

Published in final edited form as:

Health Serv Outcomes Res Methodol. 2008 December 1; 8(4): 222–269. doi:10.1007/s10742-008-0041-z.

The Analysis of Social Networks

A. James O'Malley* and Peter V. Marsden#

A. James O'Malley: omalley@hcp.med.harvard.edu; Peter V. Marsden: pvm@wjh.harvard.edu

* Associate Professor of Statistics, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115-5899

Professor of Sociology and Harvard College Professor, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138

Abstract

Many questions about the social organization of medicine and health services involve interdependencies among social actors that may be depicted by networks of relationships. Social network studies have been pursued for some time in social science disciplines, where numerous descriptive methods for analyzing them have been proposed. More recently, interest in the analysis of social network data has grown among statisticians, who have developed more elaborate models and methods for fitting them to network data. This article reviews fundamentals of, and recent innovations in, social network analysis using a physician influence network as an example. After introducing forms of network data, basic network statistics, and common descriptive measures, it describes two distinct types of statistical models for network data: individual-outcome models in which networks enter the construction of explanatory variables, and relational models in which the network itself is a multivariate dependent variable. Complexities in estimating both types of models arise due to the complex correlation structures among outcome measures.

Keywords

Correlation; exponential random graph model; latent-space model; network autocorrelation model; social relationship; social network

1. Introduction

Social network analysis studies structures of relationships linking individuals (or other social units, such as organizations) and interdependencies in behavior or attitudes related to configurations of social relations. Since many medical and health-related phenomena involve interdependent actors (e.g. patients, nurses, physicians, and hospitals), networks are of increasing interest to health services researchers. Among many other examples are social support networks that may serve to improve individual well-being by providing psychosocial or tangible resources (Berkman and Syme 1979); peer group influence networks that may heighten—or protect against—the risk of substance abuse (Unger and Chen 1999) or shape decisions about contraceptive use (Valente et al. 1997); familial and friendship networks that may influence dietary practices, exercise habits, and other behaviors affecting the risk of obesity (Christakis and Fowler 2007) or of smoking (Christakis and Fowler 2008); sexual partnership networks that may raise or reduce the risk of contracting sexually transmitted diseases (Laumann and Youm 1999); and discussion networks among professional colleagues

that may influence treatment protocols or decisions to prescribe novel drug regimens (Coleman, Katz, and Menzel 1966).

Five principal mediating pathways through which social relationships may influence the health of individuals have been posited (Berkman and Glass 2000). Prominent among these is social support, which has emotional, instrumental, appraisal (assistance in decision making), and informational aspects (House and Kahn 1985). Beyond social support, networks may also offer access to tangible resources such as financial assistance or transportation. They can also convey social influence by defining norms about such health-related behaviors as smoking or diet, or via social controls promoting (for example) adherence to medication regimes (Marsden 2006). Networks are also channels through which certain communicable diseases, notably sexually transmitted ones, spread (Klovdahl 1985), and some contend that certain network structures reduce exposure to stressors (Haines and Hurlbert 1992).

Christakis (2004) has recently suggested that health interventions may have “collateral” effects, including not only the individual targeted by an intervention, but others in the target’s social network. Such direct and indirect (or multiplier) health effects of interventions could be of interest to both doctors and patients in the selection of treatments, and merit attention from policy makers and public health practitioners when assessing the value of interventions. Social network models of how individuals influence one another offer one approach toward measuring the presence and magnitude of such collateral health effects, and a path toward assessing the total effects of interventions.

Social network analysis measures relationships among social actors, assesses factors that shape their structure, and ascertains the extent to which they affect health-related outcomes. It is related to but distinct from analyses of the mechanisms through which social support affects health. Social support studies often assess only the aggregate receipt or availability of support, not necessarily configurations of ties among specific actors, while only some social network analyses focus on health outcomes—many take the network itself as the object of study.

Several social science disciplines, especially anthropology and sociology, have long engaged in social network analyses (Freeman 2004). Many techniques and descriptive measures of networks have developed there. More recently, interest in network analysis has risen among statisticians. Advances in computing power have made possible solutions to previously intractable problems, leading to a number of new models and methods for analyzing networks. In tandem with the recognition that networks are integral components of many research questions involving the co-existence and functioning of individuals, communities, policy domains, workplaces and schools, this has led to expanded application of social network analysis.

Two distinct types of network models are common. We shall refer to these here as individual- and relational-level models, respectively. In the first, the analysis focuses on an individual-level outcome, and the network data are used to define explanatory variables. The second type of application models the relationships between individuals in a network, in essence treating it as a multivariate dependent variable with individual linkages (or ties) as its elements. Such relational analyses account for network structure using both network statistics corresponding to regularities in relational properties (i.e. dependencies among network ties) and covariates such as characteristics of the units within the network. Thus, while individual-level models make inference about attributes of the individuals, relational-level models make inference about the ties linking the individuals. Individual-level models resemble standard regression models that seek to predict the distribution of some outcome measured on a focal individual or ego, but differ in that predictors may involve characteristics measured on other individuals (often known as “alters”) in a way that involves network structure, permitting hypothesis tests

about social influence. In relational models, the dependent variable measures an aspect of the relationship between individuals in the network and hypotheses of social selection are tested. In both types of problems, a major challenge is accounting for a complex correlation structure among outcomes that arises due to the network. If there are N individuals in a data set, this is of order N^2 in an individual-level analysis, but of order $N^2 \times N^2$ in a relational-level analysis.

The next section reviews some foundations of social network analysis, introduces a network we use as an illustration throughout the article, and describes how network data sets are represented numerically and visually. Section 3 introduces basic network statistics such as density and degree, some fundamental descriptive network measures including centrality indices, and approaches to the detection of subgroups within networks. These measures provide important insights into network typology, which can be used to develop models. Our review gives particular attention to recent developments in the statistical modeling of networks. Section 4 describes statistical models for individual-level outcomes that use network data to construct explanatory variables, while in Section 5 the focus is on relational-level network models. Section 6 calls attention to some recent innovations in social network methodology and Section 7 concludes.¹

2. Background

2.1. Defining Social Networks

A social network consists of one or more sets of units—also known as “nodes,” “actors,” or “vertices”—together with the relationships or social ties among them. The units or nodes are usually individual persons, e.g. patients or clinicians. They may, however, also be other social units (such as hospitals) or objects (such as texts). Relationships often represent communication, influence, trust or affect (e.g. friendship), but can also refer to conflict (e.g. disputes). Most social network studies also include attribute data describing the nodes/actors, the relationships, or both.

Certain subnetworks are often of interest. A pair of actors is known as a *dyad* and a triple as a *triad*. A *star* consists of an actor and all relationships incident to it. An *egocentric network* consists of an actor, the other actors in its immediate locality or neighborhood, and the relationships among them.

When—as is most typical—attention centers on relationships that link elements within a set of units/actors, a network is known as *one-mode*. Most of the discussion in this article pertains to the one-mode case. Networks may involve more than one set of units/actors, however. In particular, many studies involve two distinct types of units, such as patients and physicians, or physicians and hospitals. In these *two-mode* networks, the elementary relationships of interest usually refer to affiliations of units in one set with those in the other—e.g. of patients with the physician(s) responsible for their care, or of physicians with the hospital(s) at which they are admitted to practice. Hence two-mode networks are also known as *affiliation* networks.

While most network studies focus on a single relationship or type of tie observed on one occasion, both multirelational and longitudinal social network data exist. Multirelational data recognize the multistrandedness in many social ties; the relationship between two physicians, for example, may involve both professional collaboration and personal friendship. Longitudinal data permit the study of the creation, transformation, and dissolution of social ties. Most often, measured relations are binary-valued (present/absent), but they may also be ordinal or quantitative.

¹To aid readers in applying these methods, we provide some references to network software throughout, but our coverage of software is not comprehensive. Huisman and van Duijn (2005) review software resources available earlier in this decade.

2.2. Network Study Designs

Though a few network experiments have been conducted (e.g. Friedkin and Cook 1990, Travers and Milgram 1969), most social network data are observational. Studies typically measure networks using survey and questionnaire methods. Analysts also exploit data recorded in archives, including records maintained by electronic communication systems (Marsden 1990).

“Whole network” studies seek to assemble data on relationships in a theoretical population, that is, on the ties linking all units/actors within some bounded social collective, such as all physicians within a medical practice. In such studies, it is essential that clear boundaries or rules of inclusion for units/actors be specified (Laumann, Marsden, and Prensky 1983).

Statistical models such as exponential random graph models (see Section 5.3) are usually employed to analyze whole-network data (such as those on the physician network) that provide information on relationships among all units/actors within a closed population. Inferences therefore pertain to the model postulated as having generated those data, rather than to the design used to sample relationships for study from some larger network. Most applications of such methods examine networks of modest order—including between 10 and 50 actors—though analyses of much larger-order networks have been reported (e.g. Goodreau 2007).

2.3. Example: Influential Discussions among Physicians within a Primary Care Practice

A physician influence network in a primary care practice (Keating et al. 2007) will be used as an example throughout this article. The network was measured as part of a study examining how social networks influence physicians' beliefs and the use of therapies such as hormone replacement therapy (HRT). It exemplifies a one-mode, cross-sectional, whole-network study. The actors are physicians in the practice, and the relationships are influential discussions about women's health issues. Of 38 physicians, 33 responded to a survey, reporting the number of influential discussions about women's health issues (measured ordinally, as 0, 1–3, or 4+) they had with each other physician in the practice during the prior six months. Our illustrative analyses treat these data as binary-valued, distinguishing between reports of no discussions and of one or more discussions. The survey gathered attribute data for each physician, including vignette items measuring the propensity to recommend HRT, self-assessed areas of medical expertise, and the fraction of women in her/his panel of patients. Administrative records provided information on physician gender and number of clinical sessions per week.

We create two binary-valued versions of the physician influence network using these data. In the “directed” network, a relationship from physician i to physician j is said to be present if i cites j as a partner in one or more influential discussions. Such citations need not be reciprocated. In the “undirected” network, a relationship between i and j is present if either cites the other as someone with whom an influential discussion took place. Here, the relationship is either present or absent for each dyad, lacking directionality.

2.4. Representations of Networks

Two ways of representing networks are common (Freeman 1989): as matrices, and as graphs. In a matrix representation, rows and columns correspond to units/actors; the matrix is square for a one-mode network, and rectangular for a two-mode network. Multiple matrices are required for multirelational or longitudinal data. Cell entries contain the value of the relationship linking the corresponding units/actors, so that the ij^{th} cell represents the relationship from actor i to actor j . With binary-valued ties (1s indicating the presence of a tie), the matrix representation is known as an *adjacency matrix*. Table 1 displays the adjacency matrix for the first ten physicians in the directed women's health influence network. The network appears to be quite sparse, since there are many more 0s than 1s.

Networks are often represented using graphs in which actors/units are vertices and non-null relationships are lines. Undirected relationships are known as “edges” and directed ones as “arcs”; arrows at the end(s) of arcs denote their directionality. Graphical representations are often binary, but value-weighted graphs can also be constructed by displaying non-null tie values along arcs/edges, or by letting thinner and thicker lines represent line values. Such graphic imagery is a hallmark of social network analysis (Freeman 2004); early graphical depictions of networks were known as “sociograms.”

Graphs of networks are abstract in the sense that they do not have underlying coordinate axes. Many such drawings are ad hoc renderings, constructed using aesthetic criteria (e.g., minimizing the number of crossing lines). Algorithms including multidimensional scaling (Bartholomew et al. 2002) and spring embedders (Fruchterman and Reingold 1991) now are often used to position units/actors in Cartesian space by optimizing some function of the network data and the spatial coordinates for units. For example, the Fruchterman-Reingold algorithm locates units/actors such that those connected by an edge/arc are near—but (to avoid visual clutter) not *too* near—one another, while unconnected ones lie further apart. Graphs can be enhanced by letting the sizes, shapes, colors or labels of vertices represent differing values of attributes for units/actors.

Figure 1 displays the directed physician influence network, as rendered by the Fruchterman-Reingold algorithm programmed in the statnet software package (Handcock et al. 2003). Except for omission of the directional arrows, the graph for the corresponding undirected network is identical. The 33 actors (physicians) are labeled from 1 to 33. In general, physicians who often cite, or are cited by, others as influential conversation partners (such as physicians 21 and 27) tend to appear nearer the center of the graph.

Graphic representations of networks are visually appealing and evocative, but it is important not to over-interpret them. Plotted distances do not directly correspond to measured “social distances.”² Distinct, but formally equivalent, spatial arrangements based on the same network data can influence perceptions of structural characteristics (McGrath, Blythe, and Krackhardt 1997). The many vertices and lines in graphs of large, dense networks can render them unreadable. In general, graphs are most useful for identifying distinct regions or clusters within a network, distinguishing central and peripheral nodes, and revealing intermediary nodes that link distinct regions of the network. Careful analyses of networks usually focus on their mathematical and statistical features, however, as discussed in the sections that follow.

Network visualizations can be constructed using numerous software packages. Among these are the R package sna (Butts 2007), NetDraw (Borgatti 2008), and Pajek (Batagelj and Mrvar 2003).

3. Descriptive Properties of Networks

Analysis of network data often begins by examining both actor- and network-level descriptive statistics and measures. This section reviews many of the most common of these. Wasserman and Faust (1994) provide a comprehensive introduction to descriptive network measures.

²The extent to which distances in a graphical representation correspond to the data on which they rest—dyadic measurements of social distance or proximity—depends on the objective function that serves as a fitting criterion. when the plot is constructed. The most widely-used “nonmetric” multidimensional scaling algorithm requires an ordinal correspondence between data and plotted distances; “metric” scaling uses a stronger (linear) criterion. Objective functions used by many spring-embedder methods involve a “node repulsion” term that simplifies the visual representation by discouraging co-location of vertices within a plot, but simultaneously limits the extent to which data and plotted distances correspond. Moreover, a low (ordinarily 2)-dimensional Cartesian plot may do more or less well in representing data on the relationships among N actors, which may in principle be $(N - 1)$ -dimensional.

We use the symbol y_{ij} to refer to a network variable recording data on the relationship from actor i to actor j . A matrix \mathbf{y} includes all such variables. In many applications y_{ij} is binary-valued, taking the value 1 if i is tied to j and 0 otherwise; in this article, we take y_{ij} to be binary-valued unless indicated. Self-relations y_{ij} are usually undefined. When relations are binary-valued, \mathbf{y} is the adjacency matrix.

3.1. Size and Density

Perhaps the very simplest property of a network is its number of units/actors (N), known as its *order*. For binary-valued networks, the corresponding relation-level statistic is the number of ties, known as *size* ($L = \sum_{i,j} y_{ij}$). A widely cited statistic is network *density*, defined as size relative to the number of possible ties and equal to $L/(N(N-1))$ for directed networks. More generally, for quantitative data on relations, density could be defined as the mean strength of a tie.

The 33-physician influence network is of order 33. The directed influence network has size 163, i.e., 163 nonzero ties were observed. Since $33 \times 32 = 1056$ ties were possible, network density is 0.154.

3.2 Degree and the Degree Distribution

In an undirected network, an actor's *degree* is the number of other actors to which it is directly connected. Analyses of directed networks distinguish between incoming and outgoing ties. The number of arcs oriented toward an actor is that actor's *in-degree* ($y_{+j} = \sum_i y_{ij}$), sometimes termed popularity or attractiveness; the number of arcs emanating from an actor is its *out-degree* ($y_{i+} = \sum_j y_{ij}$), also known as expansiveness. Often actors having greater degrees have prominent roles in the network; indeed, the simplest measures of centrality (Section 3.6) are based on degree (Freeman 1979).

The *degree distribution* is the frequency distribution giving the number of actors having particular numerical degrees. Its variance measures the extent to which direct connectedness varies across actors (Snijders 1981). Barabási and colleagues have focused on degree as their fundamental analytic interest (Barabási 2002; Wolfram 2002), showing that many network properties are shaped by the degree distribution. As the examples in Figure 2 illustrate, networks with the same overall density but different degree distributions may have quite different structures. A "circle" network—in which actor degree is constant (and hence, degree variance is 0)—and a "star" network—in which one actor has degree $N-1$ while all others have degree 1—lie at opposite ends of the spectrum with respect to degree variation.

Histograms of the degree distributions for the directed physician influence network are shown in Figure 3. The out-degree distribution is more uniform than the in-degree distribution, which is markedly skewed toward the right. The standard-deviation among in-degrees is 5.20, while that among out-degrees is only 3.29. Many physicians are rarely cited by others as influential discussion partners, while one has an in-degree of 24. A list of actor-level network statistics (Table 2) shows that physician 27 has out-degree 24 but in-degree of only 2. This physician directly influences most of the other physicians but is influenced by few of them. Three physicians do not influence others (have indegree 0) while two are not influenced by any others (have outdegree 0).

3.3 Paths and Geodesic Distance

Actors in networks are connected to one another indirectly via intermediaries as well as directly. Nonzero ties in the adjacency matrix give direct connections. An indirect connection is present when one or more multi-step *paths* exist from one actor to a second, in which case the latter is said to be *reachable* from the former. A length-2 path from actor i to actor j exists when there

is a third actor h such that i is adjacent to h and h in turn is adjacent to j . Paths may involve multiple adjacencies; the *length* of a path is the number of relationships or lines it contains. A *geodesic* path is a shortest-length path between a given pair of actors. *Geodesic distance*—the length of a geodesic path—is perhaps the most widely-used network-based measure of the social distance separating units/actors.

Matrix multiplication of an adjacency matrix \mathbf{y} by itself yields the number of paths of a given length between any two actors. For example, the ij^{th} element of \mathbf{y}^k contains the number of paths of length k from actor i to actor j . The geodesic distance from i to j is given by the smallest positive integer k for which the ij^{th} entry in \mathbf{y}^k is nonzero. If no path from i to j exists, the geodesic distance from i to j is said to be infinite. In directed networks, the geodesic distance from i to j need not equal that from j to i .

For the directed physician network, the number of paths of length 2 and length 3 that begin and end with actors 1 through 10 are displayed in Tables 3 and 4, respectively. There is a rapid decrease in the number of nonzero cells and a rise in the number of distance- k paths connecting most pairs of physicians as path length (k) increases. Some directly-connected pairs of physicians are not indirectly linked, however. For example, there is a direct tie from physician 2 to physician 3 (Table 1), but no path of length 2 or 3, indicating that no sequence of ties via one or two intermediary physicians leads from physician 2 to physician 3.

Table 5 displays the geodesic distances between actors 1–10 in the directed physician network. Values of -1 indicate that no path of any length exists from one physician to the other, i.e. that the geodesic distance between them is “infinite.” For example, there is no path to physician 2 from any of physicians 1–10. Similarly, physician 10 cannot reach physicians 1–9. The longest geodesic distance shown is 5, from physician 6 to 10 and from physician 7 to 10.³ The geodesic distances are not symmetric; for example, from physician 1 to 4 the distance is 3, but from 4 to 1 it is 2.

3.4. The dyad census and reciprocity

In binary-valued directed networks, three types of dyadic relationships may exist: *mutual* dyads, in which a tie from i to j is accompanied by one from j to i ; *asymmetric* dyads in which there is a relationship between i and j in one direction, but not the other; and *null* dyads in which there is no tie in either direction. The *dyad census* is the set of three network statistics giving the number of each dyad type found within a given network; for example, the number of mutual ties is $M = \sum_{i < j} y_{ij} y_{ji}$.

If all ties in a binary network are either mutual or null, the network is said to be *symmetric*, in which case the adjacency matrix \mathbf{y} and its transpose \mathbf{y}^T are identical; an undirected network is symmetric by construction. The presence and magnitude of a tendency toward symmetry or reciprocity in a directed network can be measured by comparing the number of mutual dyads to the number expected under a model in which ties are reciprocated at random. If the number of mutuals is lower than expected, there is a tendency away from reciprocation.

The dyad census for the directed physician network includes 26 mutual dyads (encompassing 52 directed ties), 111 asymmetric dyads, and 391 null dyads. The distribution of the number of mutual ties across physicians is right-skewed (Figure 3), with mean 1.45 and standard deviation 1.68. Because the proportion of mutuals among non-null dyads, $26 / (26 + 111) = 0.190$, exceeds the network density (which estimates the probability that any tie is present in a purely random network), 0.154, there appears to be a tendency toward reciprocity in this network. A formal test requires information about the distribution of the expected number of

³Note that some or all of the intermediaries along these geodesic paths may be physicians 11–33.

mutuals under the hypothesis of random reciprocation; more powerful tests condition on observed features of the network such as the degree distribution (Holland and Leinhardt 1981; Snijders 1991). Tests for reciprocity may also be conducted using regression models that control for other network effects; Section 5 presents a regression-based test for reciprocity in the physician influence network.

3.5. The Triad Census, Transitivity, and Closure

Triads in undirected, binary networks may include 0, 1, 2, or 3 relationships. Triads having 3 relationships are said to be *closed* or *transitive*, in that each pair of units/actors linked by a direct tie is also linked by an indirect path through the third unit/actor. For directed binary networks, 16 distinct triad types exist, distinguished by the number and orientation of the directed ties they include (Wasserman and Faust 1994). Of these, triad types including transitive substructures—in which the presence of a direct tie from i to j is accompanied by the presence of an indirect path from i to j via h —are indicative of network closure. The *triad census* is the set of network statistics giving the number of triads of each possible type in an observed network. The triad census is related to the mean and variance of the degree distribution, and holds strong implications for overall network structure, especially for networks of low order (Frank 1981).

One indication of transitivity can be obtained by considering the subset of triads in which one actor is connected to both of the others, and comparing the proportion of such triads that are closed (i.e., the proportion in which j and k are connected, given that both j and k are connected to i) to the network density (which estimates the probability that such a triad is closed in a fully random network). The undirected physician influence network⁴ includes 1429 triads in which one actor has ties to the two others, and 624 transitive triads. The proportion of transitive triads among those including at least two ties is 0.437 which far exceeds the undirected network density of 0.256, implying that such triads occur more often than expected by chance. This comparison is not a formal statistical test, however, as it does not consider the distribution of the number of closed and non-closed triads, nor does it condition on the dyad census, the degree distribution, or other network statistics. The papers referenced earlier describe formal tests for transitivity while Section 5 tests for it within a regression framework.

3.6. Centrality

Measures of centrality reflect the prominence of actors/units within a network. They are among the most widely-used actor-level measures that derive from network data.

Distinct centrality measures (Freeman 1979; Wasserman and Faust 1994) are sensitive to different aspects of an actor's network location. The simplest is based on an actor's degree. Separate in-degree and out-degree centrality measures exist for directed networks. Degree-based centrality reflects an actor's level of network activity or involvement. A second common measure rests on *betweenness*: the frequency with which an actor is found in an intermediary position along the geodesic paths linking pairs of other actors. In networks of communication or exchange, actors with high betweenness centrality have high capacity to broker or control relationships among other actors. A third major centrality measure, *closeness*, is based on the sum of geodesic distances from a given actor to all others; closeness-based network prominence is inversely proportional to this sum. Actors linked to others via short geodesics have comparatively little need for intermediary (broker) units, and hence have relative independence in managing their relationships. Closeness measures are defined only for networks in which all actors are mutually related to one another by paths of finite geodesic distance; this condition holds for the undirected physician network but not the directed one.

⁴Recall that the undirected physician network is identical to that shown in Figure 1, except that ties lack directionality.

Another centrality index is sensitive to the presence and/or strength of connections, as well as the centrality of those actors to which a focal actor is linked. It assumes that connections to central actors indicate greater prominence than do similar-strength connections to peripheral actors. Measures based on this conceptualization involve the eigenvector corresponding to the largest eigenvalue of a matrix representation of a network, and hence are known as *eigenvector* centrality measures (e.g. Bonacich 1987). The different centrality measures are often—though not always—well-correlated, but embody different aspects of network prominence.

The centrality measures for the directed physician network (Table 2) show that physician 19 claims to be influenced by the most others, having the largest outdegree (13), while physician 27 (with the largest indegree of 24) influences the most others.⁵ The betweenness measure reported in Table 2 has been scaled so that scores indicate betweenness as a percentage of its maximum possible level. Physicians 9 (scaled betweenness 27.6) and 27 (scaled betweenness 20.1) are most central by this definition. We do not report closeness scores for the directed network, because not all pairs of physicians are linked by finite geodesic paths.

The closeness scores for the undirected network (Table 6) are standardized for the number of physicians in the network (Beauchamp 1965); values lie between 0 and 1 with higher values reflecting greater centrality. Physician 27 (closeness 0.8) is by far the most central actor according to this measure and physicians 14, 20, and 25 are the least. Note that physician 25 is less central in terms of closeness than physician 26 although physician 25 has higher degree. Eigenvector centralities for this network⁶ also show physician 27 (0.35) to be most central, followed by physician 9 (0.29); these are scaled here such that each actor's eigenvector centrality equals the corresponding element of the first eigenvector of the adjacency matrix.

Centrality measures are often taken as indicators of an actor's network-based "structural power"; the suitability of such an interpretation depends, of course, on the substance of any particular application. Such measures are often used as explanatory variables in individual-level regression models, but such applications do not always fully account for interdependencies among the actors in whole-network data sets.

Centralization indices (Freeman 1979) are network-level statistics that resemble the degree variance, growing larger to the extent that all relationships involve a single actor (as in the "star" network shown in Figure 2).

3.7. Cliques, Components and Clusters

Descriptive analyses often use network data to assign actors to subgroups, reasoning that certain patterns in relationships reveal salient social distinctions. Often this involves a search for locally dense regions within a network, that is, subsets of actors that have strong relationships with one another. For binary-valued networks, an idealized model of such a solidary subgroup is the *clique*, a maximal subset of actors having density 1.0. This subgroup density requirement is very stringent, and analyses of empirical network data rarely find cliques of appreciable size. Other approaches to identifying cohesive subgroups relax that standard in various ways (Wasserman and Faust 1994).

Components are a much weaker subgroup concept. In a directed network, *strong components* are subsets of actors mutually linked to one another by paths of finite length. Strong components partition the actors in a network into mutually exclusive and exhaustive subsets,

⁵We calculated centrality scores using the software package UCINET 6 (Borgatti, Everett, and Freeman, 2002).

⁶Eigenvalue centrality can in principle be calculated for a nonsymmetric matrix, but the routine in UCINET 6 handles only the symmetric case.

which themselves are partially ordered. Weak components are defined similarly, except that the directionality of relationships is ignored when assessing whether two actors are connected; by construction, weak components are isolated from one another. Many networks consist of one large component, sometimes together with several smaller ones and singleton actors. A Colorado Springs study of persons at risk for HIV (Rothenberg et al. 1998) documented an overtime decline in the size of components in networks of risky (sexual, drug-using, needle-sharing) ties, connecting this to a fall in personal risk-taking.

Still another cohesion concept is the *k-connected component* (White and Harary 2001): a maximal subset of actors mutually linked to one another by at least *k* node-independent paths (i.e., paths that involve disjoint sets of intermediary actors who also lie within the subset). This notion emphasizes robustness of connections among the elements within subgroups. The mapping of actors to *k*-connected components is not mutually exclusive, and *k*-components for higher *k* are nested within those for lower *k*.

The physician network appears to be relatively cohesive. In the directed network, one main strong component includes 27 of the 33 actors. The remaining six are singletons: four of them cite physicians in the main component but receive no citations from it, while the other two are cited by main-component physicians but do not cite anyone. The undirected network consists of a single connected (weak) component. Indeed, the entire undirected network is a *bi-component*, since all pairs of physicians are connected via at least two node-independent paths. The undirected network is centered on a 7-physician clique (physicians 9, 16, 19, 21, 24, 27, and 33 in Figure 1), which is part of a 6-connected component that includes 17 physicians.

3.8. Homophily

The tendency for relationships to form between people having similar attributes is known as *homophily* (McPherson, Smith-Lovin, and Cook 2001). Homophily involves three-way statistical interactions between actor attributes and the presence of relationships, or equivalently, subgroup-specific network density statistics. With high homophily according to some attribute, networks tend toward segregation by that attribute, contributing to network closure.

Empirical studies in the network literature often report tendencies toward homophily. In their analysis of the physician network, Keating et al. (2007) documented strong tendencies toward homophily by organizational location: influential discussions tended to be held with others in a physician's clinic (subpractice) within the practice. By contrast, they found a weak and insignificant tendency toward homophily by gender.

3.9. Descriptive Properties for Egocentric Networks

Numerous properties of network structure in an actor's locality can be measured using data on that actor's egocentric network (Marsden 1987). Two very common ones are the actor's degree (often termed egocentric network size) and local network density—the extent of connectedness among the pairs of alters within a given egocentric network. High local density indicates closure within the neighborhood surrounding an actor. A betweenness centrality measure for egocentric network data is available (Marsden 2002). Actor-specific statistics summarizing the distribution of alter characteristics in an egocentric network—such as the mean and standard deviation of the ages of alters—measure network composition and heterogeneity. Burt (1992) presents a refined set of indices that measure egocentric network closure.

Such properties of egocentric networks can be derived from whole-network data, and can also be based on egocentric network data obtained within representative sample surveys. Once constructed, such indices are often used as explanatory variables in regression analyses that

seek to explain variations in some individual-level outcome such as well-being, or as dependent variables in analyses concerned with determinants of local structure. With egocentric network measures based on whole-network data (as in the physician network), however, analyses should recognize the complex pattern of interdependence between egocentric networks due to the clustering of actors. However, no special analytic problems arise when such data are assembled within sample surveys—since the alter actors in an egocentric network usually are not among the ego units sampled, so assuming independence among observations on different egocentric networks is reasonable.

Hierarchical models are often used to analyze egocentric data since observations are grouped by ego. This model uses both between-ego covariates and within-ego covariates measured on the alters or on ego-alter ties (Van Duijn, Van Busschback, and Snijders 1999). Due to the independence between egos, multilevel models whose within-ego covariance matrix captures the association between alters tied to each ego and among themselves can be applied to such data. Wellman and Frank (2001) provide an example of hierarchical modeling of this type of data in the context of social network capital. Standard hierarchical models cannot be used to analyze the data structures described in Sections 4 and 5.

3.10 Software for Descriptive Network Analysis

Most current network software is found in stand-alone programs rather than integrated software packages such as SAS or Stata, though such packages can construct many measures for egocentric network data. UCINET 6 (Borgatti et al. 2002) is relatively comprehensive and widely used in managing network data and conducting descriptive analyses. The R package sna (Butts 2007) likewise can perform most analyses discussed in this section. See Huisman and Van Duijn (2005) for other, often more specialized, network software.

4. Individual outcome regression models

Individual-outcome regression models are, as usual, primarily concerned with how the distribution of some dependent variable (e.g. an attitude or opinion) measured on a focal actor is related to one or more explanatory variables. When such attitudes or opinions are formed in part as the result of interpersonal influence, the outcomes for actors are not statistically independent as assumed by many regression models. Instead, the outcome for one actor will be related to those for the other actors who influence her or him, leading to a complex correlation structure. In theory each actor might directly or indirectly influence each other actor. Individual-outcome analyses use network data to model this correlation structure. Networks may enter through either the construction of explanatory variables or the modeling of covariances among errors.

Let \mathbf{Z} be a vector containing measures of an outcome on the N actors in a network, \mathbf{X} be a matrix whose i^{th} row contains a vector of exogenous predictor variables (e.g. gender) for the i^{th} actor, and \mathbf{W} be an $N \times N$ matrix whose elements W_{ij} measure the extent to which actor i is influenced by actor j , larger values indicating greater influence. In individual-outcome analyses, the covariates \mathbf{X} typically measure attributes of individual actors. These may include actor-level network statistics such as a focal actor's degree, centrality, or local density (see Section 3). If an analysis is based on data for actors within multiple, disjoint networks, network-level statistics such as global density or network centralization could vary among the actors and thus be used as predictors. Elements of \mathbf{W} are measured via some function of the network data (e.g., adjacency, tie strength, or inverse geodesic distance); ordinarily diagonal terms W_{ii} are set to 0. Typically, the rows of \mathbf{W} are scaled to sum to 1, so that W_{ij} is interpretable as a measure of the relative influence of j on i .

Network-related interdependence among the outcomes \mathbf{Z} may be incorporated in two distinct ways. First, one actor's outcome may depend directly on the outcomes of the alters to whom s/he is linked. The vector $\mathbf{Z}_W = \mathbf{W}\mathbf{Z}$ contains, for each focal actor, the (weighted) average value of the outcome measure for those other actors to whom that actor is linked by nonzero influence W_{ij} ; as such, outcomes for other actors contribute to \mathbf{Z}_W in proportion to their influence on the ego. Thus, \mathbf{Z}_W is a network-lagged outcome. For the special case in which \mathbf{W} is a scaled adjacency matrix (i.e. $W_{ij} = y_{i+}^{-1}$ if $y_{ij} = 1$ and 0 otherwise, where y_{i+} is the outdegree of—i.e., number of actors that influence—actor i), \mathbf{Z}_W is a vector whose i^{th} element is the unweighted average value of the outcome for alters in actor i 's egocentric network.

An *autoregressive outcome model* accounts for interdependence between outcomes by directly including \mathbf{Z}_W as a predictor. Such a regression model is

$$\mathbf{Z} = \alpha \mathbf{W}\mathbf{Z} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\boldsymbol{\varepsilon}$ denotes a vector of stochastic errors, here taken to be independent of one another, parameter α measures the magnitude of the network effect, and $\boldsymbol{\beta}$ is a vector of regression parameters.

Alternately, the errors $\boldsymbol{\varepsilon}$, rather than the outcomes \mathbf{Z} themselves, may be interdependent. Such *network autocorrelation* can be modeled via inclusion of a term $\bar{\boldsymbol{\varepsilon}}_W = \mathbf{W}\boldsymbol{\varepsilon}$ in specifying the distribution of the error term. The vector $\bar{\boldsymbol{\varepsilon}}_W$ contains, for each focal actor, the (weighted) average stochastic error for those other actors to whom that actor is linked by nonzero influence W_{ij} , again in proportion to their network-based influence on the ego. The relationship between $\bar{\boldsymbol{\varepsilon}}_W$ and $\boldsymbol{\varepsilon}$ is a second-order effect reflecting a component of correlation among elements of \mathbf{Z} due to unobserved factors. Observe that under the common assumption that the errors $\boldsymbol{\varepsilon}$ are stochastically independent of the explanatory variables \mathbf{X} , the network autocorrelation term $\bar{\boldsymbol{\varepsilon}}_W$ is likewise independent of \mathbf{X} , while the network lagged term \mathbf{Z}_W will in general be correlated with \mathbf{X} . A regression model incorporating $\bar{\boldsymbol{\varepsilon}}_W$ may be written as

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\varepsilon} = \rho \mathbf{W}\boldsymbol{\varepsilon} + \boldsymbol{\nu}, \quad (2)$$

where $\boldsymbol{\nu}$ is a vector of independent random perturbations and parameter ρ measures the strength of the network autocorrelation. The implied mean vector and covariance matrix of $\boldsymbol{\varepsilon}$ are $\mathbf{0}$ and $\text{var}(\boldsymbol{\nu})\{(\mathbf{I} - \rho \mathbf{W}^T)(\mathbf{I} - \rho \mathbf{W})\}^{-1}$ respectively. Model (2) may be rewritten as:

$$\mathbf{Z} = \rho \mathbf{W}\mathbf{Z} + \mathbf{X}\boldsymbol{\beta} - \rho \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu}. \quad (3)$$

Equation (3) reveals that model (2) differs from (1) only by the addition of the network-lagged covariate term $\rho \mathbf{W}\mathbf{X}\boldsymbol{\beta}$, which measures the effect of other actors' covariates on the outcome for an actor. Because the network-lagged outcomes and covariates have equal (though opposite) effects under model (2), model (1) is not nested in model (2). If, however, model (2) were extended by allowing different coefficients for the autocorrelation terms for the lagged outcomes \mathbf{Z} and the covariates \mathbf{X} , then both (1) and (2) would be special cases of that more general model; see Friedkin (1990) for an example.

Individual outcome models may also be specified using both \mathbf{Z}_W and $\bar{\boldsymbol{\varepsilon}}_W$. The following regression model contains both autoregressive outcomes and network autocorrelation (Anselin 1988; Burt and Doreian 1982), allowing for different weight matrices for the two:

$$\mathbf{Z} = \alpha \mathbf{W}_1 \mathbf{Z} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} = \rho \mathbf{W}_2 \boldsymbol{\varepsilon} + \mathbf{v}, \quad (4)$$

where \mathbf{W}_1 and \mathbf{W}_2 are the weight matrices for the spatially lagged network effects and network autocorrelation effects, respectively. This model includes two sources of correlation in \mathbf{Z} and one source of correlation in $\mathbf{X}\boldsymbol{\beta}$.

Several authors in the network literature (e.g., Doreian 1980, Dow 1984, Doreian 1989, Friedkin 1990) have introduced models (1) and (2), which are related to models used to account for autocorrelation in spatial data analysis. Model (2) is commonly known as a simultaneously autoregressive (SAR) model (Banerjee, Carlin, and Gelfand 2004; Waller and Gotway 2004). In purely spatial contexts, an alternative to the SAR model known as the conditional autoregressive (CAR) model is often used (Waller and Gotway 2004). The CAR model specifies the conditional probability distribution of each Z_i given all components of \mathbf{Z} other than Z_i and then uses the Hammersley-Clifford Theorem (Besag 1974) to derive the joint distribution of \mathbf{Z} , whereas the SAR model and the autoregressive outcome model in (1) specify the joint distribution of the error term $\boldsymbol{\varepsilon}$ and then induce the joint distribution of \mathbf{Z} . However, the CAR model has to date not been used in social network analysis as much as the SAR model or variants thereof.

Ordinary least squares (OLS) techniques are not suitable for estimating models (1), (2), and (4). OLS is inconsistent in the case of models (1) and (4) because \mathbf{Z} appears on both sides of the equation. In model (2), or equivalently (3), OLS is inefficient because the covariance matrix of $\boldsymbol{\varepsilon}$ is not diagonal. These models can be estimated by generalized least-squares or maximum likelihood (Waller and Gotway 2004) or instrumental variables (i.e., moment-based) methods (Anselin 1988, 1990; Land and Deane 1992). Deciding how to use network data to construct the weight matrix (or matrices) is an important step in the application of these models (Leenders 2002).

4.1. Illustrative Analysis

To illustrate the use of individual-outcome models for the physician network, we examine possible network effects on a physician's propensity to recommend HRT; denoted RecHRT, this is a summary score that averages responses to several vignette items. We hypothesized that RecHRT would increase among physicians strongly tied to others with high propensity to recommend HRT. In the autoregressive outcome model (1), the key explanatory variable (denoted AltHRT) is the average value of RecHRT among the other physicians linked to each focal physician through influential conversation ties. We constructed AltHRT and tested for network autocorrelation using two different versions of the weight matrix \mathbf{W} , one based on direct network adjacencies, the other on scaled inverse geodesic distances.⁷ Physician gender, percentage of women in a physician's patient panel, and the focal physician's outdegree serve as additional covariates.

We fit both the autoregressive outcome model and the network autocorrelation model using each of the weight matrices. For a given \mathbf{W} , the autoregressive outcome model is

$$\text{RecHRT}_i = \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{PcntWomen}_i + \beta_3 \text{OutDegree}_i + \alpha \text{AltHRT}_i + \varepsilon_i, \quad (5)$$

⁷Because two actors have outdegrees of 0, the associated rows of \mathbf{W} sum to 0 as opposed to 1. Therefore, although these actors contribute to the estimation of β and σ^2 , they do not directly contribute any information about the autocorrelation parameters α and ρ . We retained these actors in the analysis because they were cited by other physicians as influencing them and so removing them would omit information about how other actors were influenced.

where $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, and the network autocorrelation model is

$$\text{RecHRT}_i = \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{PcntWomen}_i + \beta_3 \text{OutDegree}_i + \varepsilon_i, \quad (6)$$

where $\varepsilon \sim N[\mathbf{0}, \sigma^2 \{(\mathbf{I} - \rho \mathbf{W}^T)(\mathbf{I} - \rho \mathbf{W})\}^{-1}]$.

The models in (5) and (6) may be fit by directly maximizing the respective likelihood functions of the data. When constructed using the directed network data, \mathbf{W} is asymmetric even before its rows are standardized to sum to 1. This required extending the expressions for the usual asymptotic covariance matrices of the maximum likelihood estimator of the model parameters (see Doreian (1981) and Waller and Gotway (2004), in the case of the autoregressive outcome and network autocorrelation (SAR) models respectively) to accommodate asymmetric \mathbf{W} .

The estimates for model 5A and model 5B in Table 7 suggest that AltHRT has a modest positive effect on a physician's propensity to recommend HRT. However, because the p-value for the effect of AltHRT on RecHRT is well above 0.05, further study is required before a firm conclusion can be drawn. The estimated effect size is roughly the same for the two versions of the weight matrix \mathbf{W} . The estimates for models 6A and 6B in Table 7 suggest that the residual network autocorrelation is weaker than the direct effect of AltHRT on RecHRT. Outdegree has a moderate negative coefficient in all models, suggesting that focal physicians influenced by a greater number of other physicians might be less likely to recommend the use of HRT.

The extent to which these results can be extended to other physicians and clinics depends on the similarity of the physicians, their clinics, and the extent to which differences (e.g., due to clinic characteristics or environments) affect physician behavior. If the data generating process is the same as, or at least is exchangeable with, that which generated the physician influence network, then the inferences will have relevance. However, there is no way of knowing the similarity of the clinics and their physicians without conducting a study that draws data from multiple practices (e.g. a cluster design).

4.2. Software for Individual-Outcome Analyses

Although the models fitted in Section 4.1 are non-standard in the sense that the covariance structure is a function of an unknown parameter, we found it easy (and instructive) to write our own R procedures to fit them (see Appendix). Alternatively, the *Inam* procedure in the *sna* package (Butts 2007) in R may be used to fit the autoregressive outcomes and the network autocorrelation models (see Appendix) as well as models containing both terms. Some models can also be estimated using existing software available for spatial analysis. For example, the *S+SpatialStats* package in *SPlus* can be used to fit SAR and CAR models, and the *GeoBUGS* package in *WinBUGS* will fit models with CAR terms. When applying existing packages to network data, or developing one's own code, appropriate care must be taken to accommodate the asymmetric weight matrices that commonly arise with network data.

5. Relational or dyad-level models

Relational analyses of network topology model the relationships in a social network simultaneously, recognizing the interdependencies among them. They posit that global network properties are the result of a set of localized regularities that create correlations involving subsets of network ties, e.g. within actors, dyads, triads, or tetrads (Robins, Pattison, and Woolcock 2005). Examples of such regularities are actor-level tendencies to produce and/or attract ties, dyadic tendencies toward reciprocity, and triadic tendencies toward closure or transitivity. Relational models may also incorporate attribute data on actors or relationships.

For instance, certain types of actors may tend to attract ties, actors having the same or similar attributes may tend to be linked (homophily), or actors linked in one network may also tend to be related in a second.

A relational model, in essence, specifies a set of micro-level rules governing the local structure of a network. When applied to relationships among an entire set of actors, such rules could generate many random realizations. A successful model for an observed network should produce realizations with typical properties that match the corresponding observed properties. Hence, the capacity to reproduce observed network properties—especially properties that are not explicitly modeled—signals that a model fits well. If a model does not capture a given feature of an empirical network, it surely omits some consequential rule governing network formation.

Since the 1930s, a variety of statistical methods have been used to analyze social network data (Wasserman and Faust 1994). Early models generally relied on conditionally uniform null distributions, positing that an observed network is drawn from a set of possible networks known to have particular features. Initially, models tested for reciprocity and transitivity, conditioning on lower-order network statistics. For example, Katz and Powell (1957) derived the distribution of the dyad census for a directed network given the distribution of outdegrees. Holland and Leinhardt (1976) proposed tests for transitivity (and other properties reflected in linear combinations of counts in the triad census) against a null model asserting that the distribution of networks is conditionally uniform given the dyad census. Few such distributions are analytically tractable, however. The probability mass function for a uniform distribution of networks given both the outdegrees and indegrees, for instance, cannot be written down, though it would clearly be desirable to condition on both when testing for reciprocity.

Computing power now allows enumeration (for networks of small order, say $N < 10$) or simulation of networks from heretofore intractable distributions (Snijders 1991), permitting nonparametric tests for certain network properties. To illustrate, we simulated 10,000 random binary-valued networks having the in- and outdegree distributions shown in Table 2 for the directed physician network. In these simulated networks, the mean number of mutual dyads was 15, with a maximum of 24 and a 99th percentile of 20. Since the dyad census for the actual physician network includes 26 mutual dyads, its level of reciprocity appears to be quite unusual given its degree distributions.⁸

Beginning in the 1970s and accelerating in the past decade, statisticians have formulated new parametric statistical models for relational data that can incorporate multiple network properties, as well as attribute data. The forthcoming sections review such models. We begin by defining notation. The binary random variable $Y_{ij} = 1$ if there is a network tie from actor i to actor j and $Y_{ij} = 0$ otherwise. An adjacency matrix \mathbf{Y} includes all such variables. Lower-case letters, y_{ij} and \mathbf{y} , respectively, denote realizations of these variables.

5.1. Fixed-Effect Dyad Independence Models

Some statistical models for entire networks are equivalent to models for individual ties Y_{ij} or dyads (Y_{ij}, Y_{ji}) . They emphasize degree distributions and reciprocity as features shaping network structure. Among the first statistical models to be formulated for network data, such models are comparatively simple to estimate and interpret. They specify that network variables in different dyads are conditionally independent given covariates, so the likelihood function for an observed network is the product of the probability distributions for dyads. Hence, these models can be estimated using regression techniques with ties or dyads as cases.

⁸Computations were performed using the StOCNET software package (Boer et al., 2006).

One of the simplest models corresponds to the independence digraph (Erdős and Rényi 1959) in which the presence of each possible tie is independent with $Y_{ij} \sim \text{Bernoulli}(p_{ij})$, where $\mu_{ij} = \log(p_{ij})$ denotes the logarithm of the probability of a tie from i to j . Enforcing a homogeneity assumption $\mu_{ij} = \mu$ for all i and j simplifies this to a single-parameter model, under which the probability distribution of possible networks

$$\Pr(\mathbf{Y}=\mathbf{y};\mu)=\exp(\mu t_1(\mathbf{y}))(1 - \exp(\mu))^{N(N-1)-t_1(\mathbf{y})}$$

depends only on the network statistic $t_1(\mathbf{y}) = \sum_{i,j} y_{ij}$, the total number of ties.

More general models for directed graphs specify that dyads, rather than ties, are independent. This allows the pair of ties within a dyad (Y_{ij}, Y_{ji}) to be correlated (positively so, in the case of reciprocity). Such models usually allow for correlations among ties having a source (Y_{ij}, Y_{ik}), $j \neq k$ or target (Y_{ij}, Y_{hj}), $i \neq h$ in common by introducing “sender” effects α_i and “receiver” effects γ_i , thereby fitting the degree distributions.

As a dyad has four possible states, a four-component multinomial distribution serves as the basis of a model, taking the pair of arcs in a dyad (Y_{ij}, Y_{ji}) as an independent multinomial random variable with

$$\begin{aligned}\Pr(Y_{ij}=0, Y_{ji}=0) &= \kappa_{ij}^{-1}(\boldsymbol{\theta}) \\ \Pr(Y_{ij}=1, Y_{ji}=0) &= \kappa_{ij}^{-1}(\boldsymbol{\theta}) \exp(\mu_{ij} + \alpha_i + \gamma_j) \\ \Pr(Y_{ij}=0, Y_{ji}=1) &= \kappa_{ij}^{-1}(\boldsymbol{\theta}) \exp(\mu_{ji} + \alpha_j + \gamma_i) \\ \Pr(Y_{ij}=1, Y_{ji}=1) &= \kappa_{ij}^{-1}(\boldsymbol{\theta}) \exp(\mu_{ij} + \alpha_i + \gamma_j + \mu_{ji} + \alpha_j + \gamma_i + \rho_{ij}),\end{aligned}\tag{7}$$

where $\kappa_{ij}(\boldsymbol{\theta}) = 1 + \exp(\mu_{ij} + \alpha_i + \gamma_j) + \exp(\mu_{ji} + \alpha_j + \gamma_i) + \exp(\mu_{ij} + \alpha_i + \gamma_j + \mu_{ji} + \alpha_j + \gamma_i + \rho_{ij})$ is a normalizing constant and $\boldsymbol{\theta}$ is a vector containing all model parameters. For the physician influence network, parameter μ_{ij} is a constant term reflecting the overall probability that physician i reports an influential conversation with physician j (i.e., network density), the sender effect α_i reflects the propensity for physician i to be influenced by others, the receiver effect γ_j reflects the propensity for physician j to influence others, the reciprocity parameter ρ_{ij} accounts for within-dyad dependence, and $\kappa_{ij}(\boldsymbol{\theta}) = \kappa_{ji}(\boldsymbol{\theta})$ is a normalizing constant. Model (7) is fully saturated; ordinarily it is simplified by enforcing homogeneity conditions on μ_{ij} and ρ_{ij} .

Holland and Leinhardt (1981) introduced the p_1 probability density including the homogeneity conditions $\mu_{ij} = \mu$ and $\rho_{ij} = \rho$ for all i and j , and treating the sets of parameters $\{\alpha_i\}$ and $\{\gamma_j\}$ as fixed effects. This leads to the probability density function

$$p_1(\mathbf{y}) = \Pr(\mathbf{Y}=\mathbf{y}) = \exp \left\{ \mu t_1(\mathbf{y}) + \sum_i \alpha_i t_{2i}(\mathbf{y}) + \sum_j \gamma_j t_{3j}(\mathbf{y}) + \rho t_4(\mathbf{y}) \right\} / K(\boldsymbol{\theta})\tag{8}$$

where network statistics $t_{2i}(\mathbf{y})$, $t_{3j}(\mathbf{y})$, and $t_4(\mathbf{y})$ refer to the outdegree of actor i , the indegree of actor j , and the number of mutual dyads, respectively, and $K(\boldsymbol{\theta})$ is a normalizing constant. Under this model, the probability distribution of possible networks is conditionally uniform given the two degree distributions and the dyad census.

We estimated the p_1 model for the physician network data by maximum likelihood using methods for fitting log-linear models (Fienberg and Wasserman 1981), imposing the identifying constraints $\sum_i \alpha_i = 0$ and $\sum_j \gamma_j = 0$ on the sender and receiver parameters, respectively. Estimates of these latter parameters generally correspond to the degree distributions shown in Table 2; for example, physician 27 has the largest indegree (24) and also the largest estimated receiver parameter ($\hat{\gamma}_{24} = 4.12$).⁹ The estimated reciprocity parameter (ρ) is 1.91. Interpretable as a log-odds ratio, it indicates that the predicted odds of a tie from physician j to physician i are nearly 7 times larger ($\exp(1.91) = 6.75$) if a tie from physician i to physician j is present. A likelihood ratio test statistic for reciprocity is 20.3 with 1 df. The distribution of this statistic appears to approach $\chi^2(1)$ as the number of actors (N) increases (Holland and Leinhardt 1981), suggesting (in accord with the previously-presented nonparametric test) a statistically significant tendency toward reciprocity.

Variations on the fixed-effect version of model (7), sometimes known as *a priori stochastic blockmodels*, accommodate categorical attribute data on actors. Such models may restrict model (7) by requiring that actors sharing an attribute value have identical “expansiveness” (α_i) and “attractiveness” (γ_j) parameters (Fienberg and Wasserman 1981); for example, the tendency to produce ties might be identical across male actors. Additionally, stochastic blockmodels may extend p_1 by relaxing the homogeneity constraints imposed on the density parameters μ_{ij} or the reciprocity parameters ρ_{ij} , for example by estimating separate density and/or reciprocity effects for pairs of actors who share an attribute value and those differing on the attribute (Fienberg and Wasserman 1981; Wang and Wong 1987); the density of contact or the tendency toward reciprocation might be greater for same-gender than for different-gender pairs. Such specifications imply inclusion of subgroup-specific network statistics in the p_1 density function shown above. When constraints on parameters imply that two actors have identical vectors of probabilities for their ties to others in the network, that pair of actors is said to be *stochastically equivalent* (Holland, Laskey, and Leinhardt 1983).

5.2. Mixed-Effect Dyadic Independence Models

As an alternative to the fixed effects in the p_1 model, structure may be introduced into the framework of model (7) by modeling the sender and receiver parameters using random effects together with actor-level covariates. The density and reciprocity effects remain fixed and subject to homogeneity conditions; they may, however, depend on dyadic covariates. For binary-valued network data, these specifications lead to a mixed-effect model known as p_2 (Van Duijn, Snijders and Zijlstra, 2004).¹⁰

Let vectors \mathbf{x}_{1ij} , \mathbf{x}_{2i} , \mathbf{x}_{3j} , and \mathbf{x}_{4ij} denote covariate sets that contribute to the density effect μ_{ij} , the sender effect α_i , the receiver effect γ_j , and the reciprocity effect ρ_{ij} , respectively; \mathbf{x}_{2i} and \mathbf{x}_{3j} are actor-level, while \mathbf{x}_{1ij} and \mathbf{x}_{4ij} are dyadic. The p_2 model assumes the following hierarchical structure for the parameters in (7):

$$\begin{aligned}\mu_{ij} &= \mu + \mathbf{x}_{1ij}^T \boldsymbol{\theta}_1 \\ \alpha_i &= a_i + \mathbf{x}_{2i}^T \boldsymbol{\theta}_2 \\ \gamma_j &= b_j + \mathbf{x}_{3j}^T \boldsymbol{\theta}_3 \\ \rho_{ij} &= \rho + \mathbf{x}_{4ij}^T \boldsymbol{\theta}_4,\end{aligned}\tag{9}$$

where a_i and b_j are mean-0 random effects assumed to have a multivariate normal distribution and an unrestricted covariance matrix. Wong (1987) studied a related Bayesian model that does

⁹Under p_1 , the estimate of a receiver parameter is infinitely small for actors with indegree 0; likewise, the estimate of a sender parameter is $-\infty$ when the corresponding outdegree is 0.

¹⁰The p_2 model is closely related to a social relations model developed by Kenny and La Voie (1984) for quantitative network variables.

not allow for the dependence of parameters on measured covariates. Gill and Swartz (2004) generalize the framework to other situations including *a priori* stochastic blockmodels and multirelational networks.

Estimation of the mixed-effect model specified by (9) requires methods for fitting hierarchical generalized linear models with crossed random effects. Van Duijn et al. (2004) outline an iterative generalized least squares algorithm, while Zijlstra, Van Duijn and Snijders (2006) take a Bayesian approach and suggest Markov Chain Monte Carlo (MCMC) methods for simulating the posterior distribution of the parameters in p_2 .

Keating et al. (2007) analyzed the physician network data using the p_2 model and MCMC estimation with diffuse priors. With no covariates, the median of the posterior distribution for the reciprocity parameter was 1.77 (95% credible interval (CI) 1.01 to 2.55), quite comparable to the estimate ($\hat{\rho} = 1.91$) from p_1 . Introducing covariates, they found that receiver effects were larger for physicians whose panels of patients included large percentages of women, who were self-reported experts in women's health, and who had larger numbers of patient sessions per week. The density parameter μ_{ij} was significantly larger (median 1.61, 95% CI 1.13 to 2.12) for pairs of physicians located in the same clinic within the practice. The estimated reciprocity parameter became smaller (median 1.29, 95% CI 0.50 to 2.17) after adjusting for the covariates. The residual random sender effects a_i and receiver effects b_j were uncorrelated (median covariance -0.22 , 95% CI -0.83 to 0.28).

Both the p_1 and p_2 models are restrictive because they consider only network statistics corresponding to configurations of one or two actors. However, an advantage of dyad-independence models is that the network consists of multiple independent configurations (namely dyads) and so there is a clear notion of how a sample can be drawn from the population of actors. This allows inferences and asymptotics to be treated in the usual manner. More complicated models are required to incorporate network effects involving dependencies involving multiple dyads, such as transitivity or closure. Recently-developed exponential random graph models permit such analyses of network data, although methods for sampling such data are still in their infancy (Section 6.3).

5.3. Exponential Random Graph Models (ERGMs)

ERGMs (Anderson, Wasserman, and Crouch 1999; Frank and Strauss 1986; Pattison and Wasserman 1999; Robins, Pattison, and Wasserman 1999), also known as p^* models, allow much more general forms of interdependency among network variables than those incorporated in dyadic independence models. ERGMs model the probability that a random network \mathbf{Y} is realized by an observed network \mathbf{y} as:

$$\Pr(\mathbf{Y}=\mathbf{y};\boldsymbol{\theta})=\kappa(\boldsymbol{\theta})^{-1}\exp\left(\sum_k\theta_kS_k(\mathbf{y})\right) \quad (10)$$

where $\kappa(\boldsymbol{\theta}) = \sum_{\mathbf{y} \in \Psi} \exp(\sum_k \theta_k S_k(\mathbf{y}))$ is a normalizing constant that makes the probabilities sum to 1 across possible networks, and Ψ is the set of possible networks.¹¹ The right-hand side of (10) describes a formula for producing random networks based on network statistics S_k that correspond to network features; its parameters indicate the sensitivity of the network-generating formula to particular features. A positive θ_k indicates that the rule for producing networks favors networks with feature k , while a negative value indicates that such networks tend to be avoided.

¹¹The large number of terms in $\kappa(\boldsymbol{\theta})$ complicates the estimation of ERGMs. There are $2^{N(N-1)}$ possible directed binary-valued networks; for example, with $N=10$, the number of possible networks—hence terms in $\kappa(\boldsymbol{\theta})$ —is 1.238×10^{27} .

In principle, any network statistic $S_k(\mathbf{y})$ may appear on the right-hand side of (10), and any subset of the $N(N-1)$ network variables may be conditionally dependent on one another. Many applications emphasize statistics corresponding to specific local *network configurations* consisting of small numbers of ties y_{ij} , so that $S_k(\mathbf{y}) = \prod_{y_{ij} \in k} y_{ij}$ is the binary network statistic denoting presence of configuration k . The most general ERGM allows a unique parameter for each distinct configuration (i.e., each subset of ties that takes the form of interest). Typically, however, models are simplified by imposing homogeneity constraints on parameters for isomorphic configurations, in which case the pertinent network statistics are sums over all such configurations.

Since (8) takes the form of (10), the fixed-effect p_1 model is an ERGM with parameters for configurations consisting of single directed ties y_{ij} themselves and mutual ties $y_{ij}y_{ji}$, as well as actor-level network statistics for outdegrees $\sum_i y_{ij}$ and indegrees $\sum_j y_{ij}$. Homogeneity constraints on the effects for ties and mutuality lead to the terms $S_1(\mathbf{y}) = \sum_{i,j} y_{ij}$ and $S_2(\mathbf{y}) = \sum_{i < j} y_{ij}y_{ji}$ on the right side of (10). Under p_1 , network variables are conditionally dependent if they share a sender, share a receiver, or involve reciprocity.

More general ERGMs add higher-order terms. Frank and Strauss (1986) introduced the notion of *Markov dependence*, under which two network variables y_{ij} and y_{kl} may be conditionally dependent if the two ties have any actor in common, i.e. if $i = k$, $i = l$, $j = k$, or $j = l$. This approach models degree distributions through the inclusion of statistics for “k-stars.” A k-star is a configuration in which k ties are incident to a particular actor; k-star configurations are nested within one another, so that an actor with degree m contributes k-stars for $k < m$; ¹² a positive regression parameter for such a configuration indicates a tendency for ties to cluster around a particular actor. Distinct k-out-star and k-in-star configurations exist in directed networks. For example, an indicator for the presence of a particular 2-out-star configuration is $y_{ij}y_{ih}$. Imposing a homogeneity constraint on parameters for all k-out-star configurations (for a given k) leads to the following network statistic for k-out-stars:

$$S_{3:k}(\mathbf{y}) = \sum_i \binom{y_{i+}}{k} \text{ where } y_{i+} = \sum_j y_{ij} \text{ is the out-degree of actor } i. \quad 13$$

An analogous definition holds for k-in-stars. Models typically include a small number of lower-order k-star terms rather than fitting degree distributions exactly, for parsimony and because the terms for different k are often highly collinear.

An additional configuration admissible for directed network data under Markov dependence is a “2-path” (or indirect tie), under which a given actor j is the receiver of one tie and the sender of a second; an indicator for the presence of a 2-path is the product of network variables $y_{ij}y_{jh}$, $h \neq i$. Very important in modeling networks are triadic configurations (products of three ties involving three distinct actors). In directed networks, the two triadic configurations of greatest interest are the transitive triad and the 3-cycle. With homogeneity constraints, these imply the following network statistics in (10):

$$\begin{aligned} \text{Transitive triad: } S_4(\mathbf{y}) &= \sum_{i < j < k} y_{ij}y_{ik}y_{jk} \\ 3\text{-cycle: } S_5(\mathbf{y}) &= \sum_{i < j < k} y_{ij}y_{jk}y_{ki} \end{aligned}$$

¹²For example, an actor with degree 3 contributes 1 3-star, 3 2-stars, and 3 1-stars; 1-stars are equivalent to individual edges.

¹³The set of k-star statistics is equivalent to the set of degree statistics (the number of nodes of degree k, $k = 1, 2, 3, \dots$) in that a bijection exists between the two sets of the statistics (Snijders et al. 2006).

The transitive triad is the key term for testing for tendencies toward closure in a network. Analyses of undirected networks use a single triadic “triangles” statistic.

Under Markov dependence, ties are conditionally independent unless they share at least one actor. This implies that dyads separated by at least one tie are conditionally independent given the rest of the network. An important theoretical result, the Hammersley-Clifford theorem (Besag 1974), shows that if all isomorphic graphs have the same probability under a model, then a random undirected graph is a Markov graph if and only if its probability distribution can be written as

$$\Pr(\mathbf{Y}=\mathbf{y};\boldsymbol{\theta})=\kappa(\boldsymbol{\theta})^{-1}\exp\left(\sum_{k=1}^{n-1}\theta_k S_{3:k}(\mathbf{y})+\tau S_4(\mathbf{y})\right) \quad (11)$$

where $S_{3:k}(\mathbf{y})$ is the number of k -stars and $S_4(\mathbf{y})$ is the number of triangles. Using appropriate network statistics recognizing directionality, model (11) generalizes to directed networks.

The assertion that isomorphic network configurations have homogeneous effects is often unduly restrictive. One way to relax it is by permitting the effects of a given configuration to vary with characteristics of actors. Under *Markov attribute dependence* (Robins et al. 2007), a configuration's effect may depend only on attributes of those actors involved in it, so that (e.g.) the parameter for the density configuration y_{ij} may depend on attributes of actors i and j , but not on those of actors $k \neq i, j$. The effect of any network configuration may depend on actor attributes, but applications focus on the density effect. For example, the probability that a tie is present may be greater when the receiver (j) has a particular gender or socioeconomic status x_j , implying the following network statistic for (10):

$$\text{Receiver covariate: } S_6(\mathbf{y}) = \sum_{i,j} y_{ij} x_j \cdot 14$$

Higher-way interactions between actor attributes and density are also common. For example, homophily effects (Section 3.8) can be assessed using a cross-product statistic between the density configuration and an indicator for attribute similarity:

$$\text{Nodematch covariate: } S_7(\mathbf{y}) = \sum_{i,j} y_{ij} I(x_i = x_j)$$

An ERGM model becomes non-Markovian when its network statistics involve configurations in which at least one pair of ties does not share an actor. Such configurations involve four or more actors. The number of potential statistics then escalates rapidly. One non-Markovian network configuration is a k -path (indirect path of length k); for example, an indicator for the presence of a 3-path is $y_{ij} y_{jk} y_{kh}$, $i \neq j \neq k \neq h$. Among many others is the k -cycle ($k > 3$), in which a sequence of k ties involving k distinct actors begins and ends with the same actor; the product $y_{ij} y_{jk} y_{kh} y_{hi}$, $i \neq j \neq k \neq h$ indicates that a 4-cycle is present.

ERGMs with third-order and higher terms become much more difficult to make inference about as they are essentially estimated from a sample size of 1, the observed network, which for the validity of inferences is assumed to be the whole network. If the observed network is the whole network, then inferences are to a super-population of networks that resemble the observed

¹⁴An analogous “sender covariate” statistic allows the density effect to depend on an attribute of the sender (i).

network. However, if the observed network is only a sample of the network, the model that generated the network may not have the same properties or even resemble the observed network. This incongruity between the sample network and the full network arises because there is no general way of decomposing networks into disjoint components whose sampling distribution is that of the full network (the population of interest in this context). As a consequence, some researchers advocate that only those models that can be constructed from generative processes (i.e., from assumptions about how two individual actors interact and form connections) should be used in modeling relational network data.

5.4. Model Estimation and Checking

Estimation, interpretation, and simulation for ERGMs is aided by the fact that (10) implies the following expression for the log-odds that a tie is present given the remainder of the network:

$$\text{logit}[\Pr(Y_{ij}=1|\mathbf{y}_{ij}^c)]=\boldsymbol{\theta}'\boldsymbol{\delta}(\mathbf{y}_{ij}^c) \quad (12)$$

where \mathbf{y}_{ij}^c is the realization of the network when the complement relation is applied to y_{ij} , and $\boldsymbol{\delta}(\mathbf{y}_{ij}^c)=t(\mathbf{y}_{ij}^+) - t(\mathbf{y}_{ij}^-)$ is the vector of changes in network statistics that occur if y_{ij} is 1 rather than 0.¹⁵ Multiplying a particular change statistic by the associated parameter value gives the change in the log-odds that the tie is present associated with the statistic in question, conditioned on the rest of the network (Snijders et al. 2006). For example, if a model includes the mutuality statistic $S_2(\mathbf{y})$ and the tie from j to i exists, then the presence of a tie from i to j would create an additional mutual tie, and the log-odds of observing $Y_{ij} = 1$ would increase by θ_2 , the regression coefficient for reciprocity.

Initially ERGMs were estimated using a pseudolikelihood function defined as the product of the conditional distributions implied by (12) over ordered pairs (for directed networks) or dyads (in the undirected case) (Besag 1975; Strauss and Ikeda 1990; Wasserman and Pattison 1996). Because the pseudolikelihood has the same form as the likelihood function for a logistic regression model, parameter estimates are easily obtained. However, unless dyadic independence holds, the pseudolikelihood differs from the true likelihood function, so inferences based on it can be unreliable.

Estimates with better properties can be obtained via the exact likelihood function for (10). Because the normalizing constant $\kappa(\boldsymbol{\theta})$ involves summation across the $2^{N(N-1)}$ possible (directed) networks, however, direct computation becomes intractable as the number of actors N increases. Recently developed Markov chain Monte Carlo (MCMC) methods now allow inferences to be based on the true likelihood function. One approach (Handcock 2003) relies on MCMC integration (Geyer and Thompson 1992). It is implemented in the R package Statnet (Handcock et al. 2003), which can fit models to moderate-sized networks (involving hundreds of actors). This algorithm simulates a sample of networks using a set of provisional parameter estimates; it then updates the estimates, approximating $\kappa(\boldsymbol{\theta})$ using the sampled networks and maximizing the associated likelihood function. An alternative approach (Snijders 2002) available in StocNET (Huisman and Van Duijn 2004, 2005) relies on a stochastic approximation algorithm. Obtaining convergence can be difficult using either approach because the likelihood surface based on (10) often has a highly irregular shape such that the estimation procedures become trapped at local maxima, fail to converge, or converge to inappropriate “degenerate” solutions (Handcock 2003). As an example below demonstrates, great care must be exercised when fitting ERGMs.

¹⁵ \mathbf{y}_{ij}^+ is the realization of the complement network with $y_{ij} = 1$, while \mathbf{y}_{ij}^- is the realization of the complement network with $y_{ij} = 0$.

Before interpreting or making inferences based on models fitted using MCMC, it is important to ensure that the Markov chain has converged to its stationary distribution by allowing a sufficiently long burn-in phase, and draw enough post burn-in samples to ensure that simulation error is below a specified threshold so that inferences are sufficiently precise. The coda package (Best, Cowles, and Vines 1995) in R can conduct the necessary checks in conjunction with Statnet (Handcock et al. 2003). We found that the default settings in Statnet—allowing 10,000 burn-in iterations and drawing 10,000 post-burn-in samples separated by intervals of 100, for a total of 1,010,000 iterations—were usually sufficient for models that did not contain triadic terms; the latter required longer simulations.

The overall fit of an ERGM may be quantified using statistics such as the deviance and the Bayesian information criterion (BIC). The deviance reflects the amount of variability explained by a model and so increases as terms are added. The BIC decreases as the deviance rises, tempered by a penalty reflecting the dimension (number of parameters) of the model.

Once estimates are obtained and convergence is assured, goodness of fit may be assessed by simulating a sample of networks implied by a model and then comparing observed and predicted distributions of network statistics. For statistics $S_k(\mathbf{y})$ included in the model, such comparisons are an additional diagnostic for convergence, since they tell whether the likelihood equations are satisfied stochastically. Comparisons involving statistics not in the model are signals of the adequacy of model specification. Statistics commonly used (and available in Statnet) for assessing model fit include the degree distribution, the distribution of geodesic distances between actors, and the numbers of contacts shared by dyads of actors or by those pairs linked by edges. See Hunter, Goodreau, and Handcock (2008) for a detailed discussion of methods for assessing model fit.

5.5. Illustrative Analysis: Directed Network

Our application fits models to the physician network including the density, mutuality and transitive triad configurations; we do not include k-star parameters to model the degree distributions in this illustrative analysis. We allow the density term to depend on three receiver covariates: women's health expertise (indicator variable), percentage female in a physician's panel of patients, and the number of clinical sessions per week. We estimated models using the Statnet software.

Table 8 presents estimates for four models. The first, a Bernoulli model, includes only the density (edges) statistic; its estimated coefficient is -1.701 . The associated inverse logit, 0.154 , equals the overall network density. The second model adds the mutuality statistic, which has a positive (1.187) and highly significant ($p < 0.0001$) coefficient. Using (10), we see that the log-odds that a tie is present rise if the reciprocal tie is observed, in accord with earlier observations about a tendency toward reciprocity in this network. Physicians in this practice evidently tend to regard their conversations with colleagues about women's health as being mutually influential. The estimated coefficient for the density (edges) term here (-1.952 , with inverse logit 0.124) is indicative of the density of ties in the absence of reciprocation; when a tie is reciprocated, however, predicted density rises to 0.318 , the inverse logit of the sum of the estimated density and mutuality parameters.

In many applied settings, substantive interest will focus on how attribute variables are associated with aspects of network structure. The third model illustrates this by letting the effect of the density configuration vary with the three receiver covariates. Estimates suggest that physicians with expertise in women's health, high proportions of female patients, and more clinical sessions per week were more apt to be cited as influential by their peers.

When we attempted to add the transitive triad term to the second model, we encountered difficulty in estimating model parameters by maximum likelihood. Although the resulting regression coefficients are finite and the deviance indicates that model fit improved, networks simulated using the estimated parameters tend to be extreme, often exhibiting a bimodal distribution including only fully dense or null networks. The proportions in each of the modes are such that the average simulated values of the model's three network statistics (edges, mutuality, and transitive triad) are close to the observed ones, suggesting good model fit, but this is misleading: the inability to simulate networks that resemble the observed network signals a radical discrepancy between model and data. This condition is commonly known as *degeneracy*, and is often encountered when fitting ERGMs including k-star and triadic terms. Degeneracy may arise because the network contains a high degree of structural heterogeneity (e.g., dense regions with many triangles mixed or high degree nodes mixed with low density regions), making it difficult (and perhaps impossible) to find parameter values that describe the entire network.

Because of the degeneracy with the fit of the model with the transitive triad term, Table 8 reports pseudolikelihood estimates based on (12). These estimates suggest a tendency toward network closure; i.e., that if one physician influences a second indirectly through a third, the first physician also tends to influence the second directly. Because properties of the pseudolikelihood estimates are poorly understood, interpretation and inference based on these estimates can be only tentative and cautious.

The degeneracy encountered in our attempt to fit the last model may indicate that a single homogeneous transitivity effect does not describe this network well; we note from Figure 1 that it appears to contain one or two very dense regions. Such clustering may indicate that transitive triads tend to be proximate to one another, so that higher-order terms are necessary. The following section introduces some recently developed network statistics that capture such phenomena and can be helpful when degeneracy is encountered.

5.6. Overcoming Estimation Problems: New Parameterizations

For simplicity, in this section we consider only models for undirected networks. There, the extreme lack of fit known as degeneracy is commonly confronted when fitting ERGM models including k-star or triadic terms. Attempting to model the degree distribution using a single 2-star term often leads to problems like those illustrated by the last model in Table 8, which included the transitive triad statistic.

Models including higher-order stars often yield more satisfactory estimates (Robins et al. 2005). In such models, the magnitude of the coefficients for successive star terms often declines as the order of the stars rises; moreover, the signs of these coefficients tend to alternate, so that a negative 3-star parameter tempers the tendency of ties to concentrate on particular actors implied by a positive 2-star parameter. Since the multiple k-star terms usually exhibit substantial collinearity, imposing linear restrictions on their coefficients simplifies estimation, leading to the alternating k-star statistic proposed by Snijders et al. (2006):

$$S_8(\mathbf{y}, \lambda_1) = \sum_{k=2}^{N-1} (-1)^k \frac{S_{3:k}(\mathbf{y})}{\lambda_1^{k-2}},$$

where λ_1 is a parameter (ordinarily greater than 1) governing the rate at which the magnitude of the regression coefficients for k-star terms declines as k rises.¹⁶

A similar statistic has proved useful in addressing degeneracy problems encountered when attempting to fit models to undirected networks that include a parameter for "triangle"

configurations $y_{ij} y_{ik} y_{jk}$. A higher-order configuration that captures a tendency of triangles to cluster in the vicinity of one another—as may occur in Figure 1—is known as the “*k-triangle*”: a set of k triangles resting on a common base. For example, an indicator for the presence of a 2-triangle resting on the base y_{ij} is $y_{ij} y_{ik} y_{jk} y_{ih} y_{jh}$; two triangles (involving actors i, j , and k , and actors i, j , and h) overlap in y_{ij} . The k -triangle terms may be combined into a statistic for clustering of transitive configurations that is not linear in the triangle count but instead gives lower probability to highly clustered structures. Paralleling the alternating k -star statistic, this *alternating k-triangle* statistic is defined as

$$S_9(\mathbf{y}, \lambda_2) = 3S_{4:1}(\mathbf{y}) + \sum_{k=2}^{N-2} \left(\frac{-1}{\lambda_2} \right)^{k-1} S_{4:k}(\mathbf{y})$$

where $S_{4:k}(\mathbf{y})$ is the number of k -triangles ($S_{4:1}(\mathbf{y}) = S_4(\mathbf{y})$ is the regular triangle statistic) and $\lambda_2 > 1$ again governs the rate at which the regression coefficients of k -star terms decline as the order of k -triangles rises. Models including the alternating k -star and/or alternating k -triangle parameters may fix the λ parameters or estimate them; in the latter case, the model does not take the form of (10) but becomes a *curved exponential* model (Hunter 2007).

The alternating k -triangle statistic may be rewritten using the fact that in a k -triangle, the two “base” actors have k common neighbors or “partners”. This leads to the *geometrically weighted edgewise shared partner* (GWESP) statistic

$$S_{10}(\mathbf{y}, \rho) = e^\rho \sum_{k=1}^{N-2} \{1 - (1 - e^{-\rho})^k\} EP_k(\mathbf{y})$$

where $EP_k(\mathbf{y}) = \sum_{i < j} y_{ij} I(sp_{ij} = k)$ is the number of pairs of linked actors who share k partners, $sp_{ij} = \sum_k y_{ik} y_{jk}$ is the number of partners shared by actors i and j (Goodreau 2007; Hunter 2007) and parameter ρ controls the rate at which the weights assigned to configurations having k shared partners decline with k .

5.7. Illustrative Analysis: Undirected Network

Table 9 reports estimates for five models fit to the undirected physician influence network. The first is again a Bernoulli model including only the edges term; its estimated coefficient of -1.049 has an associated inverse logit of 0.259 equal to the density of the undirected network. The second model adds the GWESP term $S_{10}(\mathbf{y}, \rho)$ with a fixed coefficient ρ of 1.2 .¹⁷ The fit of the model was nondegenerate, and the positive regression coefficient for the GWESP term offers evidence of transitivity in the undirected network.

The third model, a curved exponential model, adds 2- and 3-star terms to model the degree distribution, and also estimates parameter ρ in the GWESP term. The coefficients for the two k -star terms have opposite signs, but both are insignificant; hence there appears to be no

¹⁶An equivalent statistic based on the degree distribution itself is known as the “geometrically weighted degree” statistic; see Hunter and Handcock (2006).

¹⁷No mutuality term is included, since this is redundant with the edges term in an undirected network. Constraining the value of ρ when fitting the model with the GWESP term is often helpful; attaining adequate convergence is more difficult when it is estimated as a free parameter. We found that setting $\rho = 1.2$ served well here; the likelihood surface is relatively flat, so that using a value between 1.0 and 1.5 did not affect inferences about other parameters. Note, however, ρ was estimated at 0.93 when we left it as a free parameter in the third model in Table 9.

tendency, after adjusting for clustering, for ties in this network to concentrate on particular actors.

No actor-level covariates were directly associated with the density of ties in this network; because the Keating et al. (2007) p_2 analysis of the directed network found that actor covariates influenced the tendency to receive but not to make citations, the absence of covariate effects here may be due to the fact that the directionality of ties was removed when we constructed the undirected version of network. In common with Keating et al. (2007), however, we did find evidence that physicians of the same gender and in the same clinic within the practice tended to cite one another. In addition to the edges and GWESP terms, our fourth model introduces dyadic covariates for pairs of physicians of the same gender and in the same clinic; the “same clinic” coefficient is constrained to be the same for each of the four distinct clinics within the practice. Coefficient estimates suggest a significantly higher density among physicians of the same gender, but net of the general clustering effect modeled by the GWESP term, no similar tendency is evident for physicians in the same clinic. When we instead allow the “same clinic” effect to vary across the four distinct clinics, in the fifth model, we find a significantly elevated density within the fourth practice, which includes physicians who specialize in women’s health. Estimated parameters for density within the other three clinics are also positive, but have p values larger than 0.10. The p value for the coefficient of the “same gender” statistic in model 5 exceeds 0.3, though its estimate (0.433) is only somewhat smaller than the corresponding one in model 4 (0.532, p value 0.002); the difference may reflect a tendency for male and female physicians to be based in different clinics.

Indicators of goodness of fit point to different conclusions about which of these models best corresponds to the data. BIC prefers the more parsimonious models, taking its smallest values for models 2 and 4. Comparison of the deviance for nested models 4 and 5 suggests, however, that the latter has a better fit (difference in deviance=12.5, 3 df), and thus that the tendency toward homophily differs significantly across clinics. Observed network statistics are best reproduced via simulations based on models 4 and 5, indicating that homophily by clinic and gender play an important role in structuring this network.

5.8. Software for Relational Models

While some elementary statistical models for networks such as p_1 can be estimated via routines in standard software packages, most require specialized programs. Statnet (Handcock et al. 2003) is a suite of R packages for statistical network analysis; its “ergm” package conducts MCMCMLE estimation of ERGMs. Modules in StOCNET (Boer et al. 2006) implement several models covered here, including nonparametric tests based on enumeration or simulation, p_2 , and stochastic *a posteriori* blockmodels (see below). Its SIENA module estimates ERGMs as well as a model for longitudinal data introduced in the next section. PNET (Wang, Robins, and Pattison 2008) estimates ERGMs via the stochastic approximation algorithm used in SIENA.

6. Recent Developments for Modeling Networks

This article introduced and illustrated representations of network data, descriptive measures of networks and the two main types of statistical network models. Though we covered many widely-used network methods, we cannot be comprehensive here. We briefly touch on some additional developments in network analysis, including latent-variable models, longitudinal network analysis, and methods for network sampling.

6.1. Latent Variable Models for Network Data

Models such as ERGMs specify that interactions among observed ties, together with measured covariates, underlie observed network structures. An alternative is to posit unobserved actor-level covariates that account for observed network patterns. Such latent covariates may be either categorical or quantitative. Nowicki and Snijders (2001), for example, develop *a posteriori* stochastic blockmodels which seek to assign actors to classes of a latent categorical variable such that actors within a class exhibit stochastically equivalent relational patterns.

Latent position models (Hoff, Raftery, and Handcock 2002) introduce latent quantitative variables. One latent-position specification asserts that observed binary-valued ties Y_{ij} are conditionally independent given the locations of actors i and j in a latent space and that ties are more likely when pairs are close in the latent space. A common measure of distance is the

Euclidean distance $d(\mathbf{z}_i, \mathbf{z}_j) = \left(\sum_{k=1}^K (z_{ik} - z_{jk})^2 \right)^{1/2}$, where $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})^T$ is a latent variable locating actor i in a K -dimensional Euclidean space. The presence of ties may also depend on a vector of measured covariates, \mathbf{x}_{ij} , leading (in the case of a logit link) to a model of the form

$$\log \left(\frac{\text{pr}(Y_{ij}=1|\mathbf{z}_i, \mathbf{z}_j, \mathbf{x}_{ij})}{1 - \text{pr}(Y_{ij}=1|\mathbf{z}_i, \mathbf{z}_j, \mathbf{x}_{ij})} \right) = \alpha + \beta^T \mathbf{x}_{ij} - \left(\sum_{k=1}^K (z_{ik} - z_{jk})^2 \right)^{1/2}. \quad (13)$$

Because latent distances for a triple of actors must obey the triangle inequality, this formulation models the tendencies toward transitivity commonly found in social networks. A latent cluster model (Handcock, Raftery, and Tantrum 2007) is a variation on (10), specifying that latent positions for individual actors are mixtures of patterns associated with two or more latent categorical groups of actors. The LatentNet package in R (Handcock et al. 2007) uses Bayesian methods to fit such models.

A related generalized bilinear mixed-effect model developed by Hoff (2005) also assumes conditional independence among ties, but uses an inner-product specification for the effect of the latent quantitative variable, adds actor-level random effects for senders and receivers, and includes a hierarchical structure like that in the p_2 model (see (9)) for the effects of measured covariates. For binary-valued network data, this model takes the form:

$$\begin{aligned} \log \left(\frac{\text{pr}(Y_{ij}=1|\mathbf{z}_i, \mathbf{z}_j, \mathbf{x}_{ij})}{1 - \text{pr}(Y_{ij}=1|\mathbf{z}_i, \mathbf{z}_j, \mathbf{x}_{ij})} \right) &= \beta^T \mathbf{x}_{ij} + \varepsilon_{ij} \text{ where } \varepsilon_{ij} = a_i + b_j + \gamma_{ij} + \xi_{ij} \\ \text{with } (a_i, b_i)^T &\sim N(0, \Sigma_{ab}), \Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}, \\ (\gamma_{ij}, \gamma_{ji})^T &\sim N(0, \Sigma_\gamma), \Sigma_\gamma = \sigma_\gamma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \end{aligned} \quad (14)$$

and $\xi_{ij} = \mathbf{z}_i^T \mathbf{z}_j$ with $\mathbf{z}_i \sim N(\mathbf{0}, \Sigma_z)$. The random-effect variances σ_a^2 and σ_b^2 (respectively) quantify dependence among observations having a common sender or a common receiver, γ_{ij} represents an unconstrained sender-receiver interaction, and ρ represents reciprocity, the correlation of values of γ_{ij} within a dyad. The inner-product interaction ξ_{ij} involving latent scores \mathbf{z}_i and \mathbf{z}_j implies that the probability of a tie between actors i and j rises to the extent that the latent vectors \mathbf{z}_i and \mathbf{z}_j have similar direction and magnitude. Inclusion of ξ_{ij} —in (14) interpretable as a mean-0 random effect—models transitivity by constraining the extent to which the inner products $\mathbf{z}_i^T \mathbf{z}_k$, $\mathbf{z}_i^T \mathbf{z}_j$ and $\mathbf{z}_j^T \mathbf{z}_k$ can differ from one another. The magnitude of network transitivity may be summarized by $\text{var}(\mathbf{z}_i^T \mathbf{z}_j) = \text{trace}(\sum_z^T \sum_z)$, which reduces to $K\sigma_z^4$ in the special case

where $\sum_z = \sigma_z^2 \mathbf{I}$. The greater the magnitude of Σ_z , the greater the variation of the \mathbf{z}_i and thus of their inner products; hence, the greater the potential for transitivity (or other third-order effects such as cyclical patterns). To test for the presence of such third order effects one could compare the fit of the model with and without the inner-product term (e.g. using the deviance information criterion).

The latent locations \mathbf{z} estimated under the latent-position and generalized bilinear mixed effects models provide a statistically-grounded foundation for network visualization. Moreover, plots of MCMC-simulated values of the latent locations represent uncertainty in a spatial representation that is not captured by descriptive visualization tools.

An attractive feature of latent-variable models relying on conditional independence specifications is that they readily accommodate non-binary network data. By altering the link function on the left-hand side of (13) or (14), these models are readily adapted to the analysis of relational data in the form of quantitative variables (e.g., linear link) or counts (e.g., log-link).

6.2. Longitudinal network analysis

The vast majority of past social network research examines data from a single point in time. However, interest in longitudinal data suitable for studying network evolution is now increasing rapidly. As usual, over-time observations on a network can help allay concerns about reciprocal causation and provide a superior basis for isolating causal effects. Furthermore, some contend that since models for network change condition on the status of a network at baseline, they may be simpler to fit than that of models seeking to account for how a network came into existence (Snijders 2005).

Only a few extant models for network evolution (Doreian and Stokman 1997) adopt a statistical approach. Such models study network change within a continuous-time Markov Chain framework in which the central construct is an intensity matrix governing the rates at which ties arise and disappear. Initial efforts modeled change in tie status assuming dyadic independence (Holland and Leinhardt 1977; Wasserman 1979, 1980). In these, rates of change were dependent on a single network property such as reciprocity. Due to the simplicity of such models, closed-form expressions for the transition probabilities often exist, so that maximum likelihood estimates can be computed using standard optimization procedures. However, such models seldom provide adequate descriptions of network change.

A much more elaborate statistical model for network evolution is the actor-oriented model proposed by (Snijders 1996, 2001, 2005). This centers on an objective function for actors which may be sensitive to multiple network properties including (e.g.) reciprocity, closure, homophily, or contact with prestigious others. The model assumes that actors control their outgoing ties and alter them in order to increase their satisfaction with the network in one or more respects. Estimated parameters indicate whether changes in a given property raise or lower actor satisfaction. An important distinction from ERGMs is that the relevant network statistics in the actor-oriented model are actor-specific rather than aggregations across the network. Apart from the objective function, this model may also include a rate function describing the rate of change in an actor's outgoing ties, and a gratification function indicative of differences in satisfaction flowing from the formation and dissolution of ties.

Estimating the actor-oriented model is in general well outside the scope of standard maximum likelihood methods. MCMC simulation methods relying on a stochastic approximation algorithm are available (Snijders 2001) to support inferences via method-of-moments, maximum likelihood, or Bayesian criteria. Details on these procedures may be found in the

documentation for the SIENA module in the StOCNET software package (Huisman and Van Duijn 2004, 2005; Snijders et al. 2007).

Longitudinal modeling will be a major area of growth in network analysis. Recent elaborations of the actor-oriented model allow actors to join or leave the network over time (Huisman and Snijders 2003; Snijders 2005), certain ties to be unalterable (e.g., impossible or certainty ties), and partial imputation¹⁸ of missing tie-status over intervals of time (Snijders et al. 2007). Finally, if both the network and actors' behaviors are measured, agent-based models that allow the changing network to influence actors' behavior—while simultaneously allowing changes in actors' behavior to influence the network—can be developed. Such models treat a network as an endogenous feedback mechanism (Zeggelink 1994). Recent releases of SIENA implement all of these innovations to some extent.

6.3. Methods for sampling networks

Sampling networks can be divided into two main types: sampling partial networks and sampling whole networks. In the former the sample typically induces the network (often the full network is unobservable) while in the latter an underlying network is known but is infeasible to observe or analyze in its entirety (e.g. internet based networks).

Several schemes for sampling partial networks exist (Frank 1981). A sample-induced network is obtained by sampling from the set of units/actors, and then assembling data on relationships among the sampled units. Link-tracing or “random walk” designs randomly sample chains of relationships in a network.

Another scheme samples stars (Frank 1981). A widely-used variant on this samples from a set of units/actors and obtains data on the subnetworks surrounding them—including attributes of the “alter” units/actors to which they are linked directly, and relationships among those alter units. Such schemes have been implemented within conventional sample surveys of individuals (e.g. Marsden 1987). Studies using such *egocentric network data* seek to account for variations in local network structure, or to explain variations in attitudes or behaviors of the sampled units/actors using properties of their egocentric networks. Certain properties of whole networks, including the degree distribution (see Section 3.2), may also be approximated using egocentric network samples. The National Health and Social Life Survey (Laumann et al. 2004), for example, measured egocentric sexual and social networks.

Administrative data and electronic communication systems increasingly allow assembly of whole-network data for large networks at modest cost. If data collection is costly, however, sampling from large networks to learn about network properties is appealing. Foundational work by Ove Frank (Frank 1971, 1981, 1988) remains key in this field. Frank outlines several designs for drawing node-induced network samples—e.g. by simple random sampling from a population of actors (with or without replacement) followed by observation of relationships among the sampled nodes—and the inferences about network properties available from them. A node-induced sample, for instance, yields inferences about node, dyad, and triad totals, as well as the degree distribution (Frank 1978, 1981).

An important use of network sampling is oriented less to estimating network properties than to locating elements of rare, unlisted and/or hidden populations and estimating properties of the distribution of attributes of the nodes/actors in such populations. Generally known as snowball or chain-referral sampling, such designs draw an initial sample of actors and then trace one or more links to alters of each element in that initial sample; this link-tracing may be

¹⁸Although the missing values are replaced with non-missing values during model fitting, the statistics measuring model fit are only evaluated using actors with non-missing values throughout the corresponding interval of time. Thus, standard imputation is not performed.

repeated several times. For example, network versions of multiplicity sampling (e.g. Sudman and Kalton 1986), draw a first-stage sample from a general population by conventional methods, and then select elements in the special population of interest that are related to the first-stage elements in some well-defined way (e.g. kinship, co-residence). To develop appropriate weights, one must measure the degree (egocentric network size) of elements in the special-population sample that reflect the density of their ties to the general-population sample. Respondent-driven sampling (Salganik and Heckathorn 2004) begins with “seeds” known to be within the special population of interest, and encourages them to refer others in the population via network ties. After several waves of such referrals, this method tends toward a probability-proportional-to-degree sample from the special population even if the seeds are arbitrarily and nonrandomly chosen.

Thompson (2006) develops a very general link-tracing approach termed adaptive web sampling. It may be used to estimate both actor and network properties (such as the degree distribution). Beginning with a randomly drawn set of seeds, an adaptive web sampling design may trace a link from an already-sampled actor to a related alter, thus investigating interconnected segments of a population. With some probability, however, it also may draw new elements at random. The probability of making each kind of draw may depend on characteristics of the current sample.

7. Conclusion

Social network analysis has historically been pursued mainly in the social science disciplines, but its use has grown rapidly in recent years into many other areas. Among recent applications in health care and medicine are studies of the physician influence network by (Keating et al. 2007), the epidemic spread of obesity described by Christakis and Fowler (2007) and that of smoking (Christakis and Fowler 2008), the spread of AIDS (Morris et al. 2006), the spread of knowledge about new medical technology (Miguel and Kremer 2003), patterns of contraceptive use over time (Behrman, Kohler, and Watkins 2002), and the spread of sexually transmitted infections via sexual networks by Laumann and colleagues (Laumann et al. 2004). As has been the trend in health, most of the above applications treat the network as fixed (e.g., as in individual outcomes models) rather than modeling the network as in the analysis of relational data.

We anticipate that health-related applications of social network analysis will grow rapidly during the coming decade, since interpersonal relationships and support networks are crucial to the well-being of most persons, and because appropriate methods for addressing the difficult analytic problems posed by social network data are increasingly available. Informed applications of social network analysis in health services and outcomes research will not only yield new insights into these phenomena, but contribute toward continued improvements in social network methodology.

There are a wide range of topics that statisticians can address in the future, including some motivated by the physician network data. These include predicting how a new physician that joins a clinic will interact with those already there, and extrapolating inferences based on one network to other networks (e.g., other hospitals). Solutions to the first problem will likely require dynamic modeling of longitudinal network data to identify the effects of change in the network. The second problem requires careful consideration of the population to which the inference is being applied and the conditions under which results for one network can be extended to another; to some extent this is a sampling problem. Dynamic modeling of networks and methods of sampling networks are two areas that have not to date been heavily researched but to which statisticians can and should be heavily involved. Methods for handling missing data in network analysis are also in their infancy in contrast to mainstream statistics.

In addition to methodological work, further applied work can be undertaken by developing new applications of social network analysis. Though on the rise, applications to health care and medicine are still relatively few. Statisticians will also play an important role by highlighting limitations of models and potential traps of software packages. For example, practitioners must be sensitized to such issues as the prospect of degeneracy or poor convergence of estimation algorithms and be made aware of the need to studiously check multiple diagnostics to ensure valid interpretation of results.

Acknowledgments

We thank Nancy Keating for allowing us to re-analyze the data from the physician influence network. Research for the paper was supported by NIH grants R01 AG024448-02, P01 AG031093, and Robert Wood Johnson Foundation Award #58729.

References

- Anderson C, Wasserman S, Crouch B. A p* primer: logit models for social networks. *Social Networks* 1999;21(1):37–66.
- Anselin, L. *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1988.
- Anselin L. Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics* 1990;20(2):141–63.
- Banerjee, S.; Carlin, B.; Gelfand, A. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall; 2004.
- Barabási, A-L. *Linked: The New Science of Networks*. New York: Perseus; 2002.
- Bartholomew, D.; Steele, F.; Moustaki, I.; Galbraith, J. *The Analysis and Interpretation of Multivariate Data for Social Scientists*. New York: Chapman and Hall; 2002.
- Batagelj, V.; Mrvar, A. Pajek: Analysis and visualization of large networks. In: Jünger, M.; Mutzel, P., editors. *Graph Drawing Software*. New York: Springer; 2003. p. 77-103.
- Beauchamp M. An improved index of centrality. *Behavioral Science* 1965;10:161–63. [PubMed: 14284290]
- Behrman J, Kohler H-P, Watkins S. Social Networks and Changes in Contraceptive Use Over Time: Evidence from a longitudinal study in rural Kenya. *Demography* 2002;39:713–38. [PubMed: 12471851]
- Berkman, L.; Glass, T. Social integration, social methods, social support, and health. In: Berkman, L.; Kawachi, I., editors. *Social Epidemiology*. New York: Oxford University Press; 2000. p. 137-73.
- Berkman L, Syme S. Social Networks, Host Resistance, and Mortality: A Nine-year Follow-up Study of Alameda County Residents. *American Journal of Epidemiology* 1979;109:86–204.
- Besag J. Spatial interaction and statistical-analysis of lattice systems. *Journal of the Royal Statistical Society, Series B: Methodological* 1974;36(2):192–236.
- Besag J. Statistical analysis of non-lattice data. *Journal of the Institute of Statisticians* 1975;24:179–96.
- Best, N.; Cowles, M.; Vines, K. *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health; 1995.
- Boer, P.; Huisman, M.; Snijders, T.; Steglich, M.; Wicher, L.; Zeggelink, E. *StOCNET User's Manual*, version 1.7. Groningen, NL: ICS; 2006.
- Bonacich P. Power and Centrality: A Family of Measures. *American Journal of Sociology* 1987;92:1170–82.
- Borgatti, S. *NetDraw: Graph Visualization Software*. Lexington, KY: Analytical Technologies; 2008.
- Borgatti, S.; Everett, M.; Freeman, L. *UCINET 6 for Windows: Software for Social Network Analysis*. Lexington, KY: Analytical Technologies; 2002.
- Burt, R. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press; 1992.

- Burt R, Doreian P. Testing a structural model of perception: conformity and deviance with respect to journal norms in elite sociological methodology. *Quality and Quantity* 1982;16:109–50.
- Butts, C. sna: Tools for Social Network Analysis (release 1.5). 2007.
- Christakis N. Social networks and collateral health effects. *British Medical Journal* 2004;329(7459):184–5. [PubMed: 15271805]
- Christakis N, Fowler J. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine* 2007;357:370–79. [PubMed: 17652652]
- Christakis N, Fowler J. The Collective Dynamics of Smoking in a Large Social Network. *New England Journal of Medicine* 2008;358:2249–58. [PubMed: 18499567]
- Coleman, J.; Katz, E.; Menzel, H. *Medical Innovation: A Diffusion Study*. Indianapolis: Bobbs-Merrill; 1966.
- Doreian P. Linear-models with spatially distributed data- spatial disturbances or spatial effects. *Sociological Methods and Research* 1980;9(1):29–60.
- Doreian, P. Estimating linear models with spatially distributed data. In: Leinhardt, S., editor. *Sociological Methodology*. San Francisco: Jossey-Bass; 1981. p. 359-88.
- Doreian, P. Network Autocorrelation Models: Problems and Prospects. In: Griffith, DA., editor. *Spatial Statistics: Past, Present, Future*. Ann Arbor: Michigan Document Services; 1989. p. 369-89.
- Doreian, P.; Stokman, F. Evolution of social networks: Processes and principles. In: Doreian, P.; Stokman, F., editors. *Evolution of Social Networks*. Amsterdam: Gordon and Breach Publishers; 1997. p. 233-50.
- Dow M. A biparametric approach to network autocorrelation. *Sociological Methods and Research* 1984;13:201–17.
- Erdős P, Rényi A. On random graphs. *Publicationes Mathematicae* 1959;6:290–97.
- Fienberg, S.; Wasserman, S. Categorical data analysis of single sociometric relations. In: Leinhardt, S., editor. *Sociological Methodology*. San Francisco: Jossey-Bass; 1981. p. 156-92.
- Frank, O. *Statistical Inference in Graphs*. Stockholm: FOA Repro; 1971.
- Frank O. Sampling and Estimation in Large Social Networks. *Social Networks* 1978;11:91–101.
- Frank, O. A Survey of Statistical Methods for Graph Analysis. In: Leinhardt, S., editor. *Sociological Methodology*. San Francisco: Jossey-Bass; 1981. p. 110-55.
- Frank O. Random Sampling and Social Networks: A survey of various approaches. *Mathematiques, Informatique, et Sciences Humaines* 1988;26:19–33.
- Frank O, Strauss D. Markov graphs. *Journal of the American Statistical Association* 1986;81(395):832–42.
- Freeman L. Centrality in Social Networks, I. Conceptual Clarification. *Social Networks* 1979;1:215–39.
- Freeman, L. Social Networks and the Structure Experiment. In: Freeman, L.; White, D.; Romney, A., editors. *Research Methods in Social Network Analysis*. Fairfax, VA: George Mason University Press; 1989. p. 11-40.
- Freeman, L. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver, BC: Empirical Press; 2004.
- Friedkin N. Social Networks in Structural Equations Models. *Social Psychology Quarterly* 1990;53:316–28.
- Friedkin N, Cook K. Peer Group Influence. *Sociological Methods and Research* 1990;19(1):122–43.
- Fruchterman T, Reingold E. Graph Drawing by Force-Directed Placement. *Software-Practice and Experience* 1991;21(11):1129–64.
- Geyer C, Thompson E. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society, Series B* 1992;54(3):657–99.
- Gill P, Swartz T. Bayesian analysis of directed graphs data with applications to social networks. *Journal of the Royal Statistical Society, Series C: Applied Statistics* 2004;53:249–60.
- Goodreau S. Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network. *Social Networks* 2007;29:231–48. [PubMed: 18449326]
- Haines V, Hurlbert J. Network Range and Health. *Journal of Health and Social Behavior* 1992;33:254–66. [PubMed: 1401850]

- Handcock, M. Assessing Degeneracy in Statistical Models of Social Networks. Seattle: Center for Statistics and Social Sciences, University of Washington; 2003.
- Handcock, M.; Hunter, D.; Butts, C.; Goodreau, S.; Morris, M. Statnet: Software tools for the Statistical Modeling of Network Data (release Version 2.1). Seattle, WA: Center for Statistics and Social Sciences, University of Washington; 2003. Project home page at <http://statnetproject.org>; Software available at <http://CRAN.R-project.org/package=statnet>
- Handcock M, Raftery A, Tantrum J. Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A* 2007;170(2):301–54.
- Harville, D. Matrix algebra from a statistician's perspective. New York: Springer-Verlag Inc; 1997.
- Hoff P. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association* 2005;100:286–95.
- Hoff P, Raftery A, Handcock M. Latent space approaches to social network analysis. *Journal of the American Statistical Association* 2002;97:1090–98.
- Holland P, Laskey K, Leinhardt S. Stochastic Blockmodels: First Steps. *Social Networks* 1983;5:109–37.
- Holland, P.; Leinhardt, S. Local Structure in Social Networks. In: Heise, D., editor. *Sociological Methodology*. San Francisco: Jossey-Bass; 1976. p. 1-45.
- Holland P, Leinhardt S. A Dynamic Model for Social Networks. *Journal of Mathematical Sociology* 1977;5:5–20.
- Holland P, Leinhardt S. An exponential family of probability-distributions for directed-graphs. *Journal of the American Statistical Association* 1981;76(373):33–50.
- House, J.; Kahn, R. Measures and concepts of social support. In: Cohen, S.; Syme, S., editors. *Social Support and Health*. New York: Academic Press; 1985. p. 83-108.
- Huisman M, Snijders T. Statistical analysis of longitudinal network data with changing composition. *Sociological Methods and Research* 2003;32:253–87.
- Huisman, M.; Van Duijn, M. Software for Statistical Analysis of Social Networks. The Sixth International Conference on Logic and Methodology; Amsterdam, The Netherlands. 2004.
- Huisman, M.; van Duijn, M. Software for Social Networks Analysis. In: Carrington, PJ.; Scott, J.; Wasserman, S., editors. *Models and Methods in Social Network Analysis*. Cambridge: Cambridge University Press; 2005. p. 270-316.
- Hunter D. Curved Exponential Family Models for Social Networks. *Social Networks* 2007;29:216–30. [PubMed: 18311321]
- Hunter D, Goodreau S, Handcock M. Goodness of fit of social network models. *Journal of the American Statistical Association* 2008;103:248–58.
- Hunter D, Handcock M. Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics* 2006;15(565–583)
- Katz L, Powell J. Probability distributions of random variables associated with a structure of the sample space of sociometric investigations. *Annals of Mathematical Statistics* 1957;28:442–48.
- Keating N, Ayanian J, Cleary P, Marsden P. Factors Affecting Influential Discussions Among Physicians: A Social Network Analysis of a Primary Care Practice. *Journal of General Internal Medicine* 2007;22(6):794–98. [PubMed: 17404798]
- Kenny, D.; Voie, LLa. The Social Relations Model. In: Berkowitz, L., editor. *Advances in Experimental Social Psychology*. New York: Academic Press; 1984. p. 142-82.
- Klov Dahl A. Social Networks and the Spread of Infectious Diseases. *Social Science and Medicine* 1985;21:1203–16. [PubMed: 3006260]
- Land, K.; Deane, G. On the large-sample estimation of Regression Models with Spatial or Network Effects Terms: A Two-Stage Least-Squares Approach. In: Marsden, PV., editor. *Sociological Methodology*. Oxford: Basil Blackwell, Ltd; 1992. p. 221-48.
- Laumann, E.; Mahay, J.; Paik, A.; Youm, Y. Network Data Collection and Its Relevance for the Analysis of STDs: The NHSLs and CHSLs. In: Morris, M., editor. *Network Epidemiology: A Handbook for Survey Design and Data Collection*. New York: Oxford University Press; 2004. p. 27-41.

- Laumann, E.; Marsden, P.; Pinsky, D. The Boundary Specification Problem in Network Analysis. In: Burt, R.; Minor, M., editors. *Applied Network Analysis: A Methodological Introduction*. Beverly Hills, CA: Sage Publications; 1983. p. 18-34.
- Laumann E, Youm Y. Racial/Ethnic Group Differences in the Prevalence of Sexually Transmitted Diseases in the United States: A Network Explanation. *Sexually Transmitted Diseases* 1999;26:250-61. [PubMed: 10333277]
- Leenders R. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks* 2002;24(1):21-47.
- Marsden P. Core Discussion Networks of Americans. *American Sociological Review* 1987;52(1):122-31.
- Marsden P. Network Data and Measurement. *Annual Review of Sociology* 1990;16:435-63.
- Marsden P. Egocentric and Sociocentric Measures of Network Centrality. *Social Networks* 2002;24:407-22.
- Marsden, P. Network Methods in Social Epidemiology. In: Oakes, JM.; Kaufman, JS., editors. *Methods in Social Epidemiology*. San Francisco: Jossey-Bass; 2006. p. 267-86.
- McGrath C, Blythe J, Krackhardt D. The Effect of Spatial Arrangement on Judgments and Errors in Interpreting Graphs. *Social Networks* 1997;19(3):223-42.
- McPherson M, Smith-Lovin L, Cook J. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 2001;27:415-44.
- Miguel E, Kremer M. Networks, Social Learning, and Technology Adoption: The Case of Deworming Drugs in Kenya. *Poverty Action Laboratory*. 2003
- Morris M, Handcock M, Miller W, Ford C, Schmitz J, Hobbs M, Cohen M, Harris K, Udry J. Prevalence of HIV infection among young adults in the U.S.: Results from the Add Health study. *American Journal of Public Health* 2006;96(6):1091-97. [PubMed: 16670236]
- Nowicki K, Snijders TAB. Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association* 2001;96:1077-87.
- Pattison P, Wasserman S. Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology* 1999;52(Pt 2):169-93. [PubMed: 10613111]
- Robins G, Pattison P, Kalish Y, Lusher D. An Introduction to Exponential Random Graph (p^*) Models for Social Networks. *Social Networks* 2007;29(2):173-91.
- Robins G, Pattison P, Wasserman S. Logit models and logistic regressions for social networks: III. Valued relations. *Psychometrika* 1999;64(3):371-94.
- Robins G, Pattison P, Woolcock J. Small and Other Worlds: Global Network Structures from Local Processes. *American Journal of Sociology* 2005;110(4):894-936.
- Rothenberg R, Potterat J, Woodhouse D, Muth S, Darrow W, Klov Dahl A. Social Network Dynamics and HIV Transmission. *AIDS* 1998;12:1529-36. [PubMed: 9727575]
- Salganik M, Heckathorn D. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology* 2004;34:193-239.
- Snijders T. The Degree Variance: An Index of Graph Heterogeneity. *Social Networks* 1981;3:163-74.
- Snijders T. Enumeration and Simulation Methods for 0-1 Matrices with Given Marginals. *Psychometrika* 1991;56(3):397-417.
- Snijders T. Stochastic Actor-oriented Models for Network Change. *Journal of Mathematical Sociology* 1996;21:149-72.
- Snijders, T. The statistical evaluation of social network dynamics. In: Sobel, ME.; Becker, MP., editors. *Sociological Methodology*. Boston: Basil Blackwell; 2001. p. 361-95.
- Snijders T. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. *Journal of Social Structure* 2002;3:2
- Snijders, T. Models for longitudinal social network data. In: Carrington, P.; Scott, J.; Wasserman, S., editors. *Models and Methods in Social Network Analysis*. Cambridge: Cambridge University Press; 2005. p. 215-47.

- Snijders, T.; Pattison, P.; Robins, G.; Handcock, M. New specifications for exponential random graph models. In: Stolzenberg, R., editor. *Sociological Methodology*. Boston, MA: Blackwell; 2006. p. 99-153.
- Snijders, T.; Steglich, C.; Schweinberger, M.; Huisman, M. *Manual for SIENA version 3.2*. Groningen, The Netherlands: University of Groningen; 2007.
- Strauss D, Ikeda M. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association* 1990;85:204-12.
- Sudman S, Kalton G. New Developments in the Sampling of Special Populations. *Annual Review of Sociology* 1986;12:401-29.
- Thompson S. Adaptive Web Sampling. *Biometrics* 2006;62(4):1224-34. [PubMed: 17156298]
- Travers J, Milgram S. An Experimental Study of the Small World Problem. *Sociometry* 1969;32(4):425-43.
- Unger J, Chen X. The role of social networks and media receptivity in predicting age of smoking initiation: A proportional hazards model of risk and protective factors. *Addictive Behaviors* 1999;24:371-81. [PubMed: 10400276]
- Valente T, Watkins S, Jato M, van der Straten A, Tsitol L. Social network associations with contraceptive use among Cameroonian women in voluntary associations. *Social Science and Medicine* 1997;45:1837-43.
- Van Duijn M, Snijders T, Zijlstra B. P2: A Random Effects Model with Covariates for Directed Graphs. *Statistica Neerlandica* 2004;58(2):234-54.
- Van Duijn M, Van Busschback J, Snijders T. Multilevel analysis of personal networks as dependent variables. *Social Networks* 1999;21:187-209.
- Waller, L.; Gotway, C. *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: Wiley Interscience; 2004.
- Wang, P.; Robins, G.; Pattison, P. PNet: Program for the Simulation and Estimation of P* Exponential Random Graph Models. (release Department of Psychology, University of Melbourne; 2008.
- Wang W, Wong G. Stochastic Blockmodels for Directed Graphs. *Journal of the American Statistical Association* 1987;82:8-19.
- Wasserman, S. A Stochastic Model for Directed Graphs With Transition Rates Determined by Reciprocity. In: Schuessler, KF., editor. *Sociological Methodology*. San Francisco: Jossey-Bass; 1979. p. 392-412.
- Wasserman S. Analyzing Social Networks As Stochastic Processes. *Journal of the American Statistical Association* 1980;75:280-94.
- Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press; 1994.
- Wasserman S, Pattison P. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika* 1996;61:401-25.
- Wellman B, Frank K. Network capital in a multilevel world: getting support from personal communities. *Social Capital: Theory and Research* 2001:233-73.
- White, D.; Harary, F. The Cohesiveness of Blocks in Social Networks: Node Connectivity and Conditional Density. In: Becker, MP., editor. *Sociological Methodology*. Boston: Blackwell; 2001. p. 140-48.
- Wolfram, S. *A New Kind of Science*. Wolfram Media; 2002.
- Wong G. Bayesian Models for Directed Graphs. *Journal of the American Statistical Association* 1987;82:140-148.
- Zeggelink E. Dynamics of structure- an individual oriented approach. *Social Networks* 1994;16(4):295-333.
- Zijlstra B, Van Duijn M, Snijders T. The Multilevel p2 Model: A Random Effects Model for the Analysis of Multiple Social Networks. *Methodology* 2006;21:42-47.

Appendix A: R-code for fitting Individual Outcome Model

```
## Functions used in the analysis ##
```

```

geodist <- function(adj) {

#Derive the geodesic distance between the actors

dist <- matrix(0,nr,nr)

matpow <- diag(1,nr)

for (k in 1:nr) {

matpow <- matpow %**% adj

ind <- ifelse(dist>0,1,0)

dist <- dist*ind + k*ifelse(matpow>0,1,0)*(1-ind)

}

ind <- ifelse(dist>0,1,0)

dist <- dist*ind-1*(1-ind) #-1 indicates of not connected

diag(dist) <- 0 #Always 0 distance on the diagonal!

return(dist)

}

like.auto <- function(al,y,x,w) {

#Evaluate log-likelihood function of outcome autocorrelation model

n <- nrow(x)

icov <- diag(1,n) - al*w

z <- icov %**% y

icov <- t(icov) %**% icov

be <- solve(t(x) %**% x) %**% (t(x) %**% z) #coefficients of exogeneous
vbles

in IV model

ri <- z - x %**% be #residuals

std2 <- as.vector(t(ri) %**% ri)/n #variance of outcome

icov <- icov/std2

return(log(det(icov)) - (t(ri) %**% ri)/std2)

}

parSE.auto <- function(al,y,x,w) {

```

```

#Standard errors of outcome autocorrelation model

n <- nrow(x)

p <- ncol(x)

icov1 <- diag(1,n) - al*w

icov <- t(icov1) %*% icov1

z <- icov1 %*% y

be.be <- (t(x) %*% x)

be <- solve(be.be) %*% (t(x) %*% z) #coefficients of x

ri <- z - x %*% be #residuals

std2 <- as.vector(t(ri) %*% ri)/n #variance of outcome

G <- w %*% solve(icov1)

Gxb <- G %*% x %*% be

#The following derivation is based on Harville (1997, p. 309).

cov <- solve(icov)

wsym <- t(w)+w

wsqr <- t(w) %*% w

dicov.al <- 2*al*wsqr-wsym

dicov.alal <- 2*wsqr

t.alal <- cov %*% dicov.alal

t.al <- cov %*% dicov.al

detterm <- sum(diag(t.alal - (t.al %*% t.al)))/2

#Information matrix - extension of expression on page 370 of

#Doreian (1981) to asymmetric W

be.be <- be.be/std2

std2.std2 <- n/(2*(std2^2))

al.be <- (t(Gxb) %*% x)/std2

al.std2 <- sum(diag(G))/std2

al.al <- sum(diag(t(G) %*% G)) + (t(Gxb) %*% Gxb)/std2 - detterm

```

```

Infm <- matrix(0,p+2,p+2)

Infm[1:p,1:p] <- be.be

Infm[p+2,1:p] <- al.be

Infm[1:p,p+2] <- t(al.be)

Infm[p+1,p+1] <- std2.std2

Infm[p+1,p+2] <- al.std2

Infm[p+2,p+1] <- al.std2

Infm[p+2,p+2] <- al.al

#Estimates and covariance matrix

parms <- c(be=be, std2=std2, al=al)

covar <- solve(Infm)

SE <- sqrt(diag(covar))

names(SE) <- names(par)

tval = parms/SE

pval = 2*(1-pt(abs(tval),n-p))

return(list(estimates=parms,SE=SE,tval=tval,pval=pval))

}

like.sar <- function(rho,y,x,w) {

#Evaluate log-likelihood function

n <- nrow(x)

icov <- diag(1,n) - rho*w

icov <- t(icov) %*% icov

be <- solve(t(x) %*% icov %*% x) %*% (t(x) %*% icov %*% y)
#coefficients of

exogeneous vbles in IV model

ri <- y - x %*% be #residuals

std2 <- as.vector(t(ri) %*% icov %*% ri)/n #variance of outcome

icov <- icov/std2

return(log(det(icov)) - t(ri) %*% icov %*% ri)

```

```

}

parSE.sar <- function(rho,y,x,w) {

n <- nrow(x)

p <- ncol(x)

icov1 <- diag(1,n) - rho*w

icov <- t(icov1) %*% icov1

be.be <- (t(x) %*% icov %*% x)

be <- solve(be.be) %*% (t(x) %*% icov %*% y) #coefficients of x

ri <- y - x %*% be #residuals

std2 <- as.vector(t(ri) %*% icov %*% ri)/n #variance of outcome

G <- w %*% solve(icov1)

#The following computation is based on Harville (1997, p. 309).

cov <- solve(icov)

wsym <- t(w)+w

wsqr <- t(w) %*% w

dicov.rho <- 2*rho*wsqr-wsym

dicov.rhorho <- 2*wsqr

t.rhorho <- cov %*% dicov.rhorho

t.rho <- cov %*% dicov.rho

detterm <- sum(diag(t.rhorho - (t.rho %*% t.rho)))/2

#Information matrix - extension of expression on page 366 of

#Waller and Gotway (2004) to asymmetric W

be.be <- be.be/std2

std2.std2 <- n/(2*(std2^2))

rho.std2 <- sum(diag(G))/std2

rho.rho <- sum(diag(t(G) %*% G)) - detterm

Infm <- matrix(0,p+2,p+2)

Infm[1:p,1:p] <- be.be

```



```

Infm[p+1,p+1] <- std2.std2

Infm[p+1,p+2] <- rho.std2

Infm[p+2,p+1] <- rho.std2

Infm[p+2,p+2] <- rho.rho

#Estimates and covariance matrix

parms <- c(be=be, std2=std2, rho=rho)

covar <- solve(Infm)

SE <- sqrt(diag(covar))

names(SE) <- names(par)

tval = parms/SE

pval = 2*(1-pt(abs(tval),n-p))

return(list(estimates=parms,SE=SE,tval=tval,pval=pval))
}

## Code for loading data and fitting model ##

#Install software on R: Can comment out after first use.

install.packages("statnet",

contriburl="http://csde.washington.edu/~handcock")

install.packages("coda")

#Load functions needed for analysis

source("NetAnalFns.r")

#Load network

nr <- 33

adjdata <- scan("H:/Harvard/Christakis/HSORM/clin33correct.txt")

adjdata <- matrix(adjdata,nrow=nr,byrow=T)

adjdata[19,19] <- 0 #Tidy up data (self-connected nodes not possible)

iddata <- seq(1,33,1)

#Load node covariate data

covdata <- scan("H:/Harvard/Christakis/HSORM/nodecov.dat")

```

```

covdata <- matrix(covdata,nrow=nr,byrow=T)

nodecov <- list(male=covdata[,2], whexpert=covdata[,3],
pctwom=covdata[,4],
numsess=covdata[,5], practice=covdata[,6])

#Directed network

adjdir <- ifelse(adjdata>0,1,0)

#Standardize adjacency matrix so that row sums = 1
numalters <- apply(adjdir,1,sum)
scale=as.vector(numalters^(-1))

noalters <- ifelse(is.infinite(scale)==1,1,0)

scale[noalters*seq(1,nr)] <- 0

on=matrix(1,ncol=nr,nrow=1)

wtadjdir <- adjdir * (scale %%% on)

#Compute mean value of dependent variables for directly connected
# actors - i.e., based on the adjacency matrix
hrtalt <- wtadjdir %%% as.vector(covdata$sumhrt)

regdata <-
data.frame
(covdata,hrtalt=hrtalt,noalters=noalters,numalters=numalters)

#Compute scaled geodesic distances having row sums equal to 1

gdist <- geodist(adjdir)

igdist <- gdist^(-1)

diag(igdist) <- 0 #Always 0 distance on the diagonal!

ind <- ifelse(igdist<0,1,0) #Egos with no influential conversations

igdist <- igdist*(1-ind)

sumigeo <- apply(igdist,1,sum)

scale <- as.vector(sumigeo^(-1))

noalters <- ifelse(is.infinite(scale)==1,1,0)

```

```

scale[noalters*seq(1,nr)] <- 0 #Influence set equal to 0 if no
influential

conversations

wtigeo <- igdist * (scale %*% on)

#Augment analysis dataset with additional variables

hrtgeo <- wtigeo %*% as.vector(covdata$sumhrt)

regdata <- data.frame(regdata,hrtgeo=hrtgeo)

on <- as.vector(rep(1,nr))

x <- as.matrix(cbind(on,regdata[,c("male","pctwom","numalters")]))

#Use maximum likelihood to obtain MLEs for autoregressive outcomes and
network autocorrelation (SAR) models

# - uses optim optimization function in R.

strtval <- 0

#Fit autoregressive outcomes model

#Use scaled adjacency matrix as weight matrix

mle <- optim(par=strtval, fn=like.auto, gr=NULL, method = "BFGS",
control=list(fnscale=-1, trace=6, maxit=100, reltol=1e-16),
hessian = TRUE, y=regdata$sumhrt,x=x,w=wtadmdir)

#Compute estimates of all parameters and standard errors

Auto.adj <- parSE.auto(mle$par,y=regdata$sumhrt,x=x,w=wtadmdir)

#Use scaled geodesic matrix as weight matrix

mle <- optim(par=strtval, fn=like.auto, gr=NULL, method = "BFGS",
control=list(fnscale=-1, trace=6, maxit=100, reltol=1e-16),
hessian = TRUE, y=regdata$sumhrt,x=x,w=wtigeo)

#Compute estimates of all parameters and standard errors

Auto.geo <- parSE.auto(mle$par,y=regdata$sumhrt,x=x,w=wtigeo)

#Fit corresponding SAR models

#Use scaled adjacency matrix as weight matrix

mle <- optim(par=strtval, fn=like.sar, gr=NULL, method = "BFGS",

```

```

control=list(fnscale=-1, trace=6, maxit=100, reltol=1e-16),

hessian = TRUE, y=regdata$sumhrt,x=x,w=wtadjdir)

#Compute estimates of all parameters and standard errors

SAR.adj <- parSE.sar(mle$par,y=regdata$sumhrt,x=x,w=wtadjdir)

#Use scaled geodesic matrix as weight matrix

mle <- optim(par=strtval, fn=like.sar, gr=NULL, method = "BFGS",

control=list(fnscale=-1, trace=6, maxit=100, reltol=1e-16),

hessian = TRUE, y=regdata$sumhrt,x=x,w=wtigeo)

#Compute estimates of all parameters and standard errors

SAR.geo <- parSE.sar(mle$par,y=regdata$sumhrt,x=x,w=wtigeo)

#Print out all results

print(data.frame(Auto.adj))

print(data.frame(Auto.geo))

print(data.frame(SAR.adj))

print(data.frame(SAR.geo))

##Alternative code using lnam function in Carter Butt's sna package.
##

library(sna)

library(numDeriv)

lnam1.adj <- lnam(regdata$sumhrt,x,wtadjdir)

lnam1.geo <- lnam(regdata$sumhrt,x,wtigeo)

lnam2.adj <- lnam(regdata$sumhrt,x,NULL,wtadjdir)

lnam2.geo <- lnam(regdata$sumhrt,x,NULL,wtigeo)

#Print out all results

summary(lnam1.adj)

summary(lnam1.geo)

summary(lnam2.adj)

summary(lnam2.geo)

```

Appendix B: R-code for Relational Data Model

```
#Install software on R: Can comment out after first use.

install.packages("statnet",

contriburl="http://csde.washington.edu/~handcock")

install.packages("coda")

#Attached libraries each time use StatNet

library(statnet)

library(coda)

#Load network

nr <- 33

adjdata <- scan("clin33correct.txt")

adjdata <- matrix(adjdata,nrow=nr,byrow=T)

adjdata[19,19] <- 0 #Tidy up data (self-connected nodes not possible)

iddata <- seq(1,33,1)

#Load node covariate data

covdata <- scan("nodecov.dat")

covdata <- matrix(covdata,nrow=nr,byrow=T)

nodecov <- list(male=covdata[,2], whexpert=covdata[,3],

pctwom=covdata[,4],

numsess=covdata[,5], practice=covdata[,6])

#Make directed network

pnet <- network(adjdata,directed=TRUE,matrixtype="adjacency",

vertex.attr=nodecov,

vertex.attrnames=c("male","whexpert","pctwom","numsess","practice",

"bcma","bima","bpp","wnhlth","numcat","pctcat"))

#Plot directed network

plot(pnet,mode="fruchtermanreingold",displaylabels=T)

#Fit models to directed network
```

```

modella <- ergm(pnet~edges)

modellb <- ergm(pnet~edges + mutual)

modellc <- ergm(pnet~edges + mutual + ttriad, theta0="MPLE",
MPLEonly=TRUE)

modelld <- ergm(pnet~edges + mutual + receivercov("whexpert") +
receivercov("pctwom") + receivercov("numsess"))

#Evaluate Goodness of fit with respect to distance and degrees
dist.gof <- gof(modelld~distance, nsim=10, verbose=T)
plot(dist.gof)

ideg.gof <- gof(modelld~idegree, nsim=10, verbose=T)
plot(ideg.gof)

odeg.gof <- gof(modelld~odegree, nsim=10, verbose=T)
plot(odeg.gof)

#Define network using mutual ties.
adjmut <- ifelse(adjdata>0,1,0) + ifelse(t(adjdata)>0,1,0)
adjmut <- ifelse(adjmut>0,1,0); #One possible definition of mutual tie
#Make undirected network
pnetmut <- network(adjmut,directed=FALSE,matrixtype="adjacency",
vertex.attr=nodecov,
vertex.attrnames=c("male","whexpert","pctwom","numsess","practice",
"bcma","bima","bpp","wnhlth","numcat","pctcat"))
#Plot undirected network
plot(pnetmut,mode="fruchtermanreingold",displaylabels=T)
#Fit models to undirected network
model2a <- ergm(pnetmut~edges)
model2b <- ergm(pnetmut~edges + gwesp(1.2,fixed=T),
burnin=10000, MCMCsamplesize=10000, interval=100, maxit=3)
model2c <- ergm(pnetmut~edges + gwesp(1.2,fixed=F) + kstar(2:3),

```



```
burnin=10000, MCMCsamplesize=10000, interval=100, maxit=3)

model2d <- ergm(pnetmut~edges + gwesp(1.2,fixed=T) +
nodematch("male",diff=F) + nodematch("practice",diff=F),
burnin=10000, MCMCsamplesize=10000, interval=100, maxit=3)

model2e <- ergm(pnetmut~edges + gwesp(1.2,fixed=T) +
nodematch("male",diff=F) + nodematch("practice",diff=T),
burnin=10000, MCMCsamplesize=10000, interval=100, maxit=3)

#Test goodness of fit with respect to espartners statistic
esppart.gof <- gof(model2e~espartners,nsim=10,verbose=T)

plot(esppart.gof)
```

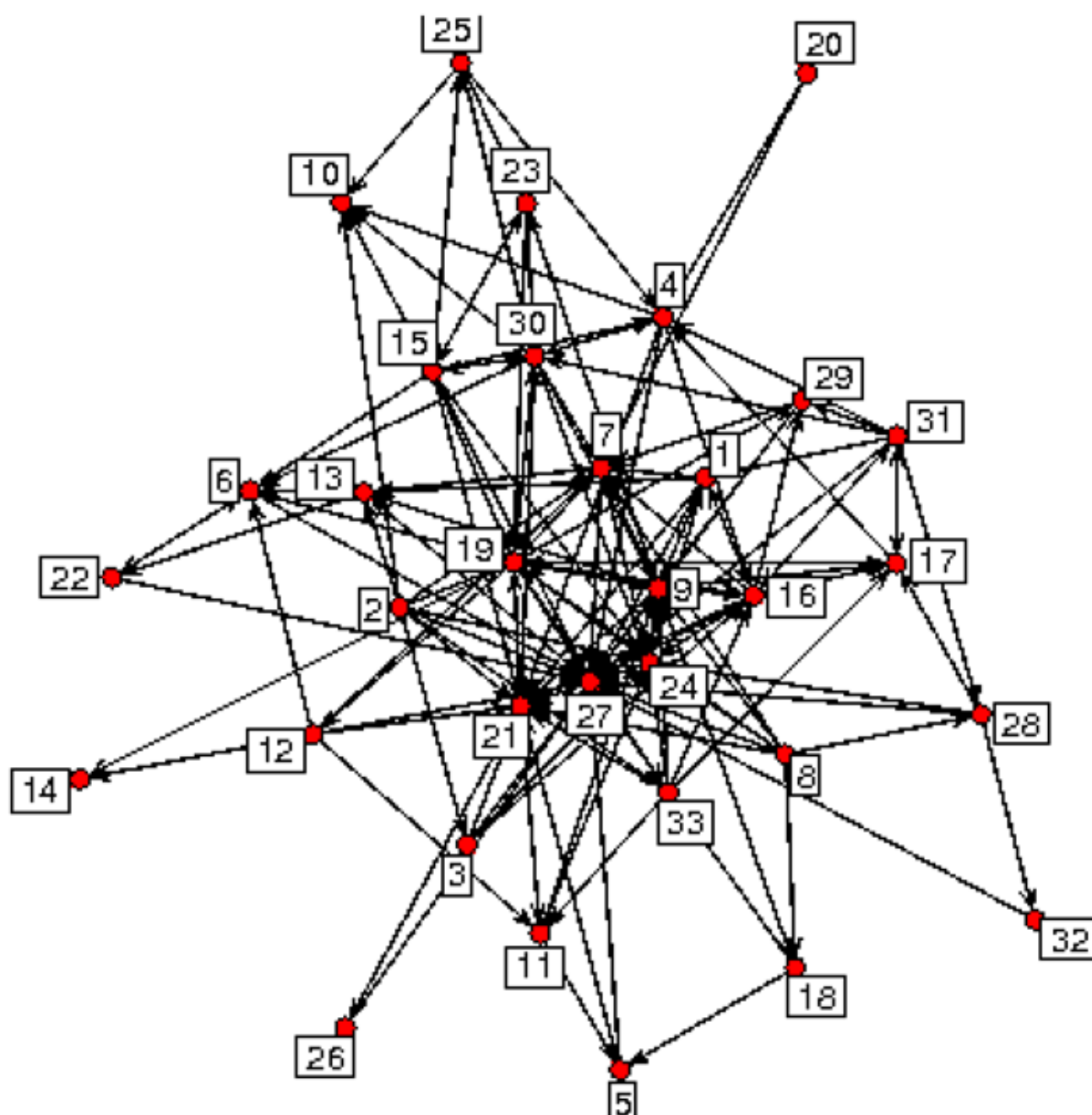
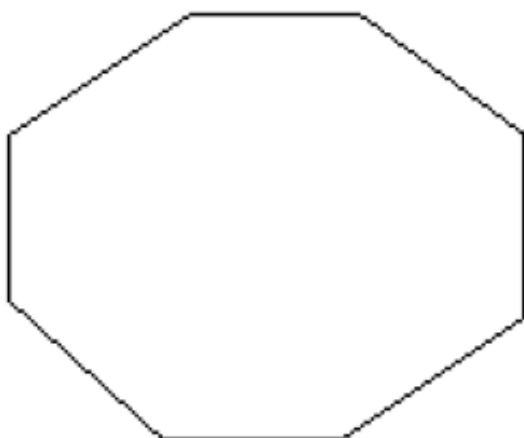


Figure 1.
 Directed physician influence network when ties defined as 1 or more influential discussions.
 Note: With the exception of the omission of the directional arrows, the graph for the corresponding undirected network is the same as that above.

Circle Network



Star Network

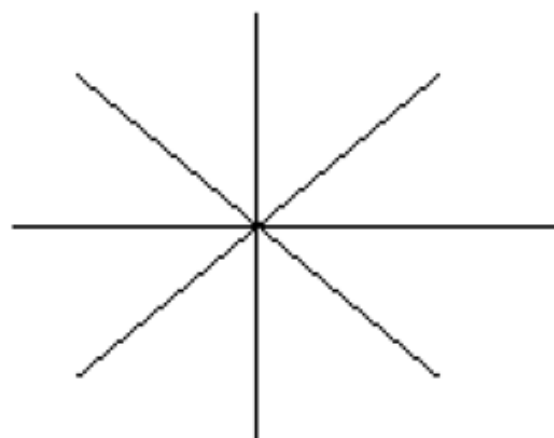


Figure 2.
Circle and star networks

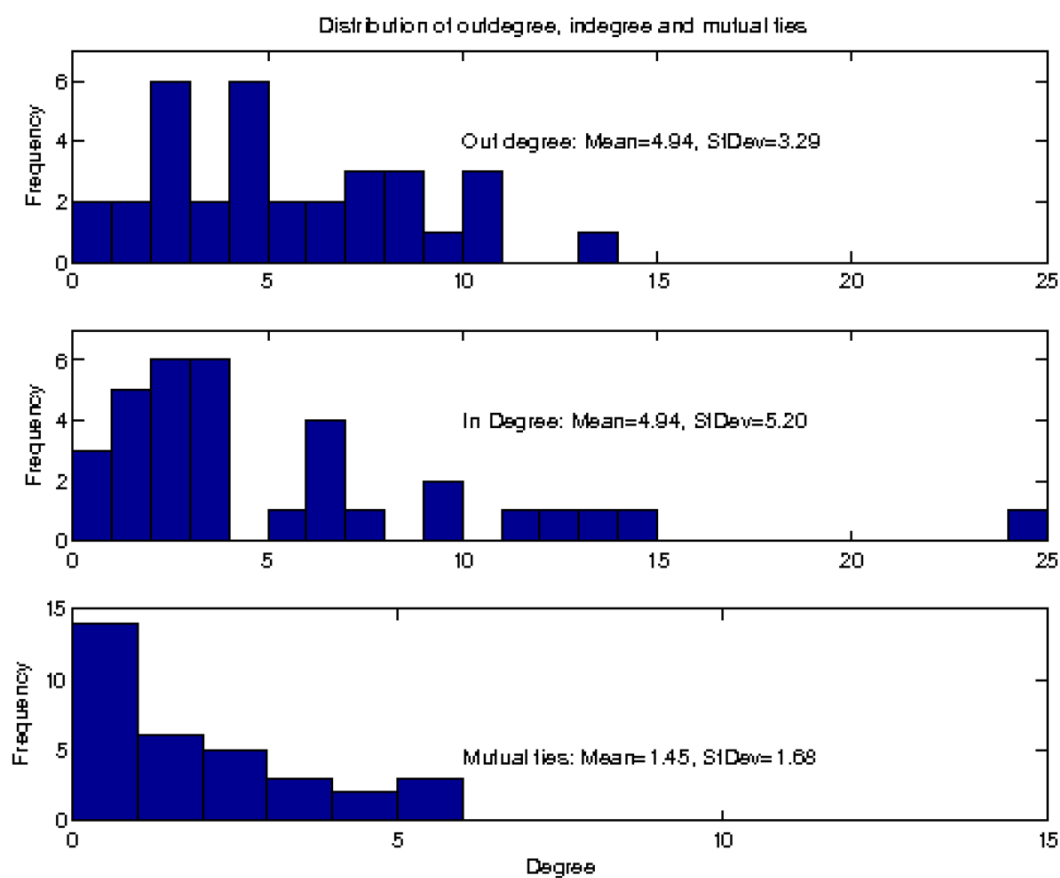


Figure 3.
Degree distributions for directed physician influence network

Table 1

Adjacency matrix (first 10 actors) for directed physician influence network

Start/End	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	1	0	1	0
2	0	0	1	0	0	0	1	0	0	1
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	1	0	1	0
9	1	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	0	0	0	0

Note: Cells containing 1 and 0 denote the presence and absence of a tie respectively.

Table 2

Node-level statistics for directed physician influence network

Node	Degree		Betweenness Centrality
	Out	In	
1	7	3	0.39
2	9	0	0
3	4	1	0.09
4	5	6	4.39
5	2	3	1.18
6	3	6	1.18
7	2	12	0.78
8	8	0	0
9	10	11	27.64
10	0	5	0
11	4	3	0.61
12	6	1	0.09
13	4	6	3.36
14	0	2	0
15	10	3	4.31
16	10	9	10.64
17	4	6	6.22
18	2	2	0.55
19	13	9	14.67
20	2	0	0
21	7	14	15.46
22	2	2	0
23	4	3	4.49
24	6	13	4.04
25	5	1	0
26	1	1	0
27	2	24	20.1
28	4	2	3.02
29	3	3	0.08
30	8	7	7.68
31	8	2	6.11
32	1	1	0
33	7	2	0.44

Table 3

Paths of length 2 (first 10 actors) for directed physician influence network

Start/End	1	2	3	4	5	6	7	8	9	10
1	2	0	0	0	1	1	4	0	3	0
2	0	0	0	0	1	2	3	0	4	0
3	1	0	0	0	1	0	2	0	3	0
4	1	0	0	1	1	1	2	0	3	1
5	0	0	0	0	1	0	0	0	1	0
6	0	0	0	0	0	2	1	0	1	0
7	0	0	0	0	0	1	1	0	1	0
8	1	0	0	0	2	1	3	0	3	0
9	1	0	0	1	2	1	4	0	5	0
10	0	0	0	0	0	0	0	0	0	0

Table 4

Paths of length 3 (first 10 actors) for directed physician influence network

Start/End	1	2	3	4	5	6	7	8	9	10
1	6	0	0	4	5	8	18	0	20	1
2	8	0	0	4	6	8	20	0	23	2
3	5	0	0	3	3	5	13	0	13	1
4	4	0	0	4	2	6	13	0	12	3
5	1	0	0	1	1	3	5	0	5	1
6	1	0	0	0	0	2	2	0	2	0
7	1	0	0	0	0	2	2	0	1	0
8	6	0	0	5	5	8	20	0	23	2
9	10	0	0	6	4	12	23	0	23	4
10	0	0	0	0	0	0	0	0	0	0

Table 5

Geodesic distances (first 10 actors) for directed physician influence network

Start/End	1	2	3	4	5	6	7	8	9	10
1	0	-1	-1	3	2	2	1	-1	1	3
2	3	0	1	3	2	2	1	-1	2	1
3	2	-1	0	3	2	3	2	-1	2	3
4	2	-1	-1	0	2	2	2	-1	2	1
5	3	-1	-1	3	0	3	3	-1	2	3
6	3	-1	-1	4	4	0	2	-1	2	5
7	3	-1	-1	4	4	2	0	-1	2	5
8	2	-1	-1	3	2	2	1	0	1	3
9	1	-1	-1	2	2	2	1	-1	0	3
10	-1	-1	-1	-1	-1	-1	-1	-1	-1	0

Note: values of -1 denote "infinite" geodesic distances.

Table 6

Node-level statistics for undirected physician influence network

Node	Degree	Closeness Centrality	Eigenvector Centrality
1	8	0.54	0.18
2	9	0.58	0.18
3	5	0.51	0.12
4	10	0.58	0.16
5	4	0.48	0.08
6	7	0.55	0.12
7	13	0.63	0.24
8	8	0.55	0.18
9	16	0.67	0.29
10	5	0.45	0.07
11	6	0.5	0.12
12	7	0.54	0.13
13	7	0.54	0.13
14	2	0.42	0.04
15	10	0.59	0.18
16	13	0.63	0.26
17	7	0.53	0.13
18	4	0.5	0.08
19	17	0.68	0.3
20	2	0.43	0.04
21	17	0.68	0.29
22	3	0.47	0.05
23	5	0.48	0.09
24	15	0.65	0.29
25	5	0.43	0.06
26	2	0.47	0.06
27	24	0.8	0.35
28	6	0.51	0.1
29	6	0.53	0.14
30	12	0.6	0.2
31	8	0.55	0.14
32	2	0.46	0.04
33	9	0.56	0.2

Table 7

Results for individual-level analyses of directed physician influence network

Model 5A: Autoregressive outcomes, adjacency weight matrix				
Term	Estimate	Std. Error	t-value	Pr(> t)
Intercept	16.078	3.897	4.126	0.000
Male	−0.206	2.331	−0.088	0.930
PercentWomenPatients	−0.052	0.048	−1.089	0.285
Outdegree	−0.353	0.198	−1.780	0.086
$\alpha(\text{AdjacencyWeightedHRT})$	0.277	0.169	1.635	0.113

Model 5B: Autoregressive outcomes, geodesic weight matrix				
Intercept	16.246	3.878	4.190	0.000
Male	−0.986	2.422	−0.407	0.687
PercentWomenPatients	−0.063	0.049	−1.289	0.207
Outdegree	−0.402	0.207	−1.940	0.062
$\alpha(\text{GeodesicWeightedHRT})$	0.360	0.199	1.806	0.081

Model 6A: Network Autocorrelation, adjacency weight matrix				
Intercept	16.471	4.051	4.066	0.000
Male	1.324	2.154	0.615	0.544
PercentWomenPatients	−0.023	0.045	−0.506	0.617
Outdegree	−0.209	0.189	−1.108	0.277
$\rho(\text{AdjacencyWeightedHRT})$	0.026	0.297	0.089	0.929

Model 6B: Network Autocorrelation, geodesic weight matrix				
Intercept	16.366	4.068	4.023	0.000
Male	1.371	2.160	0.635	0.530
PercentWomenPatients	−0.019	0.045	−0.423	0.676
Outdegree	−0.229	0.188	−1.220	0.232
$\rho(\text{GeodesicWeightedHRT})$	−0.233	0.771	−0.303	0.764

Note: The residual standard errors in models 5A, 5B, 6A, and 6B are 3.264, 3.248, 3.398, and 3.393 respectively. Outdegree is the number of distinct physicians cited by the focal physician as a partner in at least one influential conversation about women's health.

Table 8

Estimates for ERGM models fitted to directed physician influence network

Model Term	Estimate	SE	p-value	Model Statistics		
				MCMC SE	Deviance	BIC
Edges					555.4	915.5
edges	-1.701	0.085	<1e-04	NA		
Edges + Mutuality					589.3	888.5
edges	-1.952	0.107	<1e-04	0.002		
mutual	1.187	0.280	<1e-04	0.006		
Edges + Mutuality + Receiver covariates					695.4	803.3
edges	-4.894	0.443	< 1e-04	0.013		
mutual	1.244	0.282	< 1e-04	0.008		
receivercov.whexpert	1.218	0.214	< 1e-04	0.008		
receivercov.pctwom	0.029	0.005	< 1e-04	0.000		
receivercov.numssess	0.128	0.038	0.001	0.001		
Edges + Mutuality + Transitive Triad					744.3	740.5
edges	-3.158	0.169	<1e-04	NA		
mutual	0.407	0.236	0.085	NA		
transitive triad	0.336	0.030	<1e-04	NA		

Note: MCMC was not needed to fit the Edges model as logistic regression yields the MLE. The Edges + Mutuality + Transitive Triad model was fitted using pseudo-likelihood and so the standard errors may not be reliable; the deviance based on the pseudo-likelihood function is reported for that model.

Table 9

Estimates for ERGM models fitted to undirected physician influence network

Model Term	Estimate	SE	p-value	Model Statistics		
				MCMC SE	Deviance	BIC
Edges					127.4	610.8
edges	1.049	0.099	<1e-04	NA		
Edges + gwespp					291.2	453.4
edges	4.279	0.550	<1e-04	0.010		
gwepp(1.2)	0.926	0.146	<1e-04	0.003		
Edges + gwespp + kstar(2:3)					300.3	463.0
edges	3.253	1.266	0.010	0.026		
gwepp(ρ)	0.662	0.263	0.012	0.004		
ρ	0.930	0.151	<1e-04	0.002		
kstar2	0.045	0.122	0.711	0.002		
kstar3	0.006	0.013	0.624	0.000		
Edges + gwespp + Node match covariates (pooled)					302.8	454.3
edges	3.964	0.782	< 1e-04	0.017		
gwepp(1.2)	0.806	0.192	< 1e-04	0.004		
nodematch.male	0.532	0.172	0.002	0.006		
nodematch.clinic	0.008	0.173	0.962	0.006		
Edges + gwespp + Node match covariates (not pooled)					315.3	460.5
edges	4.340	0.540	<1e-04	0.028		
gwepp(1.2)	0.872	0.196	<1e-04	0.009		
nodematch.male	0.433	0.469	0.357	0.040		
nodematch.clinic.1	1.399	0.917	0.127	0.025		
nodematch.clinic.2	0.042	0.768	0.956	0.146		

Model Term	Model Statistics					
	Estimate	SE	p-value	MCMC SE	Deviance	BIC
nodematch.clinic.3	2.989	1.884	0.113	1.393		
nodematch.clinic.4	2.197	0.520	<1e-04	0.050		

Note: MCMC was not needed to fit the Edges model as logistic regression yields the MLE.