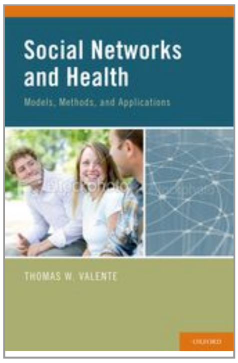


Social Networks and Health: Models, Methods, and Applications
Thomas W. Valente

Print publication date: 2010
Print ISBN-13: 9780195301014
Published to Oxford Scholarship Online: May 2010
DOI: 10.1093/acprof:oso/9780195301014.001.0001

Exponential Random Graph Models, P* and Actor Oriented Models

Thomas W. Valente

DOI:10.1093/acprof:oso/9780195301014.003.0009

Abstract and Keywords

This chapter describes in non-technical terms how researchers can determine whether empirical networks exhibit certain structural properties (centrality, triadic transitivity, etc.) using a statistical test. These models are referred to as exponential random graph models (ERGM), named after the technique used to generate simulated networks, which can then be compared to the observed network in order to statistically assess network properties. The chapter provides an exposition of the actor oriented co-evolution model, which extends the cross-sectional ERGM framework to model longitudinal processes. These models can be used to determine what behaviors drive social network evolution, and whether social relationships influence behavioral changes. Public health examples are used throughout to illustrate concepts.

Keywords: ERGM, P*, Exponential Random Graph Models, network evolution, network statistics

This chapter describes in nontechnical terms how researchers can determine whether empirical networks exhibit certain structural properties (centrality, triadic transitivity, etc.) using a statistical test. These models are referred to as exponential random graph models (ERGMs), named after the technique used to generate simulated networks, which can then be compared to the observed network to statistically assess network properties. Here we try to provide a simple introduction so researchers can seek out more advanced materials and tutorials to learn how the analysis is conducted (Burk et al., 2007; Harrigan, 2009; Snijders, 2001; Snijders et al., 2007).

ERGM is derived from the P* model, which provided the basis for early statistical analysis of networks. The critical innovation in the P* model was the use of logistic regression analysis to determine the factors (individual characteristics and network metrics) associated with a link between two nodes. This chapter first provides an introduction to the dyadic approach used for many years. The chapter then describes the exponential random graph model (ERGM) approach used primarily with cross-sectional data. The chapter then provides an exposition of the actor-oriented co-evolution model, which extends the cross-sectional ERGM framework to model longitudinal processes. These models can be used to determine what behaviors drive social network evolution and whether social relationships influence behavioral changes. This **(p.152)** actor-oriented co-evolution model is very flexible, permitting many kinds of hypothesis tests. The chapter includes some public health examples.

In Part II of this book, many social network measures were presented that treated social networks as static configurations. The network did not change and behaviors were often described as something distributed on the network. Networks were described by creating indicators for their structure such as density or centralization and network members were characterized by their positions in the network, such as their centrality. Individual and network measures could then be correlated with individual or group behavior. For example, an individual's centrality (in-degree) was found to correlate with smoking behaviors and density scores were associated with the rate of behavioral change. The language used so far often conveyed the concept of a fixed network with an idea or a behavior spreading through it much like cars moving on roads.

In reality, however, networks change. Individuals make new friends and lose track of old ones. Behaviors also change, of course, with some people trying a new activity for the first time while others quit after some experience. Sometimes relationships change because of the behavior change, while in other cases people change behaviors because their peers have changed theirs. For example, if an adolescent thinks smoking is cool, he or she may want to form friendships with smokers. Alternatively, an adolescent may start smoking because his or her friends have started. Consequently, researchers have been interested in finding ways to analyze network and behavior changes simultaneously. To study the co-evolution of behavior and social networks, researchers have developed models that attempt to estimate the probability that individuals form ties in a network based on the existing relationships and behaviors. A second limitation of much prior research on network effects is that statistical associations between network exposure (i.e., behavior of one's friends), network indicators (e.g., centrality, density), and behaviors (attributes of the individual) have not completely accounted for dependencies between the actors. For example, suppose a statistical analysis shows that people who smoke are more likely than nonsmokers to have smoking friends. This association may be a product of both the focal person and the network alters being connected to a third person who is also a smoker.

Until the relatively recent ERGM and co-evolution models, many people were unsatisfied with network analysis research without any indication of whether the network measures were expected and normal or whether one network structure could be considered to be “better” than another. The quest to develop statistical models appropriate for network data spans many decades (Holland & Leinhardt, 1979) and several teams of researchers. Fortunately, much progress has been made and a set of tools emerged that enable researchers to determine whether a given network can be considered to have **(p.153)** certain structural tendencies (e.g., Is the network transitive?), whether behaviors drive network structures (e.g., Are smokers more likely to be selected as friends?), and how network relations influence behaviors (e.g., Do popular individuals have a greater influence on their friends?) over time.

Estimating the Link

Conceptually, to test the probability of observing network as a function of the structural characteristics within the network (e.g., number of ties, reciprocal ties, transitive triads, etc.) was conducted by estimating the probability of a link. Early work used maximum pseudo-likelihood estimation (MPLE) to estimate the contribution of relevant structural properties to the observed network. Essentially, this approach uses logistic regression to estimate whether the property increases or decreases the likelihood of observing a tie between two nodes, while accounting for other structural tendencies in the network. For example, an analysis including two parameters for density and reciprocity where the estimate for reciprocity is negative would indicate that there is a tendency for ties not to be reciprocated after controlling for the total number of ties observed within the network.

A useful introduction to MPLE for estimating network parameters was provided by Crouch, Wasserman, and Contractor (1998); also see Anderson et al., 1999) in which they illustrate the technique using both hypothetical and empirical data. In their hypothetical example, a small network of six nodes with 12 links is proposed that has one binary attribute distributed within it. Crouch and others (1998) explain that this network consists of 12 ties and 18 non-ties and the matrix is reshaped to a vector (a column with 30 entries) of 1s and 0s. Additional vectors are stored next to the link vector that represent whether the two nodes share the attribute, whether the tie should exist under conditions of mutuality, whether it should exist if mutuality occurs within the attribute, and whether the tie should exist under conditions of transitivity. In other words, variables are constructed that represent each tie's contribution to the structural properties of interest and, using logistic regression, model parameters are estimated that allow one to test whether these properties contribute significantly to the presence or absence of a tie.

Vectorizing the Matrix

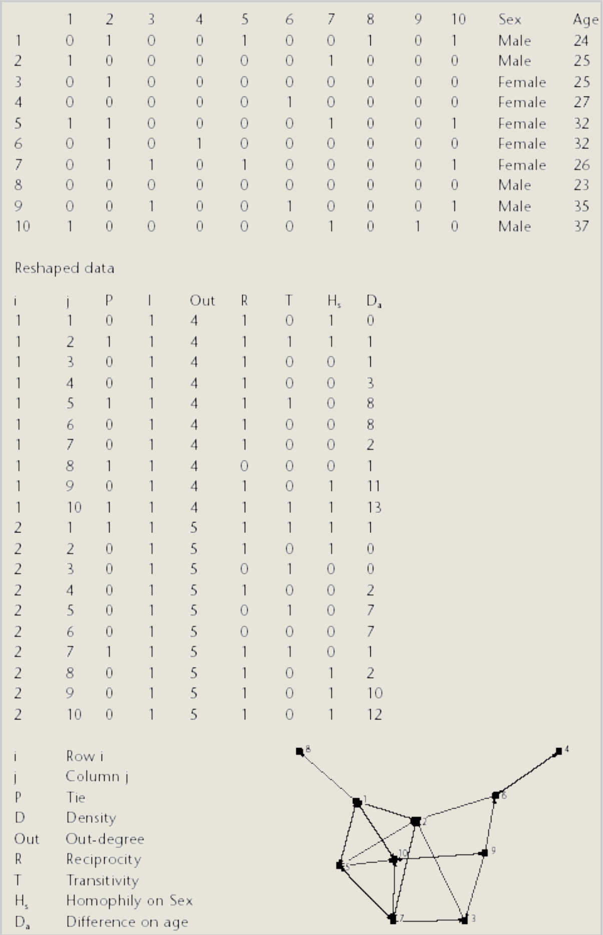
As mentioned on the estimating a link section, the process of estimating an network effects originally involved vectorizing the matrix. The matrix of *i*-to-*j* links can be converted to a vector; that is, the rows of the network **(p.154)** are stacked under one another to make one big column in which each row indicates the value of a specific *i* and *j* linkage, whether 0 or 1. This now becomes a vector that can be treated like any other vector, and we can perform a (logistic) regression to determine the things associated with elements of the vector being 1s rather than 0s. The link vector is the sequence of 0s and 1s representing who is connected to whom reformatted as a column. This link vector, the links and nonlinks in the network, is the dependent variable to be predicted with the understanding that the row entries are not independent observations. They are not independent because the same person contributes multiple cases to the dataset. This dependence violates the assumption of independence in normal regression analysis. Multilevel models (also known as random effects models) have been developed to cope with this nonindependence in survey research, and the network effects were developed to specifically model dependence in the network.

Maximum pseudo-likelihood estimates (MPLE) work by analyzing the factors associated with an element in the link vector being a 1 rather than being a 0. Node characteristics and structural properties are merged with the link vectors and statistical analysis can be conducted to determine if any of these are associated with a link. In sum, the network is converted to a vector (column) in which each element indicates whether there is a link between two nodes. This dyadic relational data is the outcome to be predicted. Vectors for the node attributes and vectors for the dependencies (network structure) are merged with the dyadic relational data. Then a logistic regression is calculated with the dyadic relational data (link vector) as the dependent variable and the other vectors as independent.

Figure 9–1 provides an illustration of how data can be reshaped for statistical analyses using a hypothetical group of 10 people with two attributes, sex and age. For example, person 1 has 4 outgoing ties, is male and 24 years old. To analyze these data, the data are converted to dyadic so each observation is the relationship between two people. The first two columns index the rows and columns to keep track of the people so the first 10 cases are person one's outgoing relations. P is the column indicating whether there is a link between the two people. Out is a count of the outdegree (number of nominations) for that person. R indicates whether the relationship between the two people is reciprocated (mutual) and T indicates whether the pair is in a triadic relationship. H_s is a variable indicating whether the pair are the same (homophilous) on sex. D_a is a variable indicating the degree of difference between the pair on age.

The statistical analysis then estimates whether these constructed variables (D, R, T, H_s, and D_a) are associated with P, the link between the two people. (Reflexive, or relations with oneself, are removed.) Of course these are not the only, nor necessarily the best variables to be constructed from this type **(p.155)**

(p.156) of data. For example, one might want to estimate whether the two people are in a transitive triad in addition to any triadic relationship. Or the researcher might want to estimate the effect of in-degree nominations. And of course with longitudinal data on the same people over time many types of interactions can be created. The conceptual point to understand is that the relationships naturally viewed as a matrix can be converted to dyadic relationships and variables indicating aspects of those relationships are then associated with each one. These challenges led to the development of exponential random graph models (ERGM).



	1	2	3	4	5	6	7	8	9	10	Sex	Age
1	0	1	0	0	1	0	0	1	0	1	Male	24
2	1	0	0	0	0	0	1	0	0	0	Male	25
3	0	1	0	0	0	0	0	0	0	0	Female	25
4	0	0	0	0	0	1	0	0	0	0	Female	27
5	1	1	0	0	0	0	1	0	0	1	Female	32
6	0	1	0	1	0	0	0	0	0	0	Female	32
7	0	1	1	0	1	0	0	0	0	1	Female	26
8	0	0	0	0	0	0	0	0	0	0	Male	23
9	0	0	1	0	0	1	0	0	0	1	Male	35
10	1	0	0	0	0	0	1	0	1	0	Male	37

Reshaped data

i	j	P	I	Out	R	T	H_s	D_a
1	1	0	1	4	1	0	1	0
1	2	1	1	4	1	1	1	1
1	3	0	1	4	1	0	0	1
1	4	0	1	4	1	0	0	3
1	5	1	1	4	1	1	0	8
1	6	0	1	4	1	0	0	8
1	7	0	1	4	1	0	0	2
1	8	1	1	4	0	0	0	1
1	9	0	1	4	1	0	1	11
1	10	1	1	4	1	1	1	13
2	1	1	1	5	1	1	1	1
2	2	0	1	5	1	0	1	0
2	3	0	1	5	0	1	0	0
2	4	0	1	5	1	0	0	2
2	5	0	1	5	0	1	0	7
2	6	0	1	5	0	0	0	7
2	7	1	1	5	1	1	0	1
2	8	0	1	5	1	0	1	2
2	9	0	1	5	1	0	1	10
2	10	0	1	5	1	0	1	12

i	Row i
j	Column j
P	Tie
D	Density
Out	Out-degree
R	Reciprocity
T	Transitivity
H_s	Homophily on Sex
D_a	Difference on age

Figure 9–1. Illustration of data re-shaping for maximum pseudo-likelihood estimation (MPLE) analysis. A hypothetical network of 10 people with sex and age attribution.

Exponential Random Graph Models (ERGM)

ERGMs provide a statistical analysis of a network that can serve two functions: (1) to determine whether network structural properties such as transitivity occur in a network more than expected by chance and (2) to determine whether there is an association between network links and behavior. It should be stressed that ERGMs are used for cross-sectional data. The statistical test measures the likelihood that the observed network could have emerged by chance. Specifically, it can be determined whether this observed network is a function of properties based on the algorithms and materials presented in Chapters 5 through 8. The algorithms and indicators presented in Chapters 5 through 8 are used to calculate the density, degree of reciprocity, transitivity, clustering, and so on that exist in that network. They do not, however, empirically assess the extent to which these properties exist, given the density and other lower-order dependencies. For example, in any network, there is likely to be some reciprocity just by chance, and in a network with more links (greater density) there will be more reciprocity just by chance. To determine if there is a tendency toward reciprocity in a network, a statistical model should control for the density of the network. ERGM provides a way to determine whether network properties occur by chance as a result of other network properties or whether the properties so measured are unlikely given other parameters of the network.

The second primary use of ERGMs, and perhaps the more substantive, is the ability to incorporate nodal attributes in model estimation. Nodal attributes are characteristics that might influence the formation or dissolution of a tie, such as whether adolescent boys are more likely to form friendships with boys than with girls. The node attribute's sex then is an important determinant of network structure and ERGMs can determine whether this tendency is exhibited in the friendship network. In addition to antecedent characteristics such as sex, researchers will often want to include a behavioral outcome variable such as smoking. For example, ERGMs can determine **(p.157)** whether friendships are more likely to form among teens when they have similar smoking status.

In sum, ERGM analysis can be used for (at least) three functions: (1) describe a network in terms of its structural properties; (2) determine if individual attributes (or node characteristics) are associated with network structural properties; and (3) determine if individual attributes are associated with behaviors controlling for items 1 and 2. So, for example, an ERGM can be estimated to determine if ties are reciprocated, if that reciprocation is greater than expected by chance, and if it is associated with sex (boys more likely to reciprocate friendships with boys). We might then include a behavior in the model to determine if smoking is more likely among friends controlling for sex and reciprocity.

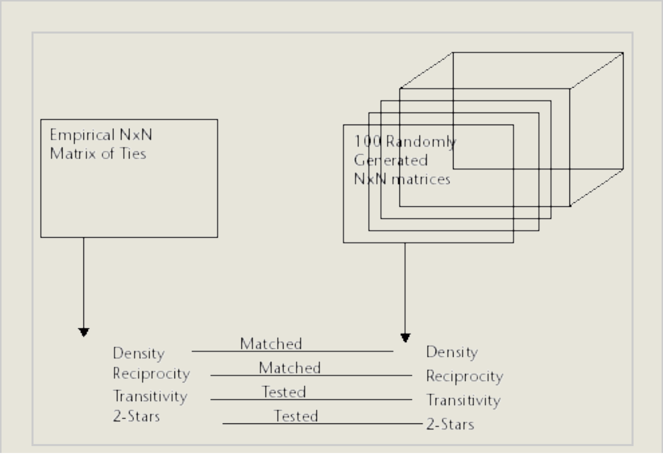
Simulation

The MPLE approach has been shown to lead to biased estimators (van Duijn, et al., 2009) and thus reflects an approximation. When the dependence among networks ties is not strong, then the pseudo-likelihood estimates will be more accurate. However, recent efforts using maximum likelihood approaches to estimate the structural tendencies in the observed network leads to more accurate estimation. This approach is based on Markov chain Monte Carlo (MCMC) techniques by which a distribution of networks are simulated, parameter estimates are obtained by comparing the observed network with the simulated networks, and this process is repeated until there is little change in the parameter estimates (Robins et al., 2007).

The simulations are run thousands of times to generate a distribution of networks based on the characteristics of the empirical networks. This distribution indicates the possible networks given the characteristics of an empirical one. For example, the distribution indicates the degree of reciprocity in a set of randomly generated networks that have the same density as the observed network. The structural properties of interest are parameters of higher order structure than those used to generate the distribution. Each parameter estimated from the randomly simulated networks is compared to the parameters found in the empirical network. If the empirical network parameter is different than the average calculated from the simulated networks then the researcher can conclude that the empirical network has the property, or more accurately, exhibits a tendency for the property.

So, for example, suppose an empirical network has 100 nodes and 500 links for a density of 5%. We might wonder if there is a tendency for reciprocity in the empirical network (if A named B, was B more likely to name A?). To determine if there is a tendency for reciprocity, hundreds or thousands **(p.158)** of networks are generated with 100 nodes and 500 links. The mean of the simulated distribution of reciprocal ties are calculated and then compared to the value in the real network. If the reciprocity of the empirical network differs from the mean reciprocity in the simulated networks more than it would be expected to by chance, then it can be concluded that there is a tendency for reciprocity in the empirical network. Figure 9–2 diagrams some of the parameters calculated during the ERGM simulation process. Researchers may optionally elect to include more parameters to guide the simulation. The network properties that generate the simulation are matched (e.g., density and reciprocity) and then higher order structural properties are compared to determine if they occur at a rate greater than expected by chance given these lower-order parameters.

It is important to realize that structural properties of networks, such as density, reciprocity, transitivity, and so on, are considered hierarchically. A test for transitivity also needs to test for density and reciprocity. In other words, a test for transitivity also generates simulated networks based on the empirical density and reciprocity. The simulated networks will have an average density and reciprocity similar to that of the empirical network because that was the basis for the simulation. The analysis then answers the question: Does the amount of transitivity in the empirical network match that of the



(p.159) simulated networks, or are they different? If different, then the observed network can be characterized as having transitivity above or below what would be expected by chance in networks of this size, density, and reciprocity.

Figure 9–2. The empirical network is described by its density, reciprocity, transitivity, and other structural indicators. Random networks are generated which have these same structural characteristics for the control structures and those which are not generated explicitly in the simulation are tested.

Of course, things can get more complicated than this for several reasons. First, it is hard to know how many networks need to be generated to get a valid distribution of parameters. Second, as networks get bigger, it becomes increasingly difficult to generate lots of networks and calculate multiple structural properties. Third, as node attributes are added to the analysis (e.g., gender, drug use), the computational effort increases further. To address these difficulties, statisticians have devised ways of generating the simulated distributions without creating whole simulated networks.

Returning to the hypothetical example of 100 nodes and 500 links, the analysis would need to generate 1,000 networks of 100 nodes and 500 links and calculate reciprocity on all of them to create a distribution of reciprocity values for networks of 100 nodes and 500 links. Rather than generate an entire matrix of random links spread among the 500 nodes, we might randomly choose links to be on or off (connected or not). Now the procedure generates random networks based on the size and density parameters of the empirical network, but rather than generate entire networks, only a sample of each network is created. This sample is then used to calculate network parameters of interest. In this way, a large distribution of network parameters can be obtained based on randomly generated networks in an efficient manner.

One issue is how to generate the random networks. In the ERGM framework, the simulated networks are created by randomly generating networks based on the structural parameters of interest (density, reciprocity, transitivity). These simulated networks are referred to as dependence networks because they specifically indicate the nonindependence that exists between cases in the network. Unlike randomly selected samples, cases in a network dataset are specifically linked to one another as indicated by the network links and the structures in that network. So the dependence graph describes the dependencies indicated by the network based on the structural model. For example, if the structural model specifies that a network is characterized by transitivity, then two nodes linked in a transitive relationship are dependent on one another according to this structure. The dependence network is the connections in a network implied by a structural model. In other words, the dependence network indicates what links would exist in a network if the structural model explained the specific set of relationships.

In sum, ERGM is used when there is an empirical network in which we want to know whether there is a tendency for some structural relations, such as transitivity. A large sample of random networks is generated (actually sampled networks because it is too time-consuming to generate the entire network), which is based on the dependence graph implied by the structural **(p.160)** model, transitivity, plus lower-order dependencies (density and reciprocity). The empirical network's transitivity is then compared to the transitivity of the simulated sample. If the simulated sample's average and empirical values are different (where a *t*-value greater than 1.96 indicates statistical significance at the 0.05 level), the researcher concludes that the empirical network exhibits a tendency toward transitivity. Note that this analysis does not prove the network is transitive because the analysis has not been able to longitudinally model the underlying processes implied by the transitivity.

One other element needs to be added to make the story complete, and that is how to seed the simulated networks to create the distribution. The structural model indicates whether links should exist, given the implied dependencies in the model. But how do we know how to start the creation of the random networks? The “burn-in phase” of network generation refers to the initial seeding of links in the simulation of randomly generated networks. Initial links are selected at random and then the rest of the network is filled in according to the structural models specified. One of the challenges facing ERGMs is discovering ways to build these networks that provide reliable parameter estimates for the simulated networks. Some earlier techniques, Bernoulli graphs and Markov random graphs, were based on simple rules that did not provide satisfactory simulated network distributions. Recent developments in model specification have resulted in programs that generate better simulated networks in the sense that they are structurally similar to the empirical networks.

New Specifications

The approach outlined thus far had been used for some time with mixed success. One problem was that simulating networks with higher-order properties (such as transitivity) degenerated. *Degeneration* refers to the scenario when the simulation parameters lead to very non-normal networks that are either too connected or too disconnected. Specifically, transitivity is a key structural feature of networks and so researchers often wanted to control for transitivity in a network. Growing simulated networks that are transitive was problematic, however, because such networks tend to have highly connected subgroups with no bridges between them. The simulated networks often resulted in completed connected subgroups or completely null networks no links at all. Thus, the simulated networks did not resemble the empirical one on which they were based.

As Harrigan (2009) points out, two recent specifications have helped to solve these earlier problems and made it possible to estimate ERGMs (Robins and Pattison, 2005; Robins, et al., 2007; Snijders et al., 2007). These two **(p.161)** advances were (a) the introduction of the four-cycle and (b) the use of a “rubbery ceiling” on structural parameters, which limited the number or density of certain specifications such as transitive triangles. The most prominent model parameter that incorporates these changes is the alternating *k*-star configuration (Snijders, 2005).

The alternating *k*-star parameter is critical to model estimation and has three features. First, all star configurations (those that involve an ego's direct ties) are incorporated into one parameter estimate. Rather than count the number of 2-stars, 3-stars, 4-stars, and so on, the *k*-star configuration estimates how many of all kinds of stars there are in a network. Second, the probability of a star of a particular order is inversely proportional to its order. So 2-stars are more likely than 3-stars, which are more likely than 4-stars, and so on. This makes sense, of course, because in most networks, 2-stars are more common than 3-stars, which are more common than 4-stars, and so on. A parameter estimate for *k*-stars captures in one estimate the distribution of stars in the network. Finally, the sign, positive or negative, for the probability of higher-order stars alternates so that if a 2-star has a positive sign, the 3-star will be negative, the 4-star will be positive, and so on.

Researchers also introduced the alternating *k*-triangle parameter, which is similar to the alternating *k*-star. The alternating *k*-triangle incorporates a four-cycle and so relies on the conditional independence assumption mentioned earlier. Finally, researchers introduced the alternating *k*-two paths parameter, which estimates the probability that two nodes share another node (i.e., two people are connected by an intermediary). In sum, estimating traditional Markov models was problematic for some time until these alternating *k*-star, *k*-triangle, and *k*-two path parameters were introduced. Once they were incorporated into the simulation of the distribution of networks, comparisons between empirical and simulated networks could be made and parameter estimates generated. Some progress has also been made at estimating the simultaneous effects of multiple networks (Koehly & Pattison, 2005).

Obesity Example

For example, Valente and others (2009) collected data from sixth-grade students in four schools in 17 classes. Students were asked to name their five closest friends in the class. Height and weight measurements were taken of all the students and their body mass index (BMI, in kg/m^2) was calculated. BMI is typically used as an indicator of body composition, with adult BMI values greater than 30 signifying obesity.

An ERGM model was used to test the probability that friendship ties existed (relative to no tie) as a function of the network change statistics of **(p.162)** the structural properties density, reciprocity, and so on. The network change statistics refer to the difference in the count of various network configuration types when the tie from node *i* to node *j* is absent to when the tie is present. This study used the latest specification for ERGMs as discussed earlier (Robins et al., 2007). A common model was applied to all 15 classes and parameter estimates and their standard errors were aggregated as in a meta-analysis to determine if the effects generalize across classes (Snijders & Baeveldt, 2003).

After controlling for structural effects, weight status similarity had a strong and statistically significant effect ($T^2 = 53.73$, $df = 15$, $p <.001$) with a mean effect size of 0.22 ($p <.001$), indicating friendships were more likely to exist between students of the same rather than different weight statuses. The estimated between-classroom standard deviation of this effect size was 0.06 (p = NS). The study also tested whether weight status was associated with naming more friends or being named as a friend. For naming friends, there was a significant main effect of weight status ($T^2 = 27.93$, $df = 15$, $p <.05$) with a mean effect size of 0.13 ($p <.05$), indicating that overweight adolescents named more friends than nonoverweight ones. The estimated between-class standard deviation of this effect size was 0.13 ($p <.05$), indicating the estimated effect size was different between classes.

Note the ERGM results do not show which direction the association occurs. That is, there is an association between friendship and weight status, which also means that nonobese students are also more likely to be friends with one another. The ERGM parameter estimate is not particularly informative as the researcher does not know how strong the association is. The ERGM analysis, however, does assure the researcher that the association between weight status and friendship is not a function of being connected to the same others in the class or other structural characteristics of the network. This study (Valente et al., 2009) also used regular random effects logistics regression to estimate the association between being at risk for obesity and having obese friends. The regression results indicated an approximate two-fold increase in obesity for those with obese friends.

In sum, ERGMs are the first building blocks of statistical estimation of network structural effects. Researchers wishing to know whether an empirical network is significant can use ERGMs to generate simulate (random) networks derived from the features of the empirical networks. Higher-order structural effects can be compared, between the empirical and simulated distributions, provided the lower-order structural effects are matched. For example, transitivity can be tested provided size, density, and reciprocity are matched (or at least the analysis is conditioned on density).

The statistical analysis conducted to test whether a network exhibits the proposed structural features is conducted by estimating how well the ties in **(p.163)** the empirical network match those generated by the simulations. The networks are simulated, a matrix representing the links implied by the structural model is constructed, and then this matrix is converted into a vector that indicates who is linked to whom if the structural model were true. This vector can then be regressed (logistically since its binary) on the actual vector of who is linked to whom to determine if the empirical data fit the model.

Actor-Oriented Model

To this point, ERGMs have been described as ways to determine whether an empirical network exhibits structural tendencies (e.g., amount of transitivity) and whether individual attributes may be associated with a link between two people. For example, do the actors tend to form reciprocal or transitive relationships, and are obese students more likely to be friends with other obese students? Researchers, however, have been interested for some time in how networks evolve and whether there are individual (or node) characteristics associated with network evolution. Testing whether there are certain properties that drive network changes associated with network evolution is often referred to as the *actor-oriented model*. The estimation procedure allows researchers to identify whether behaviors are associated with the formation and/or dissolution of a tie between actors and whether network ties lead to changes in behaviors.

The actor-oriented model follows similar logic to ERGMs except that rather than generating 500 simulated networks to compare against an empirical one, the researcher specifies the way the network at time one *evolves* to become the network at time 2. Identifying the evolution of social network preferences over time uses what have been called stochastic actor-oriented models (Burk et al., 2007; Snijders, 2001; Snijders et al., 2007). Instead of calculating frequencies of various types of social configurations as is done in ERGMs, actor-oriented models simulate dynamic processes using what is essentially a form of agent-based modeling. Current software programs used to estimate actor-oriented models require the researcher to specify the parameters that are thought to govern how the network evolves from time 1 to time 2 and then generates a simulation of networks to determine whether imposing those rules will generate networks similar to the observed network at the later time point. The challenges for the researcher are to specify the structural and behavioral tendencies in the network (this is known as the *objective function*).

Currently, there are three computer platforms used to test longitudinal network and behavior models: SIENA (Simulation Investigation of Empirical Network Analysis) STATNET, and PNET. The rest of this section is written **(p.164)** somewhat from the SIENA perspective. Specification of a SIENA model entails specification of the objective function and rate function. The objective function is the specific network structural properties thought to drive the evolution of the network, and the rate function is the estimated number of changes each actor can make between observed time points. The objective function indicates the tendency or preference for the current state of the network. Practically, the researcher specifies in the software the structural and attribute-based factors thought to affect social and behavioral evolution and does not write out the objective function. For the rate function, the researcher typically allows the software to estimate the rates without imposing a rate function.

To test an actor-oriented model using SIENA, the researcher needs to specify the data, the network matrices, and the attribute vectors. If the researcher has multiple networks, such as several schools or organizations, then he or she needs to decide whether to estimate each network separately and conduct a meta-analysis to aggregate the results, or combine the networks into one large file and insert “structural zeros” for links between separate networks. Once the data are imported, SIENA provides separate windows to conduct data transformations and/or selection (subset of the population, for example).

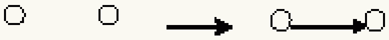
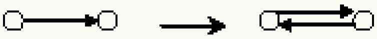
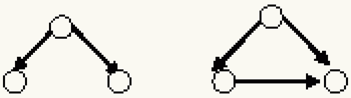
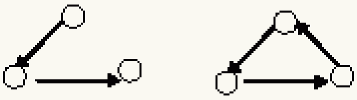
The heart of the evolutionary analysis occurs with the model specification. Researchers can specify the networks of relations as independent or dependent variables, and attributes can be constant or change over time. In addition, the researcher can specify relationship covariates that change over time (e.g., whether two individuals are married, or how far they live from one another). Finally, the researcher indicates the parameters to be estimated—structural parameters such as density, reciprocity, and transitivity—as well as the attributes that interact with these structural parameters. For example, a basic SIENA model might test whether students tend to reciprocate ties, to make transitive ties, and to make ties with others of the same gender (boys more likely to name boys and girls more likely to name girls). *Great care should be taken when specifying the model parameters as these should be guided by the theoretical model being tested*.

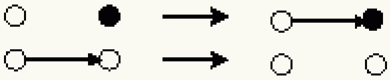
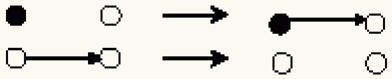
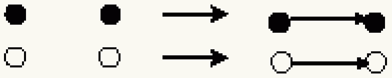
Once the model has been specified, the simulation runs by generating networks based on the model specifications. The model will converge if the simulation can generate networks that resemble the empirical ones. Often models do not converge because the parameter specification cannot generate networks that resemble the empirical ones, due in part to the empirical data (e.g., the networks are too sparse or too dense; see Chapter 8) or the model specification contains inherent contradictions.

t-Tests are provided to determine if the model converged. Here the researcher looks for non–statistically significant *t*-tests because we want the simulated networks to be similar to the empirical ones. SIENA then provides **(p.165)** estimates and standard errors to determine whether the analysis has produced significant *t*-values. Like in regression statistics, *t*-values greater than 2 signify significant estimates and the researcher can conclude that there is a tendency for the corresponding phenomenon in the data. A useful way to understand the various parameter settings to fit actor-oriented models is to diagram the parameter in terms of its relation from a focal actor and its neighbors.

Table 9–1 provides a diagram of some of the different structural properties tested in the actor-oriented evolution analysis. Each row represents a different structural parameter, which is a pattern of network relationships. The first row shows out-degree, which is an indicator of the density (the number of links) in the network. Out-degree is always included in an actor-based model to control for density. Out-degree parameter estimates are almost always negative, which suggests that ties are structured and nonrandom (a positive out-degree would indicate that network structures would tend to become decentralized with a density of 50%). The second row shows mutuality (reciprocity), where

Table 9–1. Structural Parameters Typically Estimated in ERGMs

Parameter	Illustration of Social Process	Description	Example
Out-degree (density)		The overall tendency to have ties	Actors increase their connectedness
Reciprocity		Tendency to have reciprocated ties	Actors prefer others who have selected them.
Transitive triplets		Tendency towards triadic closure of local ties	Actors prefer others who are friends of their friends
Balance effect		Tendency to have ties to structurally similar others (structural balance)	Actors prefer to have ties to others in their circle

Parameter	Illustration of Social Process	Description	Example
Attribute alter		Main effect of alter's behavior (covariate determines in in-degree)	Alters are nominated based on having an attribute or not nominated based on not having the attribute
Attribute ego		Main effect of ego's property on tie preference	Actors make ties based on having an attribute or do not make ties based on not having the attribute
Attribute similarity		Tendency to be connected to similar others	Actors prefer ties with others who are the same on an attribute.

Based on Steglich et al. (2007).

(p.166) a positive estimate would indicate that given an incoming tie from B to A, there is high chance for A to reciprocate that tie to B. The other rows in the table are structural tendencies that can be tested after these first two.

WINCART

To illustrate the co-evolution model, a study was conducted to evaluate the effects of a Community-Based Participatory Research (CBPR) intervention designed to increase linkages between community based organizations and university researchers around issues of cancer education, training, and research (Israel et al., 2000; Wallerstein & Duran, 2006; Valente et al., in press). This study reports the results of the Weaving an Islander Network for Cancer Awareness Research and Training (WINCART) initiative designed to reduce cancer disparities among Pacific Islanders in Southern California (Tanjasiri & Tran, 2008). WINCART was created as a forum for community groups to meet and establish connections between various Pacific Island community groups and involved both a scientific advisory board and a community advisory board to guide WINCART's education, research, and training activities. A stated objective of the WINCART initiative is to create linkages between community-based organizations (CBOs) and academic institutions conducting cancer research (Tanjasiri et al., 2007). These linkages would enable community organizations to disseminate information about cancer research and treatment developments to their constituents. At the same time, WINCART was designed to create connections from academic institutions and cancer researchers to CBOs so that cancer research, education, and training would be more community informed.

WINCART conducted many activities to bridge the gap between community and academia such as retreats, events, symposia, and relationship building. An actor-oriented model was used to determine if WINCART was effective in integrating these groups. There were 19 organizations in the study: 11 CBOs, 5 universities, and 3 national cancer-related organizations (e.g., American Cancer Society). Fourteen network questions were asked with regard to (1) communication, (2) formal agreements, (3) client referrals to, and (4) client referrals from; non-cancer communication or interaction regarding (5) education, (6) outreach, (7) training, (8) advocacy, and (9) research; and cancer communication or interaction regarding (10) education, (11) outreach, (12) training, (13) advocacy, and (14) research. Respondents were presented with a roster of all 19 organizations and invited to check those with which they interacted.

Electronic invitations and surveys were sent to 121 individuals in 16 organizations in June 2005 and 113 individuals in 17 organizations in July 2007. **(p.167)** Ninety-one respondents completed the survey at time 1 (75.2%) and 56 at time 2 (49.5%). At time 1, the three national organizations were not solicited to participate but invitations were sent to representatives of CIS working in the community at time two. These responses were not included in the study since these data were only available at time 2. All linkages to these three national organizations were removed from the data for this analysis because they did not make any nominations (were not invited participate in the study). At least one individual from every participating organization responded. The lower than expected response rate may have been a function of individuals within the same organization telling each other they responded for the organization. Follow-up conversations with some nonresponders also indicated that they mistook the time 2 solicitation for a reminder of the time 1 survey they already completed. The data were aggregated to the organizational level so that individual responses are unknown.

Because the number of respondents from each organization varied, links between organizations were summed and then divided by the number of respondents from each organization. The dependent variable in this case is the percent of links from one organization to another and the guiding research question is whether connectivity increased over time and became more heterogeneous on status (CBO versus university).

There was a decrease in respondents per organization from 4.79 and 2.95 at time 1 and 2, respectively. Most respondents were female, 87.9% and 81.4% at time 1 and 2, respectively. Respondents were experienced working in their organizations, averaging approximately 7 to 8 years working with their current organization. Most had participated in WINCART activities, 58.6% to 67.9%, averaging 1.71 to 2.38 activities in the past year. There were 1,426 links reported in response to the 14 network questions at baseline, and despite the fewer number of respondents, there was an increase to 1,617 links at time 2. There were 146 and 159 links within the organizations at time 1 and 2, respectively.

The linkage rate (percentage of respondents in the organization who nominated another organization) was 30% at time 1 and increased to 43% at time 2. At both time points, organizations were nominated by the most respondents in response to the question, “Which organizations have you communicated with in the past year” (44% and 54%, respectively). At time 1, the number of connections was lowest for receiving clients (19%), but at time 2 it was lowest for non–cancer-related research (34%). Figure 9–3 illustrates the cancer education network at times 1 and 2 with organizations depicted as CBOs (circles) or universities (squares).

As in the ERGM example, initial analysis consisted of estimation using ordinary least-squares regression of the connection percentage as a function of time; follow-up response rate; network question (general, noncancer, or **(p. 168)**

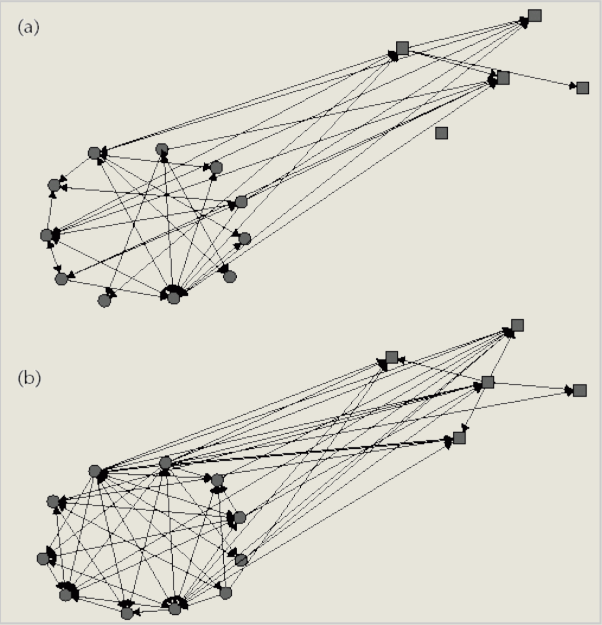


Figure 9–3. (a) Cancer Education Network at time 1. (b) Cancer Education Network at time 2. Links indicate which organization nominated which other dichotomized on the median value. Circles are CBOs and squares are universities. There is an increase in linkage from CBOs to universities but not from universities from CBOs. Figures created with Netdraw (Borgatti, 2005).

cancer); organization type; and a time-by-type interaction term. A positive and statistically significant time-by-type interaction effect indicates that linkages increased over time between and/or among types. The regression model was also reestimated using a random effects probit model to control for clustering of responses within organizations. For both of these models, the regression analysis showed a statistically significant interaction terms in which links from CBOs to CBOs and links from CBOs to universities increased over time (see Valente et al., in press).

In the statistical evolution model, parameters were included that tested whether there were tendencies for linkages between organizational types to occur after controlling for the network structural effects of density, reciprocity, and transitivity. Specifically, the objective function specified whether (1) there were more outgoing links based on organizational type, (2) there were more incoming links based on organizational type, and (3) organizations of the same type link with one another more than they link with other **(p.169)** types (similarity effects of the organizational type). The joint contribution of these organization-type effects to the objective function is (Snijders et al., in press) presented as follows:

9–1

$$\beta_e \sum_j x_{ij} v_i + \beta_a \sum_j x_{ij} v_j + \beta_s \sum_j x_{ij} I\{v_i = v_j\}$$

where β_e is a parameter for the ego effect, β_a is a parameter for the alter effect, β_s is a parameter for the same effect, x_{ij} is a tie variable from organization *i* to *j*, v_i is an ego's value of the organizational type, v_j is an alter's value of the organizational type, and $I\{v_i = v_j\}$ is an indicator function of the similarity coded as 1 if $v_i = v_j$ and as 0 otherwise. The equation representing the contribution to the objective function of the single tie from organization *i* to organization *j* (x_{ij}) that takes only the effects related to organizational type is as follows (Snijders et al., in press):

9–2

$$\beta_e(v_i - \bar{v}) + \beta_a(v_j - \bar{v}) + \beta_s I\{v_i = v_j\}$$

where $\bar{v}$ is a mean value for the centering.

The 14 networks were analyzed separately and combined using meta-analysis (Snijders & Baerveldt, 2003). Each network was dichotomized on median values of the proportion respondents nominated for each network across waves (ranged from 0.20 to 0.33). The same objective function specifications were applied to each network and the results were combined to produce vectors of parameter means and standard errors across networks (with specification of the upper bound of 5). Based on these results, ego-alter selection values were created using Equation 9-2 for the two values of organization type, v_i and v_j. Organization type was coded as 0 for the five universities and 1 for the 11 CBOs yielding a global mean $\bar{v}$ = 0.69. The centered value for university was –1.69 (= 0 – 1.69) and 0.31 (= 1 – 0.69) for CBOs. All estimation was done using SIENA (Snijders et al., 2007).

For the SIENA results, all 14 networks attained convergence with t-ratios being less than.1 in absolute value. The results of the meta-analysis across all networks indicated there was a significant ego effect for CBOs indicating they increased their outgoing linkages more rapidly than universities ($T^2 = 69.00$, df = 14, p < 0.001; mean effect size = 1.15, p < 0.001). The alter, or incoming, effects ($T^2 = 20.02$, df = 13, p = 0.10; mean effect size = –0.64, p < 0.001) and similarity effects ($T^2 =$ 14.49, df = 13, p = 0.34; mean effect size = 0.55, p < 0.01)) based on organizational type were not significant indicating no difference in nominations received by organization type and no difference in the likelihood of linkage between organizations of the same type. The estimated between- network standard deviation for the ego effect (outgoing) parameter along with **(p.170)** those for incoming and similarity parameters were negligible indicating similar effects across networks. The estimated mean effect sizes for each covariate were entered into equation (3):

9–3

$$1.15\,(v_i - 0.69) - 0.64\,(v_j - 0.69) + 0.55\, I\,(v_i = v_j)$$

Substituting the values 0 for Universities and 1 for CBOs into *v*, yields the following results for ego-alter selection tendencies as in Table 3: university to university, 0.20; university to CBO, –0.99; CBO to university, 0.80; and CBO to CBO, 0.71. Consequently, CBOs exhibited a tendency to prefer relations with universities (0.80) or other CBOs (0.71); whereas universities tended to prefer connections to other universities (0.20) and not to CBOs (–0.99). In sum, these results show that there was a tendency for CBOs to connect to one another and to universities, but universities did not demonstrate a preference for connecting to CBOs or themselves.

These results illustrate how the actor-oriented co-evolution model can be applied to a substantive public health problem, evaluating CBPR as a means to bring evidence-based public health to communities and to enable researchers to conduct community-informed research. The statistical analysis showed that linkages from CBOs to other CBOs and from CBOs to universities increased during the study. In contrast, university faculty did not increase their ties to other universities or to CBOs. This suggests that WINCART was successful at motivating network change among the community partners but not among university researchers (Valente et al., in press).

There are many other applications in public health that are just being tested. SNA provides a framework for understanding the structural processes that give rise to a specific network. The regression approach treats social relations as antecedents to behaviors. Usually in public health, we are more concerned with how a specific network structure affects the spread of disease or risk behavior, not in the mechanisms that created a specific structure. However, to the extent that attributes are included in the analysis, SNA can be useful for understanding how certain attributes interact with network structure. So for example, it can be determined whether smokers are more likely to associate with other smokers, thus testing for homophily on smoking behavior, which is critical for understanding peer influence.

Summary

This chapter has attempted to provide a nontechnical introduction to techniques used to conduct statistical analyses on networks. The ERGM was introduced as the building block for statistical estimation. ERGM is a technique in which an empirical network is compared to randomly generated **(p. 171)** ones to determine if structural properties in the empirical network occur greater than expected by chance. The critical element is to understand that structural properties are hierarchical so that density is nested within reciprocity, which is nested with triadic structures, and so on. The simulated networks match on the lower structural properties and are used to statistically assess the higher ones.

The chapter also emphasized the innovation of testing structural properties against the presence of a link in the network. Initial statistical analyses of networks involved converting the matrix of links to a vector representing each tie and regressing this on structural elements in the networks. Recent developments of new specifications for the generation of simulated networks has created the capability to test co-evolution models that simultaneously test for network and behavior changes in one statistical model. The chapter included some empirical examples. As Snijders notes, “The statistical modeling of social networks is difficult because of the complicated dependence structures of the processes underlying their genesis and development” (Snijders, 2005, p. 215). In spite of these challenges, considerable progress has been made and many new discoveries regarding network effects and processes are imminent.

Acknowledgments: The author thanks Laura Koehly, Kayo Fujimoto, and Kate Coronges for help with this chapter.