$, the largest component is almost surely $\mathcal{O}(n^{2/3})$ (Durrett, 2007). While this is a simplistic model for real-world networks, the emergence of the giant component is of practical importance when performing inference.

Perhaps the most significant criticism of this model is that real-world networks generally do not exhibit constant expected degrees across nodes; power law degree distributions (Barabási and Albert, 1999; Albert and Barabási, 2002), and departures from those (Clauset et al., 2009) are thought to be especially common. A natural extension of the Erdos–Renyi–Gilbert model, allowing for a power-law and other degree distributions, partitions the nodes into groups having different expected degrees, essentially interpolating several different Erdos–Renyi–Gilbert graphs (Watts and Strogatz, 1998). However, these models can become unwieldy; for example, efficiently generating a simple graph with a user-specified degree distribution requires sequential importance sampling (Blitzstein and Diaconis, 2011). Additionally, as we will see in Section 11.2.3, nonconstant degree distributions can be accommodated very intuitively by the latent space model.

A sample from an undirected Erdos–Renyi–Gilbert model with $n = 20$ nodes and edge probability $p = .25$ is displayed in the first panel of Figure 11.1. The expected degree for each node in the graph is $np = 4$; the observed average degree is 5.5. Estimation of the probability of edge formation under this model is straightforward via the binomial likelihood $p^e(1-p)^{n-e}$, where $e$ is the observed number of edges, and in this case $\hat{p} = .289$. The simplicity of this model and estimation procedure make it extremely appealing for inference when we cannot observe the full network, as the marginal distribution of any subgraph is easily computed, but the lack of any structure on the nodes and their degrees ensures that this model a simplification of reality in most cases.

### 11.2.2 Stochastic Blockmodel

A higher level of model complexity is achieved by the stochastic blockmodel (Holland et al., 1983; Nowicki and Snijders, 2001; Wang and Wong, 1987), which recognizes that a given

![Figure 11.1](image_url) Networks generated using the three models described in Section 11.2. They are: (a) an Erdos–Renyi–Gilbert graph with $p = 1/2$; (b) a stochastic blockmodel with probability matrix $P = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.4 \end{pmatrix}$; and (c) a directed latent space model based on the latent variable $s_{ij} = \alpha_i + \alpha_i \alpha_j + \epsilon_{ij}$ where $\alpha_i, \epsilon_{ij} \overset{iid}{\sim} \text{normal}(0, 1)$. 

174 Handbook of Big Data
network node is more likely to be connected to some nodes than to others (e.g., because it is more similar to some nodes than to others). This is codified in the assumption that the \( n \) nodes in a network are separated into \( k < n \) nonoverlapping groups and the relationship between any two nodes depends only on their group memberships. Nodes that belong to the same group are stochastically equivalent—that is, probability distributions over their edges are identical. The Erdos–Renyi–Gilbert model is a special case of a stochastic blockmodel with \( k = 1 \). By introducing structure in the form of multiple groups of nodes, the stochastic blockmodel captures an additional level of complexity and relaxes the assumption of identical expected degree across all nodes. While this model is not generally compatible with power law degree distributions, it is very flexible.

The parameter of a stochastic blockmodel is a \( k \times k \) probability matrix \( P \), where entry \( p_{ij} \) is the probability that a node in group \( i \) is connected by an edge to a node in group \( j \). Edges can be directed or undirected. The main constraint on \( P \) is that every entry is between 0 and 1. While \( p_{ij} = 0 \) or \( = 1 \) may be plausible, many estimation procedures make the assumption that \( p_{ij} \) is bounded away from 0 and 1. For undirected networks, an additional assumption is that the matrix \( P \) is symmetric (or simply upper triangular), while for directed networks, this requirement can be relaxed. In the middle panel of Figure 11.1, we see an undirected network simulated from a two block stochastic blockmodel where the probability of edges within each block is greater than that between the blocks. Additionally, one of the blocks has a higher probability of within group edges than the other.

The color coding in the figure clearly demarcates the two groups in the stochastic blockmodel and it is easy to see that each of the groups can be viewed marginally as an Erdos–Renyi–Gilbert graph. Given the group labels, we can perform inference as in the Erdos–Renyi–Gilbert case, treating each class of edges (within each group and between each pair of groups) individually to estimate the entries of \( P \). However, we rarely know group membership \textit{a priori}, making estimation of the stochastic blockmodel parameters much more complicated, since the group labels must be inferred. There are two main approaches to this estimation process. The first is a model-driven approach in which group membership is a well-defined parameter to be estimated jointly with the elements of \( P \); this approach can be viewed as a special case of the latent space model (see Section 11.2.3), where the multiplicative latent effects are \( k \) dimensional vectors with a single nonzero entry. The second approach is a heuristic approach involving spectral clustering. Given the number of clusters or blocks, \( k \), and an adjacency matrix \( A \), the first step is to find the \( k \) eigenvectors corresponding to the \( k \) largest eigenvalues (in absolute value; Rohe et al., 2011) of the graph Laplacian \( L = D^{1/2}AD^{1/2} \), where \( D_{ii} = \sum_j a_{ij} \). Treating the rows of the concatenated \( n \times k \) matrix of eigenvectors as samples in \( \mathbb{R}^k \), the next step is to run a \( k \)-means algorithm to cluster the rows into \( k \) nonoverlapping sets. These sets estimate the groups of the underlying stochastic blockmodel, and Rohe et al. (2011) provide bounds on the number of nodes that will be assigned to the wrong cluster under conditions on the expected average degree as well as the number of clusters. The power of their result lies in the fact that they allow the number of clusters to grow with sample size. Choi et al. (2012) developed similar results for a model-based approach to clustering in stochastic blockmodels. After clustering, estimation of \( P \) is straightforward.

It has recently been shown that stochastic blockmodels can provide a good approximation to a general class of exchangeable graph models characterized by a graphon: a function mapping the unit square to the unit interval and representing the limiting probability of edges in a graph (Airoldi et al., 2013). This suggests that, given the relatively weak assumption of an exchangeable graph model, a stochastic blockmodel approximation may lead to approximately valid inference. In other words, despite the simplicity of the stochastic blockmodel and the fact that the only structure it models is at the group level, it captures enough structure to closely approximate a large class of exchangeable random graph models.
Extensions to the stochastic blockmodel include mixed membership models (Airoldi et al., 2008) and degree corrected stochastic blockmodels, which induce power law distributions on the degrees (Karrer and Newman, 2011).

11.2.3 General Latent Space Model

A final level of complexity is afforded by the general latent space model. Under this model, the nodes of a network are embedded into a low-dimensional latent space, usually Euclidean, and the probability of an edge between any two nodes is a function of their latent positions. For example, in the latent distance model, the probability of a tie increases as the (Euclidean) distance between the latent positions decreases (Hoff et al., 2002). This captures both reciprocity and transitivity in the formation of network edges: since distances are symmetric, if the probability of an edge between $i$ and $j$ is high, then the probability of an edge between $j$ and $i$ will also be high, and the triangle inequality suggests that if $i$ and $j$ are close and $j$ and $t$ are close, then $i$ and $t$ are going to be close. Reciprocity and transitivity are properties that are thought to be important in real-world networks but are impossible to incorporate into the Erdos–Renyi–Gilbert model or the stochastic blockmodel. The inherent symmetry of the distance model rules out the possibility that certain nodes have a greater affinity for ties than others, and to circumvent this limitation, the general latent space model allows for asymmetric functions of the latent positions as well as for node- and dyad-specific covariates to affect the probability of tie formation. An example of a latent space model with additive and multiplicative functions of the latent positions as well as such covariates is described in detail below.

Consider an $n \times n$ asymmetric adjacency matrix $A$, representing a directed graph, and let $X$ be an $n \times n \times p$ array of observed characteristics. Each $n \times n$ slice of $X$ is either constant in the rows (representing a fixed effect that contributes to the propensity to send ties in the network, or sender effect); constant in the columns (representing a fixed effect that contributes to the propensity to receive ties in the network, or receiver effect); or neither, representing dyadic effects. We can model $a_{ij}$ as the indicator $1_{s_{ij}>0}$ that $s_{ij} > 0$, where $s_{ij} = X_{ij} \theta + \alpha_i + \beta_j + u_i^t v_j + \epsilon_{ij}$, $X_{ij}$ is the $p$-dimensional vector of covariates associated with the relationship between nodes $i$ and $j$, $\alpha_i$ is an additive sender effect, $\beta_j$ is an additive receiver effect, and $u_i^t v_j$ is a multiplicative effect (as it is the projection of $u_i$ in the direction of $v_j$ in the latent space) that captures similarity between nodes $i$ and $j$ (Hoff, 2005). This model is a generalization of the social relations model of Warner et al. (1979). Reciprocity can be introduced into the model by allowing for the error terms ($\epsilon_{ij}, \epsilon_{ji}$) to be correlated. Here $X_{ij}$ might include sender-specific information, receiver-specific information, or dyadic information. The additive latent effects $\alpha_i$ and $\beta_j$ contain information about the affinity of nodes $i$ and $j$ to send and receive ties in general, while the multiplicative effect $u_i^t v_j$ contains the information on the latent similarity of the two nodes. In particular, if the nodes are close in the latent space ($u_i^t v_j > 0$), then the probability of a tie is increased and if they are far apart ($u_i^t v_j < 0$), then it is decreased.

The third panel of Figure 11.1 displays a directed network generated from the latent class model described above (without covariates and with a one-dimensional latent space). The two sets of nodes are colored according to the sign of $\alpha_i$. The emergence of the two clusters is due to the multiplicative effect $\alpha_i \alpha_j$: ties are more likely between individuals for whom the signs of $\alpha_i$ match. This demonstrates the ability of this model to capture stochastic blockmodel behavior. Each node has its own probability of sending a tie to another node, which allows for much greater flexibility than the blockmodel. The yellow nodes send out more ties to the blue nodes than they receive from the blue nodes, due to the additional additive effect of $\alpha_i$ in the model as nodes with $\alpha_i > 0$ have a higher probability of sending out ties.
General latent space models can be fit to data from a real network via a Markov chain Monte Carlo algorithm (Hoff, 2005). A salient advantage of these models is their ability to model restrictions on the sampling design according to which data from a network is collected. For example, when collecting information about friendship networks in a survey setting, researchers often ask individuals to name their top five friends. This type of network sampling scheme is known as the fixed rank nominations scheme and is explored in detail in Hoff et al. (2013). When this sampling design is ignored, estimation of effects on the probability of edge formation is imprecise, potentially leading to incorrect inference. The ability to account for restrictions on the sampling design makes this general class of models potentially promising for causal inference using data from partially observed networks.

11.2.4 Testing for Network Structure

All three models described above make assumptions about the similarity among nodes in a network, allowing for stochastic equivalence, homophily or both (Hoff, 2007). Recent work by Bickel and Sarkar (2013) and Volfovsky and Hoff (2014) proposes two different testing procedures for the amount of similarity among nodes in a relational dataset. These tests can inform of the validity of the assumptions underlying the three models described above; rejecting the null hypothesis can be seen as an indication that the data exhibit more complexity than the null model allows. Bickel and Sarkar (2013) leverage Tracy–Widom theory to develop a hypothesis test for the null hypothesis that the graph comes from an Erdos–Renyi–Gilbert model (rejection of the null suggests the presence of blocks or clusters, as in the stochastic blockmodel), while Volfovsky and Hoff (2014) develop a test for correlation among the rows and columns of a relational data matrix within the context of the normal model (defined below) and provide an extension to the latent space model for networks that informs the choice of dimension of the latent space.

11.2.4.1 Bickel and Sarkar (2013) Main Result

Let \( A \) be the adjacency matrix of an Erdos–Renyi random graph with parameter \( p \) and let \( \hat{p} \) be the estimate of \( p \). For the centered and scaled adjacency matrix,

\[
\tilde{A} = \frac{A - (n\hat{p}1^t - \hat{p}I)}{\sqrt{(n-1)\hat{p}(1-\hat{p})}}
\]

the limiting distribution of the leading eigenvalue, \( \lambda_1(\tilde{A}) \) is given by

\[
n^{2/3}(\lambda_1(\tilde{A}) - 2) \xrightarrow{d} \text{Tracy–Widom}
\]

Given this limiting distribution, it is now easy to construct a test for the null hypothesis that a graph is generated from an Erdos–Renyi model with parameter \( p \). Bickel and Sarkar (2013) propose using this test to recursively find the blocks of the stochastic blockmodel of Section 11.2.2. In particular, if the test rejects the null hypothesis, then the algorithm splits the graph into two blocks and recurses the test on each block.

11.2.4.2 Volfovsky and Hoff (2014) Main Result

Let \( S \) be an \( n \times n \) random matrix distributed according to the matrix normal distribution with zero mean matrix and Kronecker covariance structure \( \Sigma_{col} \otimes \Sigma_{row} \). If the entries of \( S \) represent a weighted relationship between actors \( i \) and \( j \) in a network, then the identities \( E[ZZ^t] = \Sigma_{row} \text{tr}(\Sigma_{col}) \) and \( E[S^tS] = \Sigma_{col} \text{tr}(\Sigma_{row}) \) suggest the interpretation of \( \Sigma_{row} \) and
\( \Sigma_{\text{col}} \) as the covariance of the network nodes as senders and as receivers, respectively. Under the normal model, using a single sample matrix, we can construct a likelihood ratio test for row and column dependence under this model. This is because both under the null hypothesis that \( \Sigma_{\text{row}} \) and \( \Sigma_{\text{col}} \) are diagonal positive definite and under the alternative hypothesis that \( \Sigma_{\text{row}} \) and \( \Sigma_{\text{col}} \) are positive definite, the matrix normal likelihood is bounded for a single matrix \( S \).

To connect this test to binary networks we make use of the latent model of Section 11.2.3. Consider a graph with adjacency matrix \( A \) where the probability of an edge is modeled as a function of latent multiplicative sender and receiver effects, \( u_i, v_j \in \mathbb{R} \). Writing this model in the latent variable formulation we write \( a_{ij} = 1 \) if \( s_{ij} = u_i v_j + \epsilon_{ij} > -\gamma \), where \( s_{ij} \) is the product of sender and receiver effects plus a noise term. The \( n \times R \) matrices \( U \) and \( V \) formed by stacking the vectors \( u_i, i = 1, \ldots, n \) and \( v_j, j = 1, \ldots, n \), describe the row and column heterogeneity in \( S \), respectively, since \( E[SS^t] = U(V^tV)U^t + I \) and \( E[V^tS] = V(U^tU)V^t + I \). The main challenge here is choosing the appropriate rank \( R \) to use for the model. This can be achieved by performing the likelihood ratio test described above on the latent scale for a sequence of ranks. When the test stops rejecting the null of no dependence among the nodes, it means that the latent vectors capture all of the dependence between senders and between receivers present in \( A \). The above-described likelihood ratio test can be applied on the latent scale.

Testing procedures have been recently developed for more complicated models: Yang et al. (2014) proposed a novel test for a graphon, a limit object for random graphs. The models described in this section carry with them the underlying assumption of infinite exchangeability, that is, that the finite networks being modeled comprise subsamples of an infinite graph where the nodes can be relabeled without affecting the joint distribution of the graph. In a recent work, Volfovsky and Airoldi (2014) demonstrated that distributions that satisfy the substantially weaker assumption of finite exchangeability are close in total variation distance to the distribution of graphs satisfying the stronger assumption. This suggests that inference made under the stronger assumption is likely to be appropriate even when only the weaker assumption holds.

### 11.3 Observations Sampled from Network Nodes

In most of the literature on networks, the network itself is the object of study: features of network topology and probability models for edge formation, node addition, and network generation. But situations abound in which the network can be thought of as a nuisance parameter and interest is instead in the behavior and attributes of the network nodes. For social networks, which are our focus here, we might be interested in how behaviors spread from one person to their social contacts, how attributes cluster among network nodes, or how one person’s randomly assigned treatment might affect their social contacts’ outcomes. The underlying network gives important structure to the data—network ties represent opportunities for contagion or correlation—but interest is in the effects, not the determinants, of the network topology. Research methods for analyzing data on outcomes sampled from a single network is nascent and remains underdeveloped. In what follows we describe recent work in this area and conclude with a description of open problems. Most of the research in this area assumes that the network is fixed and known.
11.3.1 Terminology and Notation

In this section, we will use the terms nodes, subjects, and individuals interchangeably. This is a slight abuse of the term node but should cause no confusion as long as we are not concerned with network topology except insofar as it informs relations among the subjects and the observations they furnish. Let \( Y_i \) represent an outcome for node/subject \( i \) and let \( Z_i \) be a treatment or exposure for individual \( i \). Sometimes, we will index the outcomes with time, that is, \( Y_i^t \) is the outcome for subject \( i \) at time \( t \). If we have observed \( n \) subjects, \( Y \) is an \( n \)-dimensional vector of outcomes for subjects 1 through \( n \), possibly indexed by time, and \( Z \) is an \( n \)-dimensional vector of exposures or treatments for subjects 1 through \( n \).

Questions about the influence one subject has on the outcome of another subject are inherently questions about causal effects, which are defined in terms of potential or counterfactual outcomes (see, e.g., Hernan, 2004; Rubin, 2005). In general, a unit-level potential outcome, \( Y_i(z) \), is defined as the outcome that we would have observed for subject \( i \) if we could have intervened to set that subject’s treatment or exposure \( Z_i \) to value \( z \), where \( z \) is in the support of \( Z \). But social networks represent a paradigmatic opportunity for interference: one subject’s exposure may affect not only his own outcome but also the outcomes of his social contacts and possibly other subjects. This means that the traditional unit-level potential outcomes are not well defined. Instead, \( Y_i(z) \) is the outcome that we would have observed if we could have set the vector of exposures for the entire population, \( Z \), to \( z = (z_1, ..., z_n) \), where for each \( i \), \( z_i \) is in the support of \( Z \). In the presence of interference, there are many types of causal effects that might be of interest. Unit-level causal effects isolate the effect of a subject’s own exposure on his outcome. These effects are usually defined either holding other subjects’ exposures constant or averaging over possible exposure distributions for other subjects. Spillover effects quantify the effect on subject \( i \)’s outcome of other subjects’ exposures. For a binary treatment, the contrast \( \sum_{i=1}^{n} E[Y_i(1) - Y_i(0)] \) is often of interest. This is the difference in average potential outcomes in a world in which every subject is assigned to treatment compared to a world in which every subject is assigned to control. Eckles et al. (2014) call this particular estimand the average treatment effect (ATE); however, to differentiate this estimand from the ATE defined in the absence of interference, we will adopt the terminology of Halloran and Struchiner (1995) and refer to \( \sum_{i=1}^{n} E[Y_i(1) - Y_i(0)] \) as the overall ATE. See Ogburn and VanderWeele (2014a) and references therein for further discussion of causal effects in the presence of interference and of the distinction between interference and contagion.

We distinguish between interference, which is present when one subject’s treatment or exposure may affect others’ outcomes, and peer effects or contagion, which are present when one subject’s outcome may influence or transmit to other subjects (Ogburn and VanderWeele, 2014a). In the literature on peer effects, an ego is defined as a subject whose outcome we are interested in studying and the ego’s alters are the ego’s neighbors, that is, subjects who share an edge with the ego in the underlying network. A peer effect is a causal effect on an ego’s outcome at time \( t \) of his alter’s outcome at time \( s \) for some \( s < t \).

11.3.2 Experiments in Networks

The design of randomized experiments in networks, when subjects and their treatments may not be independent, is an area of recent but increasing interest. Without attempting to be exhaustive, this section reviews what we consider to be the most important threads in this line of research.

Experiments are difficult for two reasons: interference, which may make it impossible to simultaneously observe the treatment condition (e.g., every subject is treated) for some
units and the control condition (e.g., no subject is treated) for others, and dependence among observations, which makes the estimation of standard errors difficult. Fisherian randomization-based inference can circumvent the problem of dependence, as described in Sections 11.3.2.1 and 11.3.2.2. Other randomization-based approaches assume that the only source of dependence is interference, and therefore that observations are independent conditional on treatment. There are two approaches to dealing with the challenge of interference in the randomization literature: assumptions on the limits of interference and attempts to minimize bias due to interference. Both are discussed below.

11.3.2.1 Fisherian Hypothesis Testing

Most of the work on inference about outcomes sampled from network nodes concerns randomized experiments. But there is an important distinction to be made between inference from randomized experiments and the specific case of Fisherian randomization-based inference, pioneered by Fisher (1922) and applied to network-like settings by Rosenbaum (2007) and Bowers et al. (2013). Fisherian randomization-based inference is founded on the very intuitive notion that, under the null hypothesis of no effect of treatment on any subject (sometimes called the sharp null hypothesis to distinguish it from other null hypotheses that may be of interest), the treated and control groups are random samples from the same underlying distribution. Fisherian randomization-based inference treats outcomes as fixed and treatment assignments as random variables: quantities that depend on the vector of treatment assignments are the only random variables in this paradigm. Therefore, dependence among outcomes is a nonissue.

In an influential paper, Rosenbaum (2007) proposed the use of distribution-free statistics (statistics whose distribution under the null hypothesis can be determined a priori, without reference to the data) to perform hypothesis tests and derive confidence intervals in the presence of interference. The Mann–Whitney statistic, for example, takes all pairs consisting of one treated and one untreated subject and counts the number of times that the treated subject had a greater outcome than the untreated subject. The null distribution of this statistic is known a priori and does not depend on any aspect of the data-generating distribution of the outcomes; it is therefore agnostic to any dependence among the outcomes. After deriving the null distribution for such a statistic, one can compare the observed distribution of the statistic to the null distribution to perform a hypothesis test and to derive confidence intervals for the magnitude of the departure from the null hypothesis in the event that it is rejected.

A slightly different version of randomization-based inference is developed by Bowers et al. (2013). Under the null hypothesis, that is, supposing that we are observing outcomes in the world in which the null is true, we can generate the randomization-based null distribution of any test statistic by enumerating all of the possible treatment assignment permutations and recalculating the test statistic for each one. For example, suppose we have a network with four nodes and our experiment dictates assigning half to treatment. Our test statistic is the first in the list above—the mean outcome among treated minus the mean outcome among controls. We can compare the observed test statistic to the statistic’s null distribution and ask the question “how likely was this value to have occurred by chance, assuming that the null hypothesis is true?”
Of course, Bowers et al. (2013) consider settings considerably more complex than the simple example above. Specifically, they propose testing any parametric model \( H(y_i(z); \beta, \tau) = y_i(0) \) that maps one subject’s potential outcome under an arbitrary treatment assignment vector \( z \) to the same subject’s potential outcome under treatment assignment \( 0 \), that is under the assignment in which no subject receives treatment. The two parameters \( \beta \) and \( \tau \) index the effect of subject \( i \)'s own treatment and the spillover effect of others’ treatments, respectively, on subject \( i \)'s outcome. Assuming the correct parametric form for the model, the true values of the parameters \( \beta \) and \( \tau \) are those for which \( \{H(y_i(z); \beta, \tau) : i \text{ treated}\} \) and \( \{H(y_i(z); \beta, \tau) : i \text{ control}\} \) are both collections of estimates of \( Y(0) \) and are therefore random samples from the same underlying distribution. The authors propose tests for this using the Komolgorov–Smirnov test statistic, which is a nonparametric comparison of the probability distribution of the estimated potential outcomes \( Y(0) \) between the treated and control groups. They derive p-values and confidence intervals for \( \beta \) and \( \tau \), which facilitate testing specific hypotheses about unit-level and spillover effects of treatment.

### 11.3.2.2 Estimation of Causal Effects in Randomized Experiments on Networks

Fisherian randomization-based inference, described in Section 11.3.2.1, permits hypothesis testing but not the estimation of causal effects. A few recent papers have taken on the challenge of estimating spillover effects and the ATE of a randomized treatment in the presence of interference.

Toulis and Kao (2013) discussed estimation of the spillover effect of having \( k \) of one’s neighbors assigned to treatment compared to the null condition of having no treated neighbors. They call the state of having \( k \)-treated neighbors \( k \)-exposure. Estimation of the effect of \( k \)-exposure is challenging in part because, if nodes are randomized independently to treatment and control, there may be no nodes with \( k \)-treated neighbors and/or no nodes with no treated neighbors. Define \( V_k \) to be the set of all nodes with at least \( k \) neighbors, that is, the set of nodes eligible \( a \) priori to be exposed or unexposed to \( k \)-exposure. Toulis and Kao (2013) proposed a randomization scheme, insulated neighbors randomization, that randomly selects some portion shared neighbors (nodes that are neighbors of at least two nodes in \( V_k \)) to be assigned to control before sequentially assigning eligible nodes in \( V_k \) to either \( k \)-exposure or to the null condition. Assigning nodes to the exposure or null conditions rather than to treatment or control ensures that these conditions will be observed in the data. However, assigning a node to be in one condition or the other determines the treatment assignments of that node’s neighbors and may eliminate or reduce the possibility of finding other nodes with unassigned neighbors, that is, nodes that can be randomized in the next step of the algorithm. Assigning some portion of shared neighbors to control in a first step increases the number of nodes available to be randomized during the next steps of the algorithm.

Ugander et al. (2013) and Eckles et al. (2014) focused on estimation of the overall ATE (defined above, Section 11.3.1). These two papers propose and elaborate on a new method of randomization, graph cluster randomization, that assigns clusters of adjacent nodes to treatment or control. The informal rationale is that, for nodes in the center of the cluster, the world looks like one in which the entire network is assigned to either treatment or control. Therefore, for these nodes, the observed outcome should be equal to or close to the potential outcome that we would have observed if we had assigned the entire network, instead of just the cluster, to either treatment or control. Ugander et al. (2013) assumes that various local exposure conditions for subject \( i \) (e.g., \( i \) and all of \( i \)'s neighbors are treated or \( i \) and \( x\% \) of \( i \)'s neighbors are treated) result in an observed outcome for subject \( i \) that is precisely equal to the potential outcome we would have observed for subject \( i \) if every subject
had been treated. The overall ATE can then be estimated by comparing nodes with these exposure conditions to nodes with the prespecified null exposure conditions (e.g., the node and none of its neighbors are treated). The authors derive Horwitz–Thompson estimators of the ATE under this class of assumptions and using graph cluster randomization. Graph cluster randomization reduces the variance of standard Horwitz–Thompson estimators with respect to naive randomization schemes, in which nodes are randomly assigned to treatment conditions independently of their neighbors, because it increases the probability of observing nodes who are in the prespecified exposure conditions of interest.

Eckles et al. (2014) relax the assumption that local exposure conditions result in observed outcomes that are identical to the potential outcome had the entire network been assigned the same condition; instead, they investigate conditions under which graph cluster randomization will reduce bias in the estimated overall ATE compared to naive randomization when that assumption does not hold. Specifically, they describe a data-generating process in which one subject’s treatment affects his or her own outcome, which in turn affects his neighbors’ outcomes, then the neighbors’ outcomes can affect their neighbors’ outcomes, and so on. This is an example of interference by contagion (Ogburn and VanderWeele 2014a). Under this model, every node’s outcome can be affected by every other node’s treatment assignment, but the effect of one node’s treatment on another’s outcome is likely to decay with network distance. Intuitively, if \( i \) and all of \( i \)'s neighbors are treated, then the observed value \( Y_i \) is closer to \( Y_i(Z = 1) \) than it would have been if some or none of its neighbors had been treated. Eckles et al. (2014) formalize this intuition and prove that graph cluster randomization reduces bias compared to independent randomization. They also explore bias-variance tradeoffs: larger clusters result in observed conditions that are closer (so to speak) to the desired counterfactual world in which the entire network had the same treatment assignment; larger clusters also result in fewer units to be randomized and therefore in larger variance of Horwitz–Thompson estimators. See Eckles et al. (2014) for details and recommendations.

Two related and recent methods bear brief mentioning: Choi (2014) proposed a method for detecting clustering of outcomes that would be consistent with spillover from one subject to another and is highly robust to misspecification of the network topology. Aronow and Samii (2013) proposed a method for estimating unit-level and spillover effects of randomized treatments in relatively sparse networks. Their method is based on an exposure model, similar to the model \( H(y_i(z); \beta, \tau) = y_i(0) \) of Bowers et al. (2013), and the exposure conditions of Ugander et al. (2013), which partitions the space of all potential outcomes \( Y_i(z) \) into equivalence classes.

### 11.3.3 Observational Studies

In many settings randomization is not feasible. It may be unethical to randomize individuals to conditions that are known to be harmful, or impossible to randomize them to conditions that are more complex than a simple treatment assignment. When randomization is not possible, we rely on observational data to provide information about treatment effects and peer effects.

#### 11.3.3.1 Peer Effects

A growing body of research regresses an ego’s outcome at time \( t \) on the outcome of one of his alters at time \( t - 1 \) using generalized linear models (GLMs) and, for longitudinal data, generalized estimating equations (GEEs), reporting the coefficients on the alters’ outcomes as estimates of peer effects (Ali and Dwyer, 2009; Cacioppo et al., 2009; Christakis and Fowler, 2007, 2008, 2013; Fowler and Christakis, 2008; Lazer et al., 2010; Rosenquist
et al., 2010). This work has come under two major lines of criticism. First, there is the problem of adequate control for confounding when trying to tease apart homophily and peer influence (Cohen-Cole and Fletcher, 2008; Lyons, 2011; Manski, 1993, 2013; Noel and Nyhan, 2011; Shalizi and Thomas, 2011; VanderWeele, 2011). Homophily is the tendency of people who share similar characteristics to be adjacent in a social network (to be friends, to be related to one another, to work in the same company, etc.), and, like peer influence, it manifests as correlated outcomes among nodes that are close to one another in the network. Christakis and Fowler (2007, 2008, 2010, 2013) proposed controlling for homophily by including the alter’s and the ego’s past outcomes in their regression models, however Shalizi and Thomas (2011) demonstrated that this does not suffice to control for homophily on unobserved characteristics. In general, unless every covariate affecting friendship formation is observed and controlled for, it is not possible to differentiate between homophily and peer influence without strong parametric assumptions, and possibly not even then (Shalizi and Thomas, 2011).

An orthogonal line of criticism of the GLM/GEE approach to assessing peer influence is that these classes of models are intended for independent observations, while observations sampled from a social network are likely to be dependent (Lyons, 2011; VanderWeele et al., 2012). Indeed, the very hypothesis of peer effects entails dependence among observations. When a statistical procedure is grounded in the assumption of independence, while in fact observations are positively correlated (as we would expect them to be for contagious outcomes in a social network), then the resulting standard errors and statistical inference will be anticonservative. In some cases the assumption of independence may hold under the null hypothesis of no peer effects (VanderWeele et al., 2012), but it is unknown whether tests that rely on this fact have any power under the alternative hypothesis to detect the presence of the causal effects of interest (Shalizi, 2012).

Although problematic in myriad ways, using GLMs and GEEs to regress an ego’s outcome on his alters’ lagged outcomes is the most widespread and, until recently, the only proposal for estimating peer effects using social network data. Below we discuss two refinements to these methods (one addresses the reflection problem and the other addresses the problem of dependence) and a new proposal that avoids the methodological issues that plague the use of GLMs and GEEs but that requires a very rich set of observations over time and over network nodes.

### 11.3.3.2 Instrumental Variable Methods

O’Malley et al. (2014) proposed an instrumental variable (IV) solution to the problem of disentangling peer effects from homophily. An instrument is a random variable, $V$, that affects exposure but has no effect on the outcome conditional on exposure. When the exposure—outcome relation suffers from unmeasured confounding but an instrument can be found that is not confounded with the outcome, IV methods can be used to recover valid estimates of the causal effect of the exposure on the outcome. In this case, there is unmeasured confounding of the relation between an alter’s outcome at time $t-1$ and an ego’s outcome at time $t$ whenever there is homophily on unmeasured traits. We refer the reader to Angrist et al. (1996), Angrist and Pischke (2008), Greenland (2000), and Pearl (2000) for accessible reviews of IV methods.

O’Malley et al. (2014) propose using a gene that is known to be associated with the outcome of interest as an instrument. In their paper, they focus on perhaps the most highly publicized claim of peer effects, namely that there are significant peer effects of body mass index (BMI) and obesity (Christakis and Fowler, 2007). If there is a gene that affects BMI but that does not affect other homophilous traits, then that gene is a valid instrument for the effect of an alter’s BMI on his ego’s BMI. The gene affects the ego’s BMI only through
the alter’s manifest BMI (and it is independent of the ego’s BMI conditional on the alter’s BMI), and there is unlikely to be any confounding, measured or unmeasured, of the relation between an alter’s gene and the ego’s BMI.

There are two challenges to this approach. First, the power to detect peer effects is dependent in part upon the strength of the instrument—exposure relation which, for genetic instruments, is often weak. Indeed, O’Malley et al. (2014) reported low power for their data analyses. Second, in order to assess peer effects at more than a single time point (i.e., the average effect of the alter’s outcomes on the ego’s outcomes up to that time point), multiple instruments are required. O’Malley et al. (2014) suggest using a single gene interacted with age to capture time-varying gene expression, but this could further attenuate the instrument—exposure relation and this method is not valid unless the effect of the gene on the outcome really does vary with time; if the gene-by-age interactions are highly collinear then they will fail to act as differentiated instruments for different time points.

### 11.3.3.3 Dependence

To deal with the problem of dependent observations in the estimation of peer effects, Ogburn and VanderWeele (2014b) proposed analyzing a subsample of alter-ego pairs that are mutually independent conditional on carefully chosen observations on intermediate nodes in the network. Testing and estimation of peer effects can then proceed using GLMs or GEEs, conditional on the appropriate set of observations. Ogburn and VanderWeele (2014b) focus on two specific types of peer effects that are of interest in the context of infectious diseases, but their main idea is quite general. We briefly describe their method in the context of a toy example.

Consider the network of five nodes in a *bow tie* formation depicted in Figure 11.2.

If the only dependence among observations sampled from nodes one through five is due to peer effects, and if these effects cannot be transmitted more than once per unit of time, then \((Y_{1t}^{t-1}, Y_{2t}^{t-1})\) is independent of \((Y_{4t}^{t-1}, Y_{5t}^{t-1})\) conditional on \((Y_{3t}^{t-1}, Y_{3t}^{t-2}, ..., Y_{3t}^{0})\).

Then, we can use standard statistical models to regress \(Y_{5t}^{t}\) on \(Y_{4t}^{t-1}\) and \(Y_{1t}^{t-1}\) on \(Y_{2t}^{t}\), conditioning on the appropriate observations. In some cases conditioning on \(Y_{3t}^{t-1}\) or on a summary of \((Y_{3t}^{t-1}, Y_{3t}^{t-2}, ..., Y_{3t}^{0})\) may suffice. The requisite conditioning event could change the interpretation of the causal effect of interest (i.e., the conditional causal effect may be different from the unconditional causal effect); in these cases we can perform hypothesis tests about the unconditional causal effects but not estimate them. In other cases, conditioning does not change the interpretation and estimation and testing are both feasible.

The conditional inference proposal of Ogburn and VanderWeele (2014b) suffers from low power both because it requires subsampling from the original network, resulting in a diminished sample size, and because when the conditional and unconditional causal effects differ the former is generally biased towards the null value relative to the latter. This direction of bias is what permits valid hypothesis tests in the conditional models, but it also results in reduced power. However, this proposal is first step in the ongoing endeavor to

![Figure 11.2](image)
Networks 185
develop methods for valid inference using data collected from a single network. Furthermore, it sheds light on the issues of model misspecification and invalid standard errors for which other proposals to use GLMs and GEEs to assess peer effects have been criticized.

11.3.3.4 Targeted Maximum Loss-Based Estimation for Causal Effects in Networks

Building upon important work by Mark van der Laan on causal inference for nonindependent units (van der Laan, 2014), van der Laan et al. (forthcoming) propose targeted minimum loss-based estimation (TMLE) of average unit-level treatment effects in the presence of interference, of average spillover effects as defined by user-specified exposure functions, of peer effects, and of changes to the network itself (i.e., addition and/or removal of edges). Specifically, they propose semiparametric efficient estimators for $E \left[ \frac{1}{n} \sum_{i=1}^{n} Y_i(z) \right]$, the expected average potential outcome under exposure regime $Z = z$, where the definition of $Z_i$ is broadened to include any variable that precedes the outcome $Y_i$. It could include a list of edges emanating from node $i$, previously occurring outcomes for the alters of node $i$, or manipulated or observed exposure or treatment variables of node $i$ itself or of the alters of node $i$. A large class of causal effects can be estimated as the difference of $E \left[ \frac{1}{n} \sum_{i=1}^{n} Y_i(z) \right]$ and $E \left[ \frac{1}{n} \sum_{i=1}^{n} Y_i(z') \right]$ for exposures $z$ and $z'$ of interest. These estimators can also handle longitudinal data, that is, estimation of $E \left[ \frac{1}{n} \sum_{i=1}^{n} Y_i(z') \right]$, where $Y_i'$ is the outcome for individual $i$ at time $t$ and $Z'$ is the collection of vectors $(Z_1', ..., Z_t')$.

The details of TMLE are beyond the scope of this chapter but we refer the reader to van der Laan and Rose (2011) for an introduction. For the purposes of comparison to the other methods described above, it suffices to understand that the estimators in question are based on the efficient influence function for the observed data functional that, under some assumptions, identifies the causal estimand $E \left[ \frac{1}{n} \sum_{i=1}^{n} Y_i(z) \right]$. The assumptions required for identification are strong and require observation of a rich data structure, which may not always be feasible, affordable, or practical. When these assumptions are met and the mean counterfactual outcome is identified, the proposed estimation procedure is optimal and principled. While the other methods can be seen as somewhat ad hoc attempts to overcome suboptimal data, this is an optimal approach for very rich data.

Specifically, van der Laan et al. (forthcoming) make two strong conditional independence assumptions in addition to the usual assumption of no unmeasured confounders: first, that at time $t$, node $i$ is independent of all other nodes conditional on all the data from time $t - 1$, and second, that at time $t$, node $i$ is independent of all data from time $t - 1$ conditional on his alters’ data from time $t - 1$. Essentially, these two assumptions imply that any dependence of one node on another is due to information transmitting at discrete times along one edge at a time and that the network (and all data on every node) is observed at least as frequently as information can be transmitted. Although these assumptions are strong, they are to some extent under the researcher’s control; meeting them is partially a matter of the richness of the data and therefore of study design and/or resources. The additional assumptions required for the consistency and asymptotic normality of the estimators are relatively weak; these methods can account for the presence of a small number of highly connected and highly influential nodes and for a model of asymptotic growth in which the number of edges per node grows (slowly) with $n$.

11.4 Discussion and Future Directions

The preceding two sections comprise an overview of state-of-the-art approaches to network modeling and to causal inference using observations sampled from a fixed network. Unfortunately, in many scenarios, the network is only partially observed or is observed
with noise, so treating it as fixed and known (as most of the methods in Section 11.3 do) may be inappropriate. This makes the already-difficult project of causal inference even more challenging. The naive approach to causal inference using incomplete network data would be to impute missing data in a first step and then to proceed with causal inference as if the data estimated in the first step were fixed and known. The primary downside of this procedure is that it does not incorporate the uncertainty from the network fitting into the uncertainty about the causal effects; a procedure that performs both tasks simultaneously is highly desirable.

In Lunagomez and Airoldi (2014), the authors tackle the problem of jointly modeling the sampling mechanism for causal inference as well as the underlying network on which the data was collected. The model selected for the network in this chapter is the simple Erdős–Renyi model that depends on a single parameter $p$. Since the network is not fully observed under the sampling scheme discussed in this chapter (respondent-driven sampling), the network model is chosen to accommodate marginalizing out the missing network information in a Bayesian framework. The use of a simple network model makes computation tractable, but the framework proposed by Lunagomez and Airoldi (2014) can theoretically be relaxed to incorporate any network model.

An alternative approach could be based on the proposal of Fosdick and Hoff (2013). While these authors do not discuss estimation of causal effects, their procedure for the joint modeling of network and nodal attributes can be adapted to a model-based causal analysis. In particular, the authors leverage Section 11.2.3 to first test for a relationship between nodal attributes $Y_i$ and the latent position vector $\text{lat}_i = (a_i, b_i, U_i, V_i)$ and, when evidence for such a relationship is found, to jointly model the vector $(Y_i, \text{lat}_i)$. Considering the nodal attributes as potential outcomes and writing $(Y_i(0), Y_i(1), \text{lat}_i)$, it should in principle be possible to jointly model the full data vector using the same Markov chain Monte Carlo procedure as in Fosdick and Hoff (2013).

Without fully observing the network it is difficult to precisely define, let alone to estimate, the causal effects discussed in Section 11.3. Jointly modeling network topology, to account for missing data or subsampling, and causal effects for observations sampled from network nodes, is one of the most important and challenging areas for future research. The work of Lunagomez and Airoldi (2014) and Fosdick and Hoff (2013) point toward powerful and promising solutions, but much work remains to be done.

References


