CHAPTER 3. BUILDING A USEFUL EXPONENTIAL RANDOM GRAPH MODEL

Essentially, all models are wrong, but some are useful.

—Box and Draper (1979, p. 424), as cited in Box and Draper (2007)

For decades, network scientists struggled with statistical network models (e.g., the simple random graph model) that were not very useful in explaining the structures found in observed social networks. The employment of the Markov dependence assumption and recent advances allowing for more general and complex dependence assumptions resulted in statistical network models that are quite useful in representing, explaining, and predicting observed social structures. While the underlying dependence assumption often differs, the interpretation of an exponential random graph model (ERGM) is similar to the interpretation of a binary logistic regression model. That is, a tie in a network is the outcome, and the characteristics of network members and network structures aid in explaining or predicting the probability of a tie (Hunter, Goodreau, & Handcock, 2008).

The sections that follow demonstrate the development of a complex ERGM. The demonstration begins with exploratory analysis to identify features of the observed network and network members that may be important to capture during model development. The model is built starting with a simple random graph model only capturing network density. Main effects and interaction terms are then added to the model to represent the attributes of network members; this step results in a dyadic independence model. Finally, geometric terms are added to account for underlying network structures not captured by main effects or interaction terms for a dependence model. Diagnostic tools, strategies for assessing model fit and selecting a model, and interpretation of model results are incorporated throughout.

A list of R commands to re-create the analyses is included in Appendix A online. The commands are numbered and marked throughout the demonstration wherever commands are available. When the text refers to "Command 1," the command(s) to replicate the results shown is labeled Command 1 in Appendix A. Note that the software used to conduct the analyses shown in this monograph is open source; it can, and often does, change. The set of commands included was prepared using the software and package versions specified in the next section and may need

adjustments to work in future versions. Changes to commands can usually be identified by reading the associated help documentation for the command of interest. Occasionally in this chapter and the next chapter, commands will be included within a paragraph. R commands embedded in paragraphs of text will use courier font; when the reader should replace part of a command with a specific file name or other information, the text will be underlined. For example, the command, `read.paj('data file')` indicates that the reader should replace the words *data file* with the name of a data file in order to use the command.

Obtaining and Preparing Software

Several packages are available for estimating network models, including PSPAR, Multinet, R-statnet, RSiena, and Pnet (Shumate & Palazzolo, 2010). The analyses that follow were conducted in R-statnet, which is a suite of packages for developing ERGMs in R. A list of the developers can be found on the statnet website (http://statnet.csde.washington.edu/about_us.shtml). R is free software available from the R Project for Statistical Computing website: http://www.r-project.org/. The R software functions as a platform for developers, who can develop and disseminate statistical packages for use in the R shell. In addition to installing R, users will need to install the statnet suite, which is separate from the R shell. As its name suggests, the R-statnet suite includes packages developed by the statnet team, including ergm, network, sna, and networkDynamic. In addition, the statnet suite includes a group of packages that statnet relies on, including robustbase, Martix, lattice, trust, nlme, and coda, that were developed by others outside the statnet team. Each package includes specific features, functions, and terminology useful in developing ERGMs. Help for a specific package can be found by typing `help( package )` at the R prompt, replacing the word `package` with the name of the package. The analyses shown in this text were conducted in R version 2.15.2 using statnet version 3.0–1.

To install R-statnet, use the Packages menu in R, or type the following at the R prompt:

```
install.packages( 'statnet' )          Command 1
```

This command installs statnet from one of the many repositories of R packages. These repositories are called the Comprehensive R Archive Network (CRAN) and are located around the world (http://cran.r-project.org/). Each CRAN site contains identical material, including software

packages and documentation for R. If statnet is already installed, the update.statnet command also included in Command 1 can be used to update to the current statnet suite.

Once the statnet suite is installed, it must be loaded before it can be used *each time* you start R; only those packages that are loaded will work in any given R session. To load the statnet suite, enter the following at the R prompt:

```
library( 'statnet' )                    Command 2
```

Accessing Data

The analyses in this chapter use a network data set obtained from the National Association of County and City Health Officials (NACCHO), which is available as an R package from the CRAN. Following the same steps as above, install and open the `ergmharris` package to access the data; start by using the `install.packages( 'ergmharris' )` command, followed by the `library( 'ergmharris' )` command. Additional data sets are available as part of the statnet package or other R packages and can also be used; to view a list of the data sets available in R, enter the following at the R prompt:

```
data()                                  Command 3
```

A new window will open showing a list of all of the data sets available in R. This list will vary depending on the packages installed. Network data from outside of R can be read into R from many different types of files. For example, files saved in the Pajek network software as .paj or .net files can be read using the `read.paj()` function; files saved in the common network edgelist format can be read in as matrices.

Once a network file is read in, depending on its format, it may need to be converted to the network data type, or *network class* in R parlance, in order to use statnet. To check the class of imported data, enter `class(data name)` at the R prompt. If 'network' is not returned, the file will need conversion to a network format before network modeling can be conducted with statnet. Depending on the format of the data, this may be as simple as using the command `as.network(data name)`. For instructions on conversion of different data types to a network object, see Butts (2008).

The NACCHO data set used for the remainder of this chapter is a network of communication relationships among local health department leaders

nationwide. The data were collected in 2010 through a survey sent to all local health departments (LHDs) in the NACCHO directory (http://www.naccho.org/about/LHD/). The survey instrument included questions about LHD structure, finances, leadership, and staffing, along with the types of health programming conducted by LHDs at a local level. To open the LHD network data in the ergmharris R package for use in the following tutorial, use Command 4 and type the following at the R prompt:

```
data( lhds )                                    Command 4
```

After the data are loaded, type the name of the network object, `lhds`, at the prompt to check that the data loaded properly:

```
lhds                                            Command 5
```

The output resulting from Command 5 shows descriptive information about the network, including the network size (vertices = 1,283), whether the network is directed (directed = FALSE), how many edges the network includes (n = 2,708), and additional information about the network. Next in the output are names of the attributes of the network members (Vertex attribute names) included as part of the network object for use in the tutorial. In this case, there were five attributes of the 1,283 local health departments stored with the network: state, nutrition, hivscreen, popmil, and years. These attributes were defined as follows:

state: the state where the LHD is located

nutrition: binary variable indicating whether the LHD does nutrition programming (nutrition = Y) or not (nutrition = N)

hivscreen: binary variable indicating whether the LHD does human immunodeficiency virus (HIV) screening (hivscreen = Y) or not (hivscreen = N)

popmil: LHD jurisdiction population in millions

years: number of years the current LHD leader has been in his or her position in categories of 1 to 2 years (years = 0), 3 to 5 years (years = 1), 6 to 10 years (years = 2), and >10 years (years = 3)

Following Goodreau and colleagues (2008), a more complete summary of the network and its attributes can be obtained using Command 6 (Table 3.1). The summary shown in Table 3.1 starts with general network characteristics including network size, density, and whether the network is directed (e.g.,

Table 3.1 Partial R output summarizing network data.

```
Network attributes:
  vertices = 1283
  directed = FALSE
  hyper = FALSE
  loops = FALSE
  multiple = FALSE
  bipartite = FALSE
  title = lhds
  total edges = 2708
  missing edges = 0
  non-missing edges = 2708
  density = 0.00329279

Vertex attributes:

 hivscreen:
   character valued attribute
   attribute summary:
  N   Y
461 804

 nutrition:
   character valued attribute
   attribute summary:
  N   Y
326 941

 popmil:
   numeric valued attribute
   attribute summary:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.00055  0.01722  0.04094  0.15860  0.12870 10.11000


 state:
   character valued attribute
   attribute summary:
   the 10 most common values are:
MO OH MA IL KS NJ WI NC FL MN
73 72 66 63 63 62 62 57 49 49
  vertex.names:
   character valued attribute
   1283 valid vertex names

 years:
   integer valued attribute
   1283 values

No edge attributes


Network edgelist matrix:
        [,1] [,2]
   [1,]    2   10
   [2,]    2   11
```

directed = FALSE). After the general network information, there are descriptive statistics for the five attributes assigned to the network object. The attributes show that most local health departments are conducting HIV screening (Y = 804; N = 461) and nutrition programming (Y = 941; N = 326), the popmil attribute indicates that LHDs have between 550 and 10.1 million constituents, and Missouri and Ohio have the most LHDs in the network with 73 and 72, respectively.

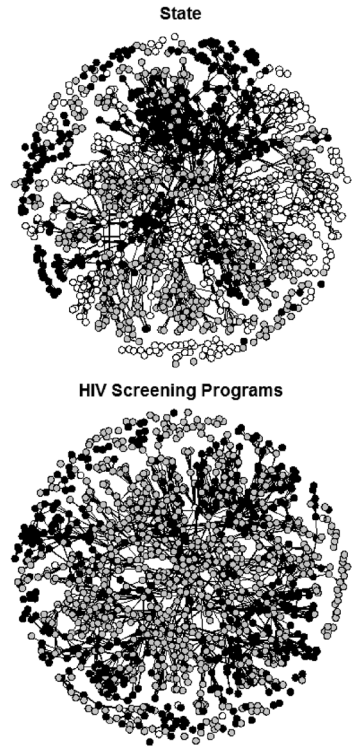
Following the vertex attributes, there is a section for edge attributes. Edge attributes are additional characteristics of each edge in the network. For example, if the file included information on how far apart two LHDs were in miles, an edge attribute might indicate a distance in miles. The LHD network data set includes no edge attributes. The last piece of information in the table is a full list of the edges in the network. For example, the first edge in the list is a link between the LHDs represented by node 2 and node 10; this list was truncated in Table 3.1 due to its length but should be visible in R after running Command 6.

Exploring Network Data

As with all statistical modeling, data exploration is advised prior to network model development. In the case of network data, visualization and descriptive statistics can give some insight into the structure of a network that can be helpful during the development and evaluation of a statistical model.

Using the set of commands labeled Command 7 in Appendix A, a visualization of the network with node color showing LHD attributes can aid in identifying patterns of ties among LHDs with different characteristics. There is some apparent clustering in the network graphic showing state; groups of nodes with the same color, indicating they are in the same state, are clustered together. This may indicate that an LHD is more likely to communicate with another LHD when the two are located in the same state (Figure 3.1). The network shaded by HIV screening in Figure 3.1 also demonstrates some clustering; the LHDs with lighter shading seem to cluster toward the middle of the network, while those with darker shading seem to be more on the periphery of the network. One hypothesis stemming from these visuals is that same-state communication appears more likely than would happen just by chance. Likewise, perhaps leaders in LHDs conducting the same types of programming (e.g., HIV screening) are more likely than expected to communicate. Note that Command 7 will not produce the exact spatial properties of the plots seen in Figure 3.1; the attributes of the nodes and the ties among them represent the data, but the location of each node in space is partially arbitrary and should not be interpreted as having a specific meaning.

Figure 3.1 LHD network depicting communication among health departments with nodes shaded to show characteristics of the LHDs.



Having a large number of nodes in the network can sometimes obscure important patterns in a network graph. Displaying the largest component (i.e., largest connected group of nodes) in a network can aid in clarifying patterns. The largest component can be isolated and graphed using Command 8. In this case, the largest component contains most of the nodes in the network (n = 1,083). Figure 3.2 shows the largest component with nodes shaded by whether or not the LHD has an HIV screening program. Note that Figure 3.2 now includes a legend that is specified in Command 8; there are many options for creating and placing a legend in R. For a list of options and instructions on creating a graph legend in R, use `help( legend )` at the R prompt. Other options for visual display of networks during exploratory analysis can be found in Goodreau and colleagues (2008), Butts (2008), and tutorials on the statnet website (http://statnet.csde.washington.edu/).

Vertex size and vertex shape are other ways to visually discern patterns in network structure. Nodes should be sized by continuous or ordinal attributes only; nominal variables may be used for node color and shape. In the LHD network, most of the attributes are nominal; however, jurisdiction population or LHD leader experience might be used to size the nodes. Network measures such as degree can also be used to size nodes in a network. Degree is the number of links a network member has. In this case, degree would represent the number of communication relationships for each LHD in the network. Unfortunately, use of Command 9 to create a degree attribute and plot the network using this attribute results in very large node sizes that overlap and obscure the ties. The size of the nodes can be reduced by reducing the values of degree in the attribute. Command 10 divides the original degree attribute by 6 and plots the network with larger nodes that have more connections to other LHDs. The graphic (Figure 3.3) now shows some patterns around HIV screening programming; there appear to be more large white nodes than large black nodes, indicating that LHDs doing HIV screening (white) have higher degree (more connections with other LHDs) than LHDs not doing HIV screening (black). One possible explanation for this is that LHDs conducting HIV screening programs are in big cities and, therefore, may be more visible and well connected to other LHDs.

In addition to visualization, examining network and node characteristics can provide some insight into network structures and possible modeling strategies (Goodreau et al., 2008). Network size and density are shown in Table 3.1; the average number of links per node (mean degree), the frequency

Figure 3.2 The largest component in the LHD network shaded by whether or not the LHD is conducting HIV screening.

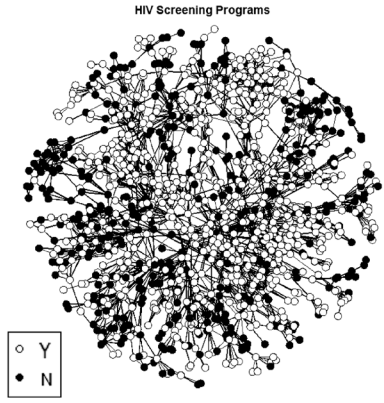
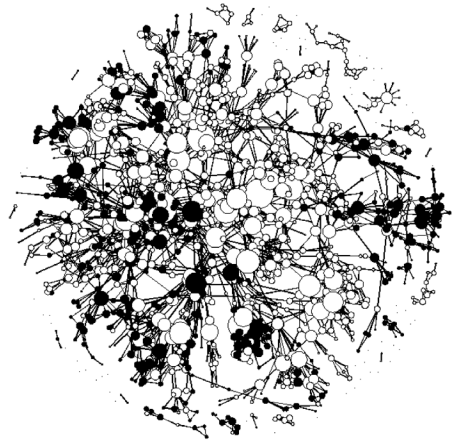


Figure 3.3 LHD network shaded by HIV screening program with nodes sized by the number of connections the LHD has within the network (degree).



of each degree value, and the distribution of triads can be obtained using Command 11 (see Figure 1.5 for a graphic depiction of the four triad types). Note that the degree command is set to default to consider the network as directed; to specify that a network is undirected, use the gmode argument with the "graph" indicating undirected ("digraph" indicates directed).

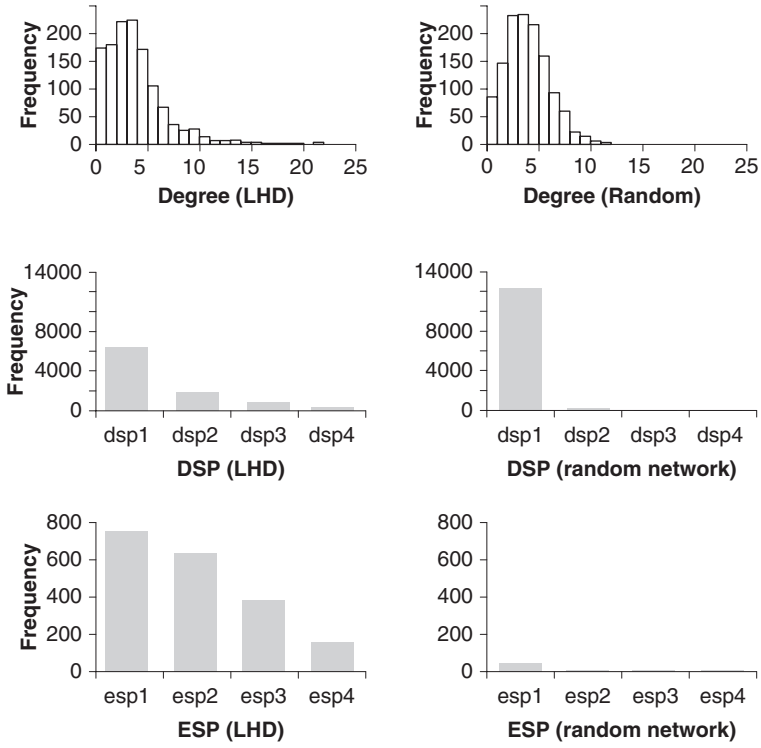
```
> mean( degree( lhds, gmode = "graph" ) )
[1] 4.221356

> sd( degree( lhds, gmode = "graph" ) )
[1] 2.895897

> table( degree( lhds, gmode = "graph" ) )

  0   1   2   3   4   5   6   7   8   9  10
 58 117 182 223 226 172 104  67  35  25  26
 11  12  13  14  15  16  17  18  19
 14   8   6   8   4   3   1   1   1
 20  22
  1   1

> triad.census( lhds, mode = "graph" )
              0       1    2    3
[1,] 347709795 3445061 9788 1437
```

The LHD network had an average degree of 4.22 (*SD* = 2.90); so, LHDs are connected to, and therefore communicating with, an average of 4.22 other LHDs. With 1,283 LHDs and 2,708 links between them, an average of 4.22 may seem higher than you would expect. However, a single link

includes two LHDs, so the link between A—B would be included in the degree count for A and B. That is, each of the 2,708 links contributes to the degree count for two LHDs, for a total of 5,416 degrees (2,708 * 2) across 1,283 LHDs. The 5,416 is divided by the 1,283 for an average degree per LHD of 4.22. The degree distribution table output above shows 58 LHDs with zero connections, 117 with one connection, and so on. The triad census table shows 347,709,795 triads with no edges, 3,445,061 triads with one edge, 9,788 triads with two edges, and 1,437 complete triangles.

Graphic examination of degree, edgewise shared partners (ESP), and dyadwise shared partners (DSP) are also useful in understanding underlying network structures. As is common in observed networks, the distribution of degree in the LHD network shows many low-degree nodes and few high-degree nodes compared with a random network of the same size and density (Command 12; Figure 3.4). Note that Command 12 may take 10 minutes or longer to run.

Figure 3.4 Plots of degree and shared partnerships in the observed LHD network (left) and a randomly generated network of the same size and density (right).



Edgewise and dyadwise shared partner distributions also differ in the observed and random networks (see Figure 2.3 for examples of ESP and DSP), with the observed LHD network having more network members with multiple DSP and ESP compared with the random network, which shows a large number of nodes with a single shared partner and little else in terms of shared partners.

The network visualizations above showed some potential clustering; another option for identifying clustering is to examine mixing matrices and correlation coefficients. Following Goodreau and colleagues (2008), a mixing matrix can be used to examine the number of connected dyads (pairs of LHDs) for each possible combination of levels for a categorical node attribute. For example, how many connected dyads have both LHDs doing HIV screening, or how many connected dyads are there with one LHD in Missouri and the other LHD in California? In the visualizations above, we found some evidence of clustering by state and by program area (HIV screening). Mixing matrices can help to confirm these patterns and explore other node attributes as well (Command 13; Table 3.2).

Table 3.2 Mixing matrices for HIV screening, nutrition programming, and years of leader experience.

```
> mixingmatrix( lhds, "hivscreen" )
    N    Y
N 526  632
Y 632 1498
> mixingmatrix( lhds, "nutrition" )
    N    Y
N 216  648
Y 648 1812
> mixingmatrix( lhds, "years" )
    0   1   2   3
0  71 190 207 283
1 190 120 259 355
2 207 259 225 516
3 283 355 516 389
```

The matrices show each level of each attribute in the network as both a column and a row; the numbers in the matrices represent the number of connected dyads with the corresponding row and column attribute. For instance, a connected dyad where both LHDs were conducting nutrition programming would be counted among the 1,812 dyads in the lower right corner of the second mixing matrix shown in Table 3.2. A dyad where one LHD is not doing nutrition programming and the other is doing nutrition programming would be counted as one of the 648 dyads on the off-diagonal. The state attribute mixing matrix is not shown here given the large size of the matrix

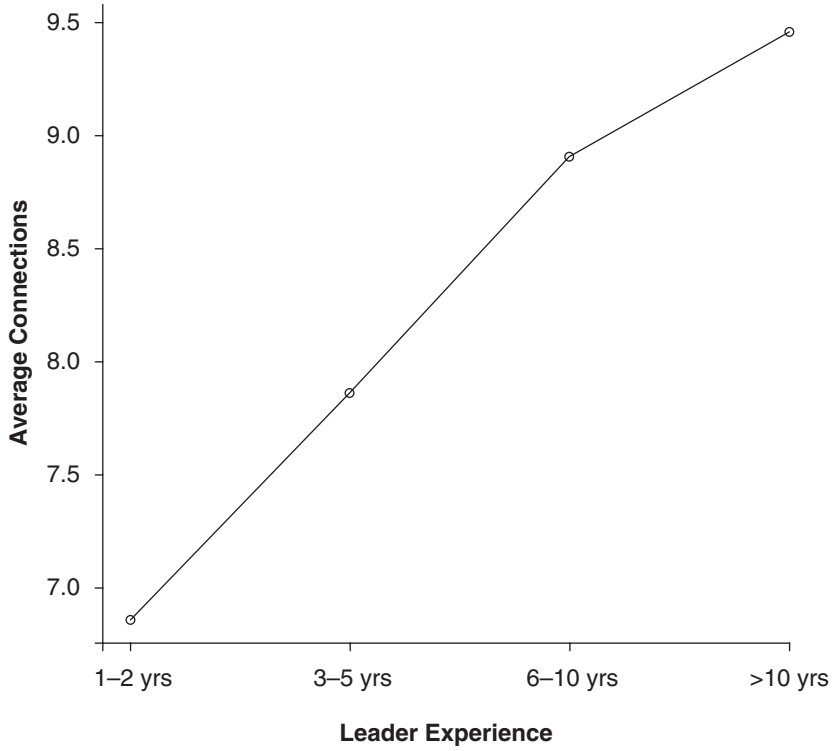
(49 rows and 49 columns), but this will show in the output window after running Command 13. A review of this large matrix will show that connections tend to be between LHDs in the same state.

A few patterns emerge in the mixing matrices. Of the connected dyads, 1,812 are connections between two LHDs doing nutrition programming, while 648 connected dyads include an LHD doing nutrition programming and one not doing nutrition programming, demonstrating a propensity for LHDs with similar programming to be connected (homophily of programming). The HIV mixing matrix shows a similar pattern, although not as clearly, with 1,498 connected dyads where both LHDs are doing HIV screening. However, the leader years of experience matrix shows a different connection pattern. In this mixing matrix, LHDs with the more experienced LHD leaders seem to have more connections with others in all experience groups. For example, of the 751 connected dyads including a leader with the lowest experience level (coded 0), 283 (38%) are connections with an LHD with a leader in the top experience category (coded 3); in contrast, of the 1,543 connected dyads that include an LHD with a highly experience leader, just 18.3% include an LHD with a low-experience leader. The average number of connections for each experience category can be calculated and plotted to further examine the relationship between leader experience and network structure (Command 14; Figure 3.5); this graph shows that LHDs with more experienced leaders have more connections than those with less experienced leaders.

Further examination of node characteristics could provide additional insight into network structure. For example, for continuous attributes such as popmil, it may be useful to examine the correlation between the attribute and degree (Command 15). The resulting correlation coefficient of .27 indicates that, as population in an LHD jurisdiction increases, the number of connections also increases. In addition, in some networks, it may be useful to be able to examine the average number of connections for nodes with different attributes. Command 16 demonstrates an additional way to explore the attribute data using two-way tables.

These exploratory analyses aid in identifying numerous characteristics of the LHD network that may be useful during the model-building process. First, the network shows extensive homophily by state and moderate homophily by program area. Second, LHDs with more experienced leaders have more connections; likewise, LHDs serving larger populations have more connections. Finally, underlying structural features in the LHD network are notably different from those in a random network of the same size and density. Specifically, the degree distribution is nonuniform, including more nodes with low degree and a few nodes with very high degree. There are also more nodes with multiple ESP and DSP in the LHD

Figure 3.5 LHDs with more experienced leaders have more links to other LHDs.



network compared with the random network, indicating higher rates of transitivity and pretransitivity than expected by chance. The presence of homophily, a nonuniform degree distribution, and transitivity in the LHD network are all consistent with current network theory and modeling strategies; each of these qualities can be incorporated into a statistical network model that could be used to better understand the social forces underlying the observed network structure.

Model Building

The Null Model

As with many forms of model building, statistical network modeling should begin with a null model. The null model is a simple random graph

model, the simplest model described in Chapter 2, consisting of a single term representing the edges, or number of connections, in the network (Goodreau et al., 2008). The statistical model shown in Equation 6 can be modified to depict the null model:

$$\text{logit}\left(P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)\right)=\theta_{\text{edges}}\delta_{\text{edges}} \tag{12}$$

where δ_{edges} is the change statistic for the edges term, and θ_{edges} represents the coefficient of the edges term. The null model estimated in R (Command 17; Table 3.3) for the LHD network includes the edges coefficient (θ_{edges}= –5.71272) along with several other pieces of information.

Table 3.3 LHD network null model.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges

Iterations:  20

Monte Carlo MLE Results:
      Estimate Std. Error MCMC % p-value
edges -5.71272    0.01925     NA  <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   36365  on 822402  degrees of freedom
          Deviance: 1103728  on      1  degrees of freedom

AIC: 36367    BIC: 36379
```

Using Equation 7, the probability of a tie in this network can be calculated from the information in this table. There is only one term in this model, the edges term. The column labeled "Estimate" is the column where the coefficient (θ) for each term in the model is found. In this case, the coefficient for the edges term is negative (–5.71272), indicating that the density of the network is below 50%; an edges term of 0 would represent a 50% or .5 density. A negative edges coefficient is a typical feature of an observed network; very few observed networks have a density of .5 or higher. Most network models will contain negative edges terms.

Remember that the change statistic (δ) represents the change in the statistic of interest when an edge is added (as Y_{ij} goes from 0 to 1) (Hunter, Goodreau, & Handcock, 2008). The edges term will always have the same change statistic, $\delta_{edges} = 1$, because the edges term accounts for the number of edges in the network, and the addition of one edge to the network changes the number of edges in the network by 1. The logistic function on the right-hand side can be calculated as usual for a logistic regression model: $\frac{1}{1+e^{-(\theta_1 X_1)}}$ (Field, 2009).

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\left(\theta_{\text{edges}}\delta_{\text{edges}}\right)$$

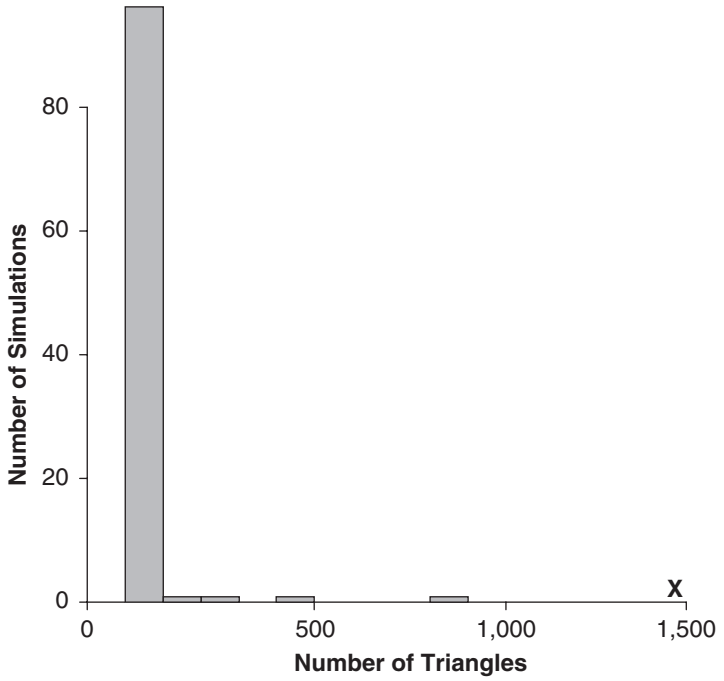
$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\left(-5.71272*1\right)$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \frac{1}{1+e^{-(-5.71272*1)}} = 0.003293$$

The resulting probability of a link is, as expected, the same as the density of the LHD network, .0033. This model was estimated using the same maximum likelihood estimation methods used in standard binary logistic regression. Because the null model is a simple random graph model, there are no complex dependence assumptions to account for. While estimating a null model may seem like a complex way to demonstrate the simple network characteristic of density, the null model provides a baseline for assessing model fit that is useful as more complex models are built.

Although the null model is a good representation of the observed density of the LHD network, it is unlikely to be a good representation of other observed network characteristics. Plots of network measures from simulated networks based on the null model can aid in understanding how well this model captures the structures (e.g., triangles) comprising the observed network. Following Goodreau and colleagues (2008), 100 networks can be simulated and the distribution of triangles across the simulated networks graphed using Command 18. Figure 3.6 shows the distribution of triangles across 100 networks simulated from the null model; an X marks the spot for the 1,437 triangles in the observed LHD network, which is much higher than the number of triangles in any of the 100 simulated networks. Clearly, a more complex model is needed to capture transitivity in this network.

Figure 3.6 Number of triangles in 100 networks simulated based on the null model; X marks the number of triangles in the observed LHD network.



Adding Node Attributes

The first thing that researchers often consider to improve model fit is whether node attributes influence the likelihood of a link. In this case, do the characteristics of the LHDs and their leaders influence the likelihood they form communication ties? Descriptive statistics indicated that leader experience and jurisdiction population may influence the number of ties an LHD has. To examine the influence of these node attributes on the likelihood of a tie, they are added to the model as main effects. Hypotheses testing the main effects of LHD attributes on the likelihood of a connection might be worded as follows:

H_0: There is no association between jurisdiction population and the likelihood of an LHD to form ties.

H_A: There is an association between jurisdiction population and the likelihood of an LHD to form ties.

In adding main effects, it is important to use the command appropriate for the data type; Morris, Handcock, and Hunter (2008) present a comprehensive list of available statnet terms and specific directions for their use. For the LHD network model, leader experience will be included as a categorical variable, while jurisdiction population will be added as a continuous predictor (Command 19). In statnet, categorical main effects are added using `nodefactor`, and continuous main effects are included using `nodecov`. Nodefactor adds multiple statistics to the model, each one equal to the number of times a node with the specified attribute is at one end of an edge. The nodecov main effect term adds one network statistic to the model that sums the attribute of interest for the two nodes comprising the endpoints of each edge in the network. For example, if edge *ij* consisted of an LHD with a population of 1.2 million and an LHD with a population of .5 million, the edge would add 1.2 + .5 = 1.7 to the network statistic representing jurisdiction population.

Table 3.4 LHD network main effects model.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years")

Iterations:  20

Monte Carlo MLE Results:
                  Estimate Std. Error MCMC % p-value
edges             -6.22545    0.06353     NA < 1e-04 ***
nodecov.popmil     0.19663    0.01431     NA < 1e-04 ***
nodefactor.years.1 0.14379    0.04509     NA 0.00143 **
nodefactor.years.2 0.27927    0.04216     NA < 1e-04 ***
nodefactor.years.3 0.33689    0.03983     NA < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   36166  on 822398  degrees of freedom
          Deviance: 1103927  on      5  degrees of freedom

AIC: 36176    BIC: 36234
```

To interpret the results in Table 3.4, it is necessary to know the reference group for any categorical variables. In the statnet package, the reference group in an ERGM defaults to the first group in the list shown in the summary of the network (see Table 3.1). In this case, 1 to 2 years of experience is the reference group for leader experience. As usual for a logistic regression model, the main effects model omits the reference groups and produces estimates for the other categories. Changes to the reference group can be made using the base argument. For example, note the form of the nodefactor years term in Command 19: `nodefactor( 'years' )`. The term has no arguments other than the name of the attribute, so the default value (first category) will be automatically selected as the reference group. To select the last category, which is >10 years of experience, add a base argument of 4 to indicate that the fourth category should be used, like this: `nodefactor( 'years', base = 4 )`, and run the model and summary command as before (Command 20). The summary table produced will include estimates for the first three experience categories, omitting the highest experience category as the reference group. Notice that there are four statistically significant main effects in Table 3.4 (population in millions and three categories of leader experience).

One of the benefits of working in R is the ability of the user to modify the underlying code directly, depending on code access provided by the package authors. It may be useful to customize the information included in the model summary to fit reporting preferences. For example, if reporting test statistics is a standard practice in a field for reporting logistic regression results, the ergm summary function can be edited to report the test statistic for each coefficient every time the summary function is called, reducing the number of additional commands needed to produce this information. To revise the R code for the ergm summary function, use the`fix()` command (Appendix B, online).

Follow the instructions in the first section of Appendix B to modify the ergm summary code and rerun the summary line only in Command 19 to obtain a main effects model summary, including Wald test statistics. There is no need to reestimate the model; the edits made changed only what is printed in the model summary. As a reminder, access to the underlying code through the fix command may change if the package developers modify settings in future versions of statnet. Alternatively, to obtain the test statistics, the second half of Command 21 can be used (note that this code will only work after the first half of Command 21 has been run).

In the main effects model (Table 3.4), all estimates are significant and positive. This indicates an increased likelihood to form ties for LHDs when the LHD leader had more experience, compared to those with 1 to 2 years of experience, and for LHDs with larger jurisdiction populations. These results are consistent with the mixing matrices and correlation coefficient for jurisdiction population and leader experience.

In addition to general interpretation of the significance and direction of the coefficients, the coefficients and their corresponding standard errors can be transformed for interpretation as odds ratios and confidence intervals for each individual attribute (Command 21; Table 3.5). To transform the coefficient, simply use an exponential transformation (e^{θ}). Typically, odds ratios are reported with confidence intervals demonstrating the significance and precision of the estimate. The 95% confidence interval can be calculated using

$$95\%\text{CI}_{\theta} = e^{\theta \pm 1.96 s.e._{\theta}} \tag{13}$$

For confidence intervals (CIs) that are larger or smaller (e.g., 99% CI; 90% CI), replace the 1.96 with the appropriate value of *z,* which would be 2.56 and 1.28, respectively. Odds ratios are interpreted with respect to the reference group for categorical variables. For continuous variables, an odds ratio is defined as the increase (or decrease) in odds of the outcome with each one-unit increase in the variable of interest. An odds ratio more than 1 indicates increased odds, while odds ratios less than 1 indicate decreased odds. An odds ratio of 1 indicates no association; confidence intervals including 1, therefore, indicate nonsignificant relationships. Nonsignificant odds ratios and the odds ratio for the edges term may be reported but are not typically interpreted.

Table 3.5 Odds ratios and 95% confidence intervals for main effects model parameters.

	Lower	OR	Upper
edges	0.0017	0.0020	0.0022
nodecov.popmil	1.1836	1.2173	1.2519
nodefactor.years.1	1.0570	1.1546	1.2613
nodefactor.years.2	1.2173	1.3222	1.4361
nodefactor.years.3	1.2954	1.4006	1.5143

According to the main effects model, LHDs with leaders who have 3 to 5 years of experience are 1.15 times more likely to form a tie with a given LHD compared to LHDs with leaders with 1 to 2 years of experience, all other network properties held constant. The 95% confidence interval ranges from 1.06 to 1.26, indicating the range in which the true value of this relationship likely lies. Likewise, LHDs with leaders who have more than 10 years of experience are 1.40 times as likely to be connected to a given LHD as LHDs with leaders who have 1 to 2 years of experience, all else held constant.

In addition to the coefficients and standard errors used to calculate odds ratios and confidence intervals, the R-ergm procedure produces many other objects that may be useful in interpreting and reporting a model. To obtain a list of the objects produced with an R-ergm model, use the names command (Command 22). Each of the objects listed is described in detail in the R-ergm help documentation.

Odds ratios (ORs) are commonly reported for logistic models in some fields; it may be worth incorporating columns of odds ratios and confidence intervals into the default summary statistics to print when the summary function is called for any given model. The second section of Appendix B provides instructions for modifying the summary ergm table to include odds ratios and confidence intervals. Table 3.6 shows the expanded version of Table 3.4.

With the summary output, we can now report the results of our original hypothesis test:

H_0: There is no association between jurisdiction population and the likelihood of an LHD to form ties.

H_A: There is an association between jurisdiction population and the likelihood of an LHD to form ties.

Based on the main effects model, *the null hypothesis is rejected* in favor of the alternate hypothesis ($p < .05$). There is a significant association between the likelihood of forming a tie and LHD jurisdiction population. For every additional 1 million people living in a jurisdiction, the likelihood of forming a tie increases 1.22 times, all else held constant (OR = 1.22; 95% CI = 1.18–1.25).

Predicting Probabilities Using the Model

Like the null model, the main effects model can also be used to predict the probability of tie formation between any two network members. Because attributes of the network members are now included in the model, the predicted probability of a tie can now be calculated for network members with specified characteristics. For main effects predictors, the change statistic for each term is relatively straightforward. If the predictor is categorical, the value of the change statistic is 0, 1, or 2. If neither of the network members in the dyad has the characteristic of interest, it is 0. A value of 1 indicates that *one* of the nodes in the dyad has the characteristic; a 2 indicates that both nodes in the dyad have the characteristic. In the LHD network, then, for a link between two LHDs with highly experienced leaders, the coefficient for nodefactor.years.3 would be multiplied by the change statistic of 2 to represent two LHDs with this characteristic; for a link between an LHD with a very experienced leader and one with a new

Table 3.6 Expanded summary table for main effects model.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years")

Iterations:  20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC %     Lower        OR Upper p-value
edges              -6.225446   0.063528     NA  0.001747  0.001978 0.002 < 1e-04 ***
nodecov.popmil      0.196630   0.014310     NA  1.183626  1.217293 1.252 < 1e-04 ***
nodefactor.years.1  0.143793   0.045086     NA  1.056988  1.154645 1.261 0.00143 **
nodefactor.years.2  0.279269   0.042164     NA  1.217290  1.322163 1.436 < 1e-04 ***
nodefactor.years.3  0.336894   0.039827     NA  1.295417  1.400590 1.514 < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   36166  on 822398  degrees of freedom
          Deviance: 1103927  on      5  degrees of freedom

AIC: 36176    BIC: 36234
```

leader, the nodefactor.years.3 coefficient would be multiplied by 1 to represent the one LHD with the experienced leader; and so on. Following the notation of Hunter, Goodreau, and colleagues (2008), the δ corresponding to a categorical node attribute might be summarized as follows:

$$\delta_{\text{cat}} = \begin{cases} 2 \text{ if both nodes } i \text{ and } j \text{ have the characteristic} \\ 1 \text{ if } i \text{ or } j \text{ has the characteristic} \\ 0 \text{ if neither } i \text{ nor } j \text{ has the characteristic} \end{cases}$$

For continuous predictors, δ represents the sum of the characteristic for the two LHD leaders in the dyad. In the case of the LHD network, jurisdiction population is a continuous predictor, so if one LHD had 1 million constituents and the other had .5 million, the δ for population (δ_{popmil}) would be 1 + .5 or 1.5.

To find the predicted probability of a tie between (1) an LHD in a jurisdiction with 2 million people (popmil = 2) with a leader who had been there 7 years (years = 2) and (2) an LHD in a jurisdiction with 100,000 people (popmil = .1) and a leader there 1 year (years = 0), the coefficients in the Estimate column of Table 3.6 would be multiplied by the change statistics for each term.

$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\begin{pmatrix} \theta_{\text{edges}}\delta_{\text{edges}} + \theta_{\text{popmil}}\delta_{\text{popmil}} + \theta_{3-5\text{years}} \\ \delta_{3-5\text{years}} + \theta_{6-10\text{years}}\delta_{6-10\text{years}} + \theta_{>10\,years}\delta_{>10\,years} \end{pmatrix}$$

$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\left(-6.23{*}\delta_{\text{edges}} + .20{*}\delta_{\text{popmil}} + .34{*}\delta_{6-10\text{years}}\right)$$

$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\left(-6.84 * 1 + .20 * 2.1 + .34 * 1\right)$$

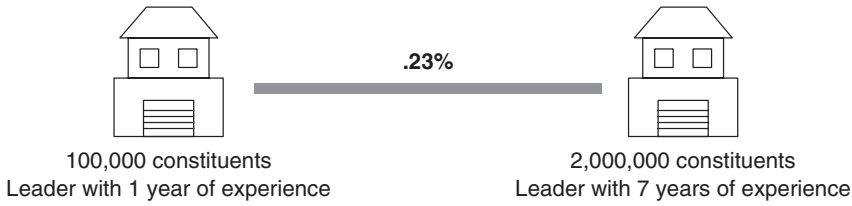
$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\left(-6.08\right) = \frac{1}{1 + e^{-(-6.08)}} = .0023$$

The probability of a tie between these two LHDs with the characteristics specified is .0023 or 0.23% (Figure 3.7). Although this may seem low, remember the density of this network was .0033, indicating that about one third of 1% of possible linkages actually exist in the LHD network. So, the likelihood of this particular connection is a little lower than might be expected.

Adding Interaction Terms

While node attributes account for characteristics of each individual network member, interaction terms for nodal attributes account for the attributes of both members of a dyad (Morris et al., 2008). The most commonly used interaction terms may be those accounting for homophily (two nodes sharing an attribute, e.g., both male) and heterophily (two nodes different on an attribute, e.g., one male and one female).

Figure 3.7 Probability of a tie between two LHDs based on the main effects model.



Interaction terms continue to treat each dyad as independent; models including interaction terms continue to be dyadic independence models. Remember, dyadic independence models assume that each dyad is independent of all other dyads in the model, so the likelihood of a link between Pam and Michelle would be considered independent from the likelihood of a link between Phil and Pam, even though Pam is in both dyads.

Based on the exploratory analysis, LHDs appear to form more connections with other LHDs in the same state and conducting the same programs. That is, there seems to be some homophily of state and programming in connected dyads across the LHD network. Interaction terms for each of these attributes can be used to test this hypothesis in a new model (Command 23).

The resulting model shows significant positive coefficients for state and programming homophily in the network. That is, two LHDs in the same state are more likely to be connected, as are two LHDs conducting the same programming (Table 3.7); homophily terms are indicated in the R commands and output by "nodematch" and are highlighted in Table 3.7. All of the main effects (nodefactor and nodecov terms) remain positive and significant. Note that main effects for programming are not entered into this model; because each LHD only has two possible values for each program (Y, N), it is not possible to enter both the main effects and interaction terms for a program given the limited degrees of freedom available. Many other terms are available for testing additional main effects and interactions (Goodreau et al., 2009; Morris et al., 2008).

Table 3.7 Model of LHD network including homophily terms.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years") + nodematch("hivscreen") +
    nodematch("nutrition") + nodematch("state")

Iterations:  20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC %      Lower         OR   Upper  p-value
edges              -9.537e+00  1.133e-01     NA  5.775e-05  7.212e-05   0.000  < 1e-04 ***
nodecov.popmil      3.643e-01  1.939e-02     NA  1.386e+00  1.440e+00   1.495  < 1e-04 ***
nodefactor.years.1  1.810e-01  4.743e-02     NA  1.092e+00  1.198e+00   1.315 0.000135 ***
nodefactor.years.2  3.249e-01  4.435e-02     NA  1.269e+00  1.384e+00   1.510  < 1e-04 ***
nodefactor.years.3  3.040e-01  4.207e-02     NA  1.248e+00  1.355e+00   1.472  < 1e-04 ***
nodematch.hivscreen 2.889e-01  4.668e-02     NA  1.218e+00  1.335e+00   1.463  < 1e-04 ***
nodematch.nutrition 2.754e-01  4.665e-02     NA  1.202e+00  1.317e+00   1.443  < 1e-04 ***
nodematch.state     6.279e+00  8.491e-02     NA  4.514e+02  5.332e+02 629.736  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   19568  on 822395  degrees of freedom
          Deviance: 1120525  on      8  degrees of freedom

AIC: 19584    BIC: 19677
```

There was some indication in the exploratory analyses that program homophily was different for LHDs conducting the program compared with those not conducting the program; LHDs were more likely to have contact with other LHDs conducting the same programs, but the converse was not necessarily the case. That is, LHDs *not* conducting the program were not necessarily more likely to be linked to other LHDs *not* conducting the program. Homophily can be estimated for each level of a categorical variable; this is called differential homophily. By specifying differential homophily for programs, homophily terms will be included separately for conducting and not conducting the program (Command 24; Table 3.8). Differential homophily is shown in the command by the use of diff=T following the name of the attribute in a nodematch term.

At this point, if you have not closed and reopened R since beginning analyses, you may be experiencing slowness in model estimation or even memory allocation errors. There are a few strategies you can use to improve the speed without closing and reopening R. First, you can remove any objects you no longer need. To do this, use the `ls()` command to list all objects currently open in R. From this list, identify objects you no longer need and use the command `remove( object )` to delete each one. Once you have removed objects you no longer need, you can use the garbage collector to clean up the memory by running the command `gc()`. Finally, once objects are removed and the garbage collector is done, you can increase the allocation of memory by using `memory.size( 8000 )`.

Note that the output now shows the category for each homophily term, highlighted in Table 3.8. The model including differential homophily terms demonstrates a significant increase in the likelihood of a tie between two LHDs both conducting the specified programming, but not between two LHDs that are both *not* conducting the programming.

It may be useful to keep only the terms for program homophily when the LHDs are both conducting the program and to drop homophily terms for LHDs not conducting programming. To do this, specify which terms you want to keep in the nodematch command. In this case, not doing a program is coded as "N" and doing a program is coded as "Y"; since N comes before Y, the first homophily term estimated by nodematch will be two LHDs not doing programming (N-N), and two LHDs doing programming (Y-Y) will be the second. To keep only the homophily term for two LHDs both conducting programs, add keep=2 to the nodematch command (Command 25).

Homophily and differential homophily change statistics are similar to main effects change statistics, although since the unit of interest is now the dyad, there are now only two possible values. Following Hunter, Goodreau, and colleagues (2008) and Goodreau and colleagues (2009), the change statistics can be denoted as follows.

Table 3.8 Homophily model with differential homophily terms for state and programming.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years") + nodematch("hivscreen",
    diff = T) + nodematch("nutrition", diff = T) + nodematch("state")

Iterations:  20

Monte Carlo MLE Results:
                       Estimate Std. Error MCMC %      Lower         OR   Upper  p-value
edges                -9.548e+00  1.131e-01     NA  5.714e-05  7.132e-05   0.000  < 1e-04 ***
nodecov.popmil        3.306e-01  2.009e-02     NA  1.338e+00  1.392e+00   1.448  < 1e-04 ***
nodefactor.years.1    1.757e-01  4.748e-02     NA  1.086e+00  1.192e+00   1.308 0.000215 ***
nodefactor.years.2    3.234e-01  4.445e-02     NA  1.267e+00  1.382e+00   1.508  < 1e-04 ***
nodefactor.years.3    3.468e-01  4.236e-02     NA  1.302e+00  1.415e+00   1.537  < 1e-04 ***
nodematch.hivscreen.N -2.571e-02  6.203e-02     NA  8.630e-01  9.746e-01   1.101 0.678461
nodematch.hivscreen.Y  4.490e-01  4.990e-02     NA  1.421e+00  1.567e+00   1.728  < 1e-04 ***
nodematch.nutrition.N  1.013e-02  8.302e-02     NA  8.585e-01  1.010e+00   1.189 0.902902
nodematch.nutrition.Y  2.495e-01  4.861e-02     NA  1.167e+00  1.283e+00   1.412  < 1e-04 ***
nodematch.state        6.313e+00  8.440e-02     NA  4.675e+02  5.516e+02 650.802  < 1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   19457  on 822393  degrees of freedom
          Deviance: 1120635  on     10  degrees of freedom


AIC: 19477    BIC: 19593
```

Table 3.9 Homophily model with differential homophily terms keeping only one category.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years") + nodematch("hivscreen",
    diff = T, keep = 2) + nodematch("nutrition", diff = T, keep = 2) +
    nodematch("state")

Iterations:  20

Monte Carlo MLE Results:
                      Estimate Std. Error MCMC %      Lower          OR   Upper  p-value
edges                -9.556e+00  1.100e-01     NA  5.707e-05  7.080e-05   0.000  < 1e-04 ***
nodecov.popmil        3.310e-01  2.005e-02     NA  1.339e+00  1.392e+00   1.448  < 1e-04 ***
nodefactor.years.1    1.756e-01  4.748e-02     NA  1.086e+00  1.192e+00   1.308 0.000216 ***
nodefactor.years.2    3.238e-01  4.443e-02     NA  1.267e+00  1.382e+00   1.508  < 1e-04 ***
nodefactor.years.3    3.463e-01  4.233e-02     NA  1.301e+00  1.414e+00   1.536  < 1e-04 ***
nodematch.hivscreen.Y 4.587e-01  4.352e-02     NA  1.453e+00  1.582e+00   1.723  < 1e-04 ***
nodematch.nutrition.Y 2.496e-01  4.504e-02     NA  1.175e+00  1.284e+00   1.402  < 1e-04 ***
nodematch.state       6.310e+00  8.411e-02     NA  4.666e+02  5.502e+02 648.811  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   19457  on 822395  degrees of freedom
          Deviance: 1120635  on      8  degrees of freedom

AIC: 19473    BIC: 19566
```

Homophily change statistic:

$$\delta_{\text{hom}} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ have the same value for a categorical covariate} \\ 0 \text{ otherwise} \end{cases}$$

Differential homophily change statistic:

$$\delta_{\text{diff}} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ have the same value for a certain category} \\ \quad \text{of a categorical covariate} \\ 0 \text{ otherwise} \end{cases}$$

Calculating the probability of connection between the two LHDs in Figure 3.7 can demonstrate the use of dyad-level terms. In addition to each LHD having its own characteristics for leader experience and jurisdiction population, the two LHDs in this dyad may match on state (i.e., both LHDs in Missouri) and nutrition programming but may not match on HIV screening (perhaps the larger LHD offers HIV screening, the smaller does not). There are 10 terms in this model; the full model is shown first, but only the terms that apply to the LHDs in question are shown with substituted values. The probability that the two LHDs are linked is calculated (for the sake of brevity, homophily is abbreviated as "Hom"):

$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\begin{pmatrix} \theta_{\text{edges}}\delta_{\text{edges}} + \theta_{\text{popmil}}\delta_{\text{popmil}} + \\ \theta_{3-5\text{years}}\delta_{3-5\text{years}} + \theta_{6-10\text{years}}\delta_{6-10\text{years}} + \\ \theta_{>10\text{years}}\delta_{>10\text{years}} + \theta_{\text{HIVHom}}\delta_{\text{HIVHom}} + \\ \theta_{\text{NutritHom}}\delta_{\text{NutritHom}} + \\ \theta_{\text{StateHom}}\delta_{\text{StateHom}} \end{pmatrix}$$

$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\begin{pmatrix} -9.56*\delta_{\text{edges}} - .33*\delta_{\text{popmil}} + .32*\delta_{6-10\text{years}} + \\ .25*\delta_{\text{NutritHom}} + 6.31*\delta_{\text{StateHom}} \end{pmatrix}$$

$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = \text{logistic}\left(-9.56*1 - .33*2.1 + .32*1 + .25*1 + 6.31*1\right)$$

$$P\left(Y_{ij} = 1 \mid n \text{ actors}, Y_{ij}^{c}\right) = .033$$

This is much higher than the probability calculated with the main effects model due primarily to the large coefficient for being in the same state. Based on this model, these two LHDs have a 3.3% chance of being connected (Figure 3.8).

Although many of the predictors in the model demonstrate statistical significance and reflect the patterns seen in exploratory analyses, giving face validity to the model, it is important to more systematically examine how well the estimated model actually captures observed network structure.

Model Fit

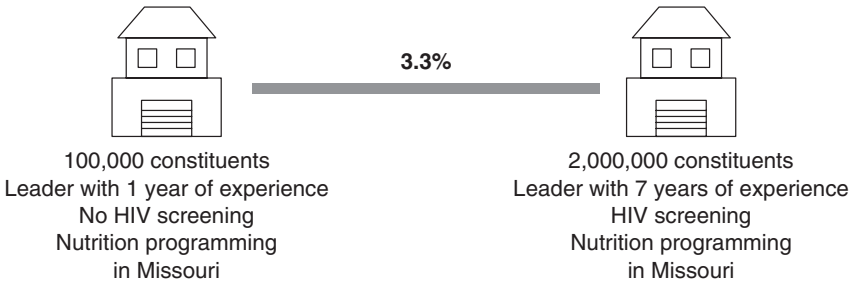
There are several ways to examine model fit for statistical network models. The simplest way is to compare the statistical measures of log-likelihood (LL) and related measures of deviance (–2LL), the Akaike information criterion (AIC), or the Bayesian information criterion (BIC). Log-likelihood is calculated by summing the differences between the predicted probabilities of Y_{ij} and the observed value of Y_{ij} (Field, 2009).

$$\text{log-likelihood} = \sum_{i=1}^{N} [Y_{ij} \ln(P(Y_{ij})) + \left(1 - Y_{ij}\right) \ln\left(1 - P\left(Y_{ij}\right)\right)] \tag{14}$$

In a nutshell, the LL sums the product of the differences between the predicted probability of a tie for each dyad and the actual observed presence or absence of a tie in each dyad.

For example, consider a situation where there was an existing connection (Y_{ij} = 1) between the two LHDs shown in Figure 3.6. The differential homophily model estimated a .23% chance of these two LHDs being

Figure 3.8 Probability of a tie between two LHDs based on the differential homophily model.



connected ($P(Y_{ij})$ = .0023). If a tie existed between these two in the observed network, the contribution of this dyad to the log-likelihood score for the model would be

$$Y_{ij}\ln(P(Y_{ij})) + (1 - Y_{ij})\ln(1 - P(Y_{ij}))$$

$$1 * \ln(.0023) + (1 - 1) * \ln(1 - .0023) = -6.07.$$

If there were no tie (Y_{ij} = 0) between these two LHDs, the contribution of this dyad to the overall log-likelihood score would be

$$Y_{ij}\ln(P(Y_{ij})) + (1 - Y_{ij})\ln(1 - P(Y_{ij}))$$

$$0 * \ln(.0005) + (1 - 0) * \ln(1 - .0005) = -.0023.$$

Because the predicted probability of a tie between these two was very low (.23%), the magnitude of the contribution to the log-likelihood score in the event that they were actually connected was much greater than the magnitude for no connections. The log-likelihood, therefore, grows in magnitude when the predicted probability is far from the observed value; the worse the predicted values are, the larger the magnitude of the log-likelihood. In quantifying the lack of fit, the LL is conceptually similar to the residual sum of squares in linear regression (Field, 2009). The LL is often negative, making comparisons intuitively more difficult. To combat this, deviance is often used in place of the LL. Deviance is simply the LL multiplied by –2, usually resulting in a positive value. Deviance is also considered a measure of lack of fit; the larger the deviance, the greater the lack of fit.

The deviance of a larger model can be compared with the deviance of a smaller *nested* model to determine whether the larger model is statistically significantly better than the smaller model in terms of fit. The difference between two deviance scores for nested models follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters in the two models. In this case, the main effects model has a deviance of 1,103,927 with 5 degrees of freedom; the differential homophily model has a deviance of 1,120,635 with 10 degrees of freedom. The difference between the two is 16,708 with 10 – 5 = 5 degrees of freedom. Comparing this value to a chi-squared distribution, we find a *p* value less than .0001, indicating that the differential homophily model is

a significantly better fit for the data than the main effects model ($\chi^2(5) = 16{,}708$; $p < .0001$). Adding homophily terms to the model significantly improved model fit.

AIC and BIC are two additional measures of fit that are variations on the deviance (–2LL) of a model. Deviance will always be smaller when more parameters are added to a model; the AIC and BIC account for this by penalizing models with more parameters that do not explain enough additional information to be considered a better fit (Akaike, 1973; Schwarz, 1978). In this way, they serve a somewhat similar function as the adjusted R^2 in linear regression, although they have no direct interpretation and therefore are used only to compare models. In Equation 15, p stands for the number of parameters in the model, and N represents the sample size.

$$\begin{aligned} \text{AIC} &= \text{Deviance} + 2p. \\ \text{BIC} &= \text{Deviance} + p*\ln(N). \end{aligned} \tag{15}$$

AIC and BIC are more flexible than deviance since they can be used to compare nonnested models. In the case of the LHD models, the null model had an AIC of 36,367, the main effects model had an AIC of 36,176, the differential homophily model AIC dropped to 19,477, and the second differential homophily model AIC dropped just slightly to 19,473. Based on these AIC values, the second differential homophily model is the best fit so far.

Given that these measures of fit were developed for data meeting the independence of observations assumption, other measures of model fit are generally considered a better choice for assessing how well an ERGM captures the observed network characteristics. So far, the null, main effects, and homophily models meet the assumption of independence for dyads, so deviance, AIC, and BIC are still useful; once more complex models that assume dyadic or other dependencies are developed, simulation-based assessments of model fit should be used instead.

One simple way to use simulation to assess model fit is to simulate a single network based on the model and compare the characteristics of the simulated network with the characteristics of the observed network. Simulation of one network based on each of the models developed so far can be examined and compared using this strategy (Command 26).

Note that there are some differences between the simulated networks and the observed network (Table 3.10); for example, they all have fewer isolates and fewer triangles than the observed LHD network (highlighted in the first row). While there is still room for improvement, it is also important

Table 3.10 Number of edges, nodes with degrees from 0 to 5, and triangles in the LHD network and simulated networks based on estimated models.

	edges	degree0	degree1	degree2	degree3	degree4	degree5	triangle
lhds	**2708**	**58**	**117**	**182**	**223**	**226**	**172**	**1437**
Null	2647	18	97	159	243	276	196	17
Main effects	2660	29	95	166	243	246	202	32
Homophily	2704	48	127	149	234	244	168	1223
Diff homophily	2707	45	125	169	224	231	174	1249
Diff homophily 2	2713	48	112	182	222	233	170	1249

to note that the characteristics of the simulated networks are getting considerably closer to those of the observed network with each addition of terms accounting for some underlying social process. Take triangles, for example; there is no triangle term in the models, but the number of triangles in the simulated networks goes from 17 in the simple random graph model (null) to 1,249 in the differential homophily models.

Increasing the number of simulations such as the one used for Table 3.10 can provide additional insight into model fit. Simulating 10 networks (or any number) from a model allows comparison of average network statistics from simulations with network statistics from the observed network (Command 27). For example, partial output from Command 27 shows the nodefactor.popmil (jurisdiction population) network statistic as ranging from 1,285.733 to 1,351.685 in 10 simulated networks from the second differential homophily model:

```
Stored network statistics:
      edges nodecov.popmil
 [1,]  2701       1333.655
 [2,]  2689       1315.555
 [3,]  2704       1338.370
 [4,]  2710       1351.434
 [5,]  2711       1345.763
 [6,]  2722       1347.097
 [7,]  2710       1351.685
 [8,]  2720       1329.238
 [9,]  2717       1287.713
[10,]  2719       1285.733
```

The network statistic for a continuous main effect (nodecov) equals the sum of the values of the variable for each time the node is at one end of an edge. To find the observed value for jurisdiction population in the observed LHD network, use Command 28 to obtain

```
nodecov.popmil
      1345.815
```

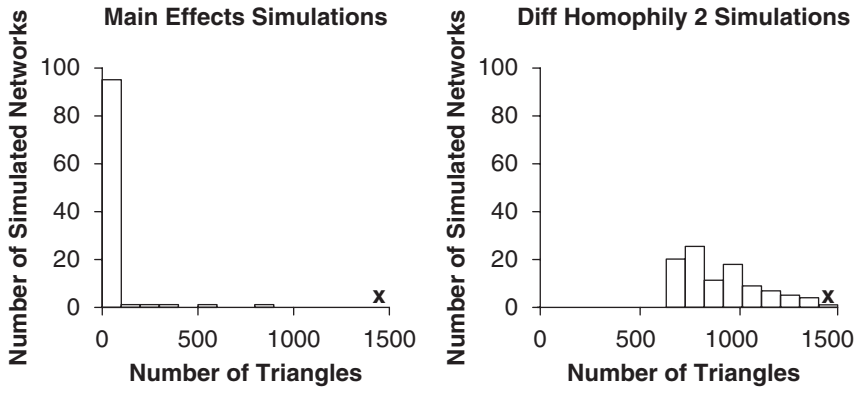
The network statistic in the observed LHD network for jurisdiction population is 1,345.815, which is in the range (1,285.733–1,351.685) for the simulated networks based on the second differential homophily model.

A comparison of how well the main effects and second differential homophily models capture the transitivity in the LHD network can be examined by comparing the observed number of triangles in the network with the distribution of the number of triangles in 100 simulated networks (or any number of simulations) based on the main effects model and 100 simulated networks based on the second differential homophily model (Commands 29–31; Figure 3.9) (Goodreau et al., 2008).

Of the 100 simulated networks from the main effects model, only 5 included more than 100 triangles. All simulated networks based on the second differential homophily model included 500 or more triangles, which was an improvement over the main effects model. However, the second differential homophily model still demonstrated an underestimation of the number of triangles in the observed LHD network, with 1,437 triangles (marked in Figure 3.9 by an X). This indicates that transitivity in the LHD network is not well represented by either model.

Model simulations built into the goodness-of-fit procedures can compare other network characteristics for simulated networks and the observed network. Comparisons of simulated and observed network degree distribution,

Figure 3.9 Distribution of triangles in 100 simulations for the main effects and second differential homophily models. The X marks the observed number of triangles in the LHD network.



edgewise shared partners, and dyadwise shared partners are built into a goodness-of-fit procedure within the ergm package. Using the results of this procedure, goodness of fit can be assessed in two ways. First, observed frequencies for each network statistic can be compared with frequencies in the simulated models (Command 32).

Table 3.11 Goodness of fit for the differential homophily model.

```
Goodness-of-fit for degree
   obs min   mean max MC p-value
0   58  32  58.78  78       0.96
1  117 111 134.76 176       0.12
2  182 159 190.23 217       0.62
3  223 177 205.45 233       0.16
4  226 151 188.28 237       0.08
5  172 120 152.33 196       0.28
6  104  86 116.71 144       0.22
7   67  59  84.98 109       0.04
8   35  29  55.45  74       0.02
9   25  20  37.61  60       0.08
10  26  15  24.70  42       0.80
11  14   8  14.97  26       0.84
12   8   2   9.03  17       0.88
13   6   1   4.43  10       0.54
14   8   0   1.96   7       0.00
15   4   0   1.05   4       0.08
16   3   0   0.50   3       0.02
17   1   0   0.50   3       0.76
18   1   0   0.28   3       0.50
19   1   0   0.23   2       0.44
20   1   0   0.22   1       0.44
21   0   0   0.11   1       1.00
22   1   0   0.11   2       0.20
23   0   0   0.10   1       1.00
24   0   0   0.11   1       1.00
25   0   0   0.07   1       1.00
26   0   0   0.04   1       1.00
29   0   0   0.01   1       1.00

Goodness-of-fit for edgewise shared partner
      obs min    mean  max MC p-value
esp0  696 923 1652.45 1808       0.00
esp1  750 647  723.50  805       0.56
esp2  630 153  232.33  578       0.00
esp3  382  33   63.65  322       0.00
esp4  156   5   15.72  109       0.00
esp5   56   0    4.50   47       0.00
esp6   25   0    1.04   18       0.00
esp7    8   0    0.32    7       0.00
esp8    3   0    0.09    2       0.00
esp9    0   0    0.05    1       1.00
esp10   1   0    0.03    1       0.06
esp11   1   0    0.03    1       0.06
```

```
Goodness-of-fit for dyadwise shared partner
          obs     min       mean     max MC p-value
dsp0   813034 811054 811708.89 812789        0.00
dsp1     6329   6795   8477.22   9143        0.00
dsp2     1928   1543   1767.79   1929        0.02
dsp3      732    270    367.36    649        0.00
dsp4      253     40     66.60    204        0.00
dsp5       80      3     12.55     71        0.00
dsp6       33      0      2.03     27        0.00
dsp7        9      0      0.36      7        0.00
dsp8        3      0      0.09      2        0.00
dsp9        0      0      0.05      1        1.00
dsp10       1      0      0.03      1        0.06
dsp11       1      0      0.03      1        0.06
```

These comparisons include five columns of information: obs, min, mean, max, and MC *p* value (Table 3.11). The first column lists the value of each statistic (degree, ESP, DSP). The *obs* column shows the number of nodes in the observed LHD network with the value listed in the first column. *Min* shows the minimum number of nodes with the specified degree, ESP, or DSP across the *simulated* networks. The *mean* column shows the average number with the value of degree, ESP, or DSP across the *simulated* networks. The *max* column shows the maximum number of degree, ESP, or DSP with each value in the *simulated* networks. The *MC p value* column is the *proportion of the simulated values of the statistic that are at least as extreme as the observed value.* Large values of the MC *p* value are indicators that the simulated networks are similar to the observed network on the characteristic of interest (i.e., not significantly different). Small *p* values show a difference between observed and simulated frequencies; *p* values less than .05 would therefore be interpreted as demonstrating a significant difference between the simulated and observed networks. This would indicate the model is not fitting the data well. All *p* values less than .05 in Table 3.11 are shaded, showing the observed network characteristics that the simulations failed to accurately represent.

The goodness of fit for degree in Table 3.11 shows that the observed LHD network had 58 isolates (nodes with a degree of 0) according to the first row of values in the table. The simulated networks had an average of 58.78 isolates and a range of 32 to 78 isolates. The range and average number of isolates show that the simulated networks are doing a good job of capturing this observed network characteristic. Consequently, the MC *p* value for a degree of 0 is .96; the observed network and the simulated networks are not significantly different in the number of nodes that have a

degree of 0. The simulated networks are good at capturing the number of nodes for most values of degree. One hundred seventeen nodes in the observed network have a degree of 1, and 134.76 nodes on average in the simulated networks also have a degree of 1. The *p* value of .12 indicates that the simulated networks represented this well; the observed and simulated networks are not significantly different. Generally speaking, the fewer the *p* values less than .05 in these tables, the better the model fits.

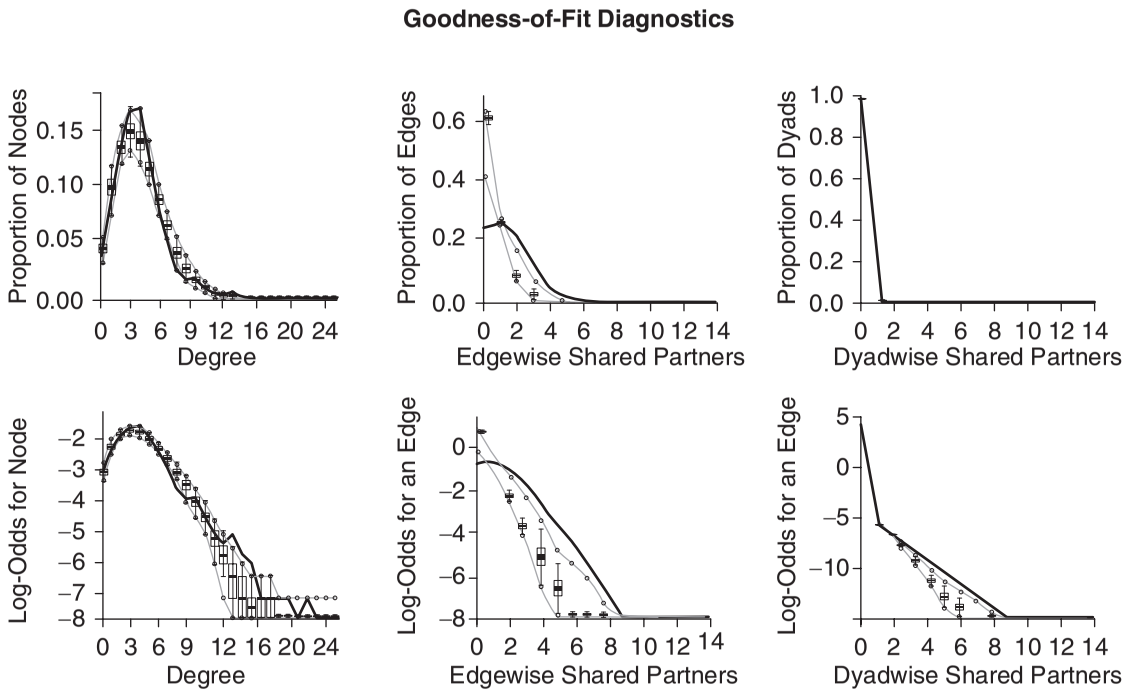
Table 3.11 demonstrates some lack of fit for ESP and DSP. Few of the edgewise shared partner frequencies were well captured, and only dyads with 9, 10, or 11 shared partners (DSP = 9, DSP = 10, DSP = 11) were appropriately represented by the second differential homophily model. Given that ESP and DSP are indicators of transitivity, the poor fit of the simulations for these two measures is consistent with the lack of triangles in the simulated networks compared with the observed network (Figure 3.9). This is further evidence that the second differential homophily model is not capturing observed transitivity.

It is important to note that the tables do not show all the values possible for each network statistic. For example, each node in the LHD network could have up to 1,282 links to others since the network has 1,283 members, but the output above only shows a degree of 0 to 29. Command 33 can be used to see all rows in each table if this is of interest; the output will be extremely long.

The goodness-of-fit procedure in statnet also includes triangle census and geodesic distance options. For measures not available in the built-in process, a separate simulation procedure can be used to assess how well the model captures the measures (e.g., see Commands 29–31 for triangle simulations).

In addition to comparing the frequencies of the observed and simulated values for each network statistic, the goodness-of-fit procedure produces graphics. Instead of comparing the *frequency* of each network statistic value, the graphics compare the *proportion* of nodes in the observed network with the *proportion* of nodes in the simulated networks with the same characteristic (Command 34). Or, if the plot setting is changed, the graphics compare the log-odds for each parameter in the observed network and the range of log-odds in the simulated networks. For example, the tables compare the number of isolates in the observed network ($n = 58$) with the number of isolates in the simulated networks ($n = 32–78$); the graphs compare the proportion of nodes that are isolates (4.5%) with the proportion that are isolates in the simulations (Figure 3.10, top row). The bottom row in Figure 3.10 compares the log-odds of each measure in the observed networks with the range of log-odds of the

Figure 3.10 Goodness-of-fit for measures from simulated networks based on the differential homophily model. Black lines represent the observed value; gray lines and boxplots represent simulated network measures. Top row shows proportions, while the bottom row shows log-odds.



same measure across simulated networks. Because the fit is easier to discern in the log-odds plots, this format will be used for remaining goodness-of-fit plots.

In Figure 3.10, the thick black line represents the observed LHD network, and the gray lines show the 95% confidence interval of simulated network measures. When the black line falls between the gray lines, the simulated networks are capturing the characteristics of the observed network. In this case, based on simulations, degree and DSP seem well explained by the model, but there are some problems with fit for ESP.

Note that, although there should be consistency in model fit between Table 3.11 and Figure 3.10, these two displays of simulated data are demonstrating different things. Table 3.11 compares *frequencies* of specific values of statistics between observed and simulated networks, while the Figure 3.10 graphs compare *proportions* (or log-odds) of observed and simulated networks with specific values of statistics.

Three ways to examine goodness of fit have been demonstrated:

1. AIC and BIC are included in the modeling output and can be used to compare models with one another, with lower AIC and BIC indicating better fit.

2. Simulating one or more networks and comparing the characteristics to those of the observed network can provide some insight into how well basic characteristics (e.g., degree and triangles) are being captured by the model.

3. Goodness-of-fit procedures in statnet provide tables and graphics comparing observed measures to simulated measures for the structural characteristics of degree, distance, edgewise shared partners, dyadwise shared partners, and triangle census. The tables provide a p value and the graphs provide a confidence interval that can be used to determine whether the observed and simulated measures come from the same distribution. These built-in procedures are similar to the second strategy in this list.

Although the second differential homophily model has so far demonstrated better fit than the main effects model based on several measures, it was not generally a great fit for the observed network. This is a common occurrence for dyadic independence models; although the maximum likelihood estimation procedure finds the model with highest possible probability of replicating the observed network, this might still be a very low probability (Hunter, Goodreau, et al., 2008). The goodness-of-fit measures show that the models have failed to accurately represent transitivity through main effects

and homophily terms. Terms accounting for underlying distributions and complex dependencies may aid in improving model fit.

Adding Dependence Terms

To account for complex dependencies in an observed network, Snijders and colleagues (2006) proposed three terms subsequently modified by Hunter and Handcock (2006) to simplify interpretation: geometrically weighted degree (GWD), geometrically weighted edgewise shared partnerships (GWESP), and geometrically weighted dyadwise shared partnerships (GWDSP). The modified terms were implemented in statnet and are used here to estimate dependence models. The three modified terms accounted for the degree distribution and transitivity stemming from complex patterns of dependence in observed networks (see Chapter 2 for more information on these terms).

The maximum likelihood estimation used for the dyadic independence models is computationally prohibitive for dyadic dependence models. For instance, calculating the constant in Equation 6 would require summing over all possible network configurations, which consists of $2^{\binom{n}{2}}$ networks. For a network with just nine nodes, this is 68,719,476,736 configurations (Cranmer & Desmarais, 2011). Models incorporating dyadic dependence terms therefore use a Markov chain Monte Carlo (MCMC) parameter estimation algorithm to calculate an approximate log-likelihood (Snijders, 2002). By default, maximum pseudolikelihood is used to determine the starting values for model estimation. The MCMC algorithm then works by selecting a network from all possible realizable networks, randomly selecting a dyad or dyads from the network, toggling the dyad or dyads from 0 to 1 or from 1 to 0, and comparing the new network to the pretoggle network to determine if it is a better fit. The algorithm then accepts the new network or keeps the pretoggle network and draws another random dyad or dyads to toggle for the next proposal. This propose-compare-decide process is repeated until the specified MCMC chain length is reached (Morris et al., 2008).

As discussed in Chapter 2, even models incorporating homophily and other terms may exhibit problems with degeneracy, indicating that observed structures were not adequately captured. Degeneracy in network modeling is often manifested by a model that produces simulated networks that are either nearly empty or nearly complete (see Robins et al., 2007, Figure 1, for an excellent visual example). Adding geometric terms accounts for structures that, in early statistical network modeling, often resulted in these degenerate models; the geometric terms (GWD, GWDSP, GWESP) are part of *what* is being modeled.

In addition to the *what,* changes can be made to *how* the model is estimated. Specifically, several additional steps can be taken to reduce the likelihood of model nonconvergence, including selecting adequate MCMC sample size, burn-in, and interval (Goodreau et al., 2008; Morris et al., 2008). The sample size controls how many networks will be sampled in the MCMC chain (the length of the chain described in the previous paragraph), the burn-in indicates how many networks to ignore at the beginning of the sample, and the interval specifies how many networks to skip over between sampled networks. If a network model exhibits signs of degeneracy, increasing the value of each of these settings and reestimating the model may aid with convergence. Note that, with increased MCMC sample size, most users will notice a *substantial* increase, often hours, in the amount of time it takes R to estimate a model. These three settings are specified in the control.ergm command. To produce model estimates that can be replicated, a seed value can also be added to the command, directing the model to start at the same place each time. The seed value is also added as part of the control.ergm command.

Following advice from Goodreau and colleagues (2008; see Chapter 2) that α be selected by starting at .1 and increasing until the log-likelihood ceases to improve, models with geometric terms were estimated for several values of α starting with α = .1 (Command 35). For readers following along using commands, please note that *each* of these models may take a *very long time* to converge (i.e., hours). While parallel processing is possible in R in some forms, for many users, R has limitations that permit the use of only one core regardless of how many cores a computer has. Without advanced computing capabilities, the speed of model estimation is difficult to increase. The statnet development team has begun to work on adding parallel processing functionality; see ergm-parallel in the R documentation for the ergm package (http://cran.r-project.org/web/packages/ergm/ergm.pdf).

Although graphic measures of fit are preferred for dependence models, AIC and BIC are usually consistent with the graphic measures and may suffice for quick comparisons during the selection of α. The model estimates from Command 35 resulted in AIC of 18019, 17943, 17875, 17814, 17759, 17732, 17700, **17660,** 17667 for α of .1, .2, .3, .4, .5, .6, .7, **1,** and 1.1, respectively. The BIC demonstrated a similar trajectory. In this case, the best fit based on AIC and BIC was α = 1 (see bold AIC above). The three geometrically weighted terms using α = 1 were included with the other variables from the second differential homophily model to estimate a dyadic dependence model (Table 3.12).

The dependence model includes positive significant coefficients indicating an increased likelihood of a tie for state homophily, program homophily, jurisdiction population, years of leader experience, GWD, and GWESP.

Table 3.12 Dependence model with $\alpha = 1$ specified.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodefactor("years") + nodecov("popmil") + nodematch("hivscreen",
    diff = T, keep = 2) + nodematch("nutrition", diff = T, keep = 2) +
    nodematch("state") + gwdegree(1, T) + gwesp(1, T) + gwdsp(1,
    T)

Iterations:  20

Monte Carlo MLE Results:
                      Estimate Std. Error     MCMC %      Lower          OR    Upper p-value
edges                -1.007e+01  3.621e-01  7.900e+01  2.090e-05  4.250e-05    0.000 < 1e-04 ***
nodefactor.years.1    1.426e-01  4.983e-02  1.600e+01  1.046e+00  1.153e+00    1.272 0.00420 **
nodefactor.years.2    2.541e-01  4.691e-02  1.800e+01  1.176e+00  1.289e+00    1.413 < 1e-04 ***
nodefactor.years.3    2.969e-01  4.795e-02  2.100e+01  1.225e+00  1.346e+00    1.478 < 1e-04 ***
nodecov.popmil        2.005e-01  2.304e-02  1.000e+00  1.168e+00  1.222e+00    1.278 < 1e-04 ***
nodematch.hivscreen.Y 1.750e-01  4.200e-02  1.000e+00  1.097e+00  1.191e+00    1.293 < 1e-04 ***
nodematch.nutrition.Y 1.897e-01  4.208e-02  1.300e+01  1.113e+00  1.209e+00    1.313 < 1e-04 ***
nodematch.state       5.018e+00  1.210e-01  4.700e+01  1.192e+02  1.511e+02 191.504 < 1e-04 ***
gwdegree              1.851e+00  2.193e-01  6.800e+01  4.144e+00  6.369e+00    9.790 < 1e-04 ***
gwesp.fixed.1         9.588e-01  2.941e-02  2.900e+01  2.463e+00  2.609e+00    2.763 < 1e-04 ***
gwdsp.fixed.1        -3.936e-02  1.377e-02  6.400e+01  9.358e-01  9.614e-01    0.988 0.00427 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   17638  on 822392  degrees of freedom
          Deviance: 1122455  on     11  degrees of freedom

AIC: 17660    BIC: 17788
```

A positive and significant coefficient for a geometric term indicates that, given the distribution of degree, ESP, or DSP in the observed network, the likelihood of a tie between any two given LHDs is greater than would happen by chance, all else held constant. Calculations demonstrating the influence of these terms on tie likelihood are found in the next section.

MCMC Model Diagnostics

In addition to checking model fit using the strategies discussed, model diagnostics can help determine whether the estimating algorithm has converged or there are degeneracy problems and if the model itself or the estimation settings need adjustment. The first strategy is to examine the changes in log-likelihood during the iterations; these are printed as the estimation is proceeding if `verbose = T` is included in the model command, as in Command 35. The amount the log-likelihood improves is an indicator of how far the iterations were from the starting values; large improvement numbers indicate the starting values for the estimation were off. The MCMC algorithm stops if the changes in LL are greater than 20 for any iteration, which may indicate a degenerate model or a model with starting values that are very far from the final estimates. Generally, changes should be small and decreasing with each iteration.

In addition to examining changes in LL, it may be useful to examine graphic MCMC diagnostics (Command 36). Graphic diagnostics show what is going on in the model during the final iteration (Figure 3.11); the graphs on the left in Figure 3.11 show the MCMC chain as a time series for each statistic in the model, while the right shows the same chain in a histogram (Goodreau et al., 2008).

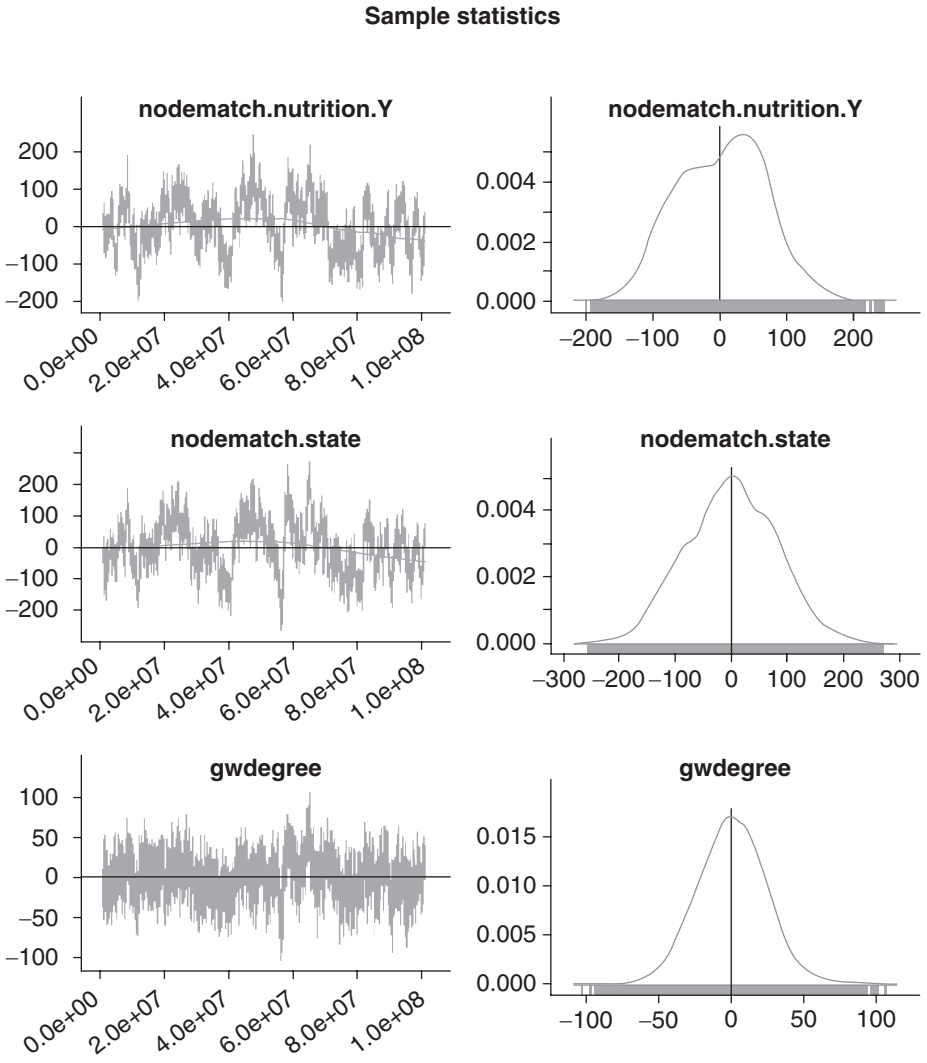
If the model has converged, these graphics should show each statistic varying stochastically around a mean of 0, where 0 represents the value of the statistic in the observed data. In this case, the graphics appear to be varying stochastically around 0 for most of the measures, with the exception of nutrition programming homophily and few others that have a slight skew. Overall, the diagnostics indicate a stable model.

For interested readers, more technical information about the MCMC diagnostics implemented in statnet can be found in Plummer and colleagues (Plummer, Best, Cowles, & Vines, 2006), and information about MCMC diagnostics in general can be found in Cowles and Carlin's (1996) review of the topic.

Curved Exponential Family Model

Rather than selecting an α a priori, the α resulting in the best-fitting model can be estimated during model estimation. Models that estimate the α, rather than specifying it a priori, are called curved exponential family

Figure 3.11 MCMC diagnostics for several terms in the dependence model with $\alpha = 1$.



models (CEF); try reestimating the model as a CEF model using Command 37. Note that the process of selecting α based on fitting several models shown previously may be considered, by some, as an alternate way to estimate the α for a CEF model (Hunter, 2007).

Table 3.13 Curved exponential family model.

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodefactor("years") + nodecov("popmil") + nodematch("hivscreen",
    diff = T, keep = 2) + nodematch("nutrition", diff = T, keep = 2) +
    nodematch("state") + gwdegree(1, F) + gwesp(1, F) + gwdsp(1,
    F)

Iterations:  20

Monte Carlo MLE Results:
                      Estimate Std. Error     MCMC %      Lower         OR   Upper p-value
edges                -9.118e+00  7.742e-01  8.700e+01  2.405e-05  1.097e-04   0.001 < 1e-04 ***
nodefactor.years.1    1.312e-01  4.234e-02  2.000e+00  1.049e+00  1.140e+00   1.239 0.00195 **
nodefactor.years.2    2.402e-01  4.142e-02  7.000e+00  1.172e+00  1.272e+00   1.379 < 1e-04 ***
nodefactor.years.3    2.791e-01  4.081e-02  5.000e+00  1.220e+00  1.322e+00   1.432 < 1e-04 ***
nodecov.popmil        2.253e-01  3.043e-02  2.400e+01  1.180e+00  1.253e+00   1.330 < 1e-04 ***
nodematch.hivscreen.Y 2.079e-01  3.603e-02  1.000e+00  1.147e+00  1.231e+00   1.321 < 1e-04 ***
nodematch.nutrition.Y 1.884e-01  4.348e-02  2.000e+00  1.109e+00  1.207e+00   1.315 < 1e-04 ***
nodematch.state       4.923e+00  9.404e-02  5.000e+00  1.143e+02  1.374e+02 165.187 < 1e-04 ***
gwdegree              1.109e+00  1.841e-01  3.100e+01  2.113e+00  3.031e+00   4.348 < 1e-04 ***
gwdegree.decay        8.380e-01  8.388e-01  8.600e+01  4.467e-01  2.312e+00  11.965 0.31774
gwesp                 9.690e-01  3.487e-02  6.000e+00  2.461e+00  2.635e+00   2.822 < 1e-04 ***
gwesp.alpha           9.451e-01  3.057e-02  0.000e+00  2.424e+00  2.573e+00   2.732 < 1e-04 ***
gwdsp                -8.237e-02  7.573e-02  6.900e+01  7.939e-01  9.209e-01   1.068 0.27677
gwdsp.alpha           1.822e+00  1.196e-01  3.600e+01  4.894e+00  6.187e+00   7.822 < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   16987  on 822389  degrees of freedom
          Deviance: 1123106  on     14  degrees of freedom

AIC: 17015    BIC: 17178
```

The estimated α values for the geometric terms in the CEF model are shown just after the corresponding geometric term. In this case, the estimated α for GWD is listed in Table 3.13 as gwdegree.decay and is .838, the GWESP α is listed as gwesp.alpha and is .9451, and the GWDSP α is listed as gwdsp.alpha and is 1.822. The three estimated α values are highlighted in Table 3.13.

There was some difference in the magnitude of the covariates between the dependence and CEF models. The AIC and BIC both decreased in the CEF model as compared with the dependence model, indicating an increase in model fit. Adding the dependence and CEF models to Table 3.11, we can examine the change in fit for some of the basic network structures (Command 38; Table 3.14).

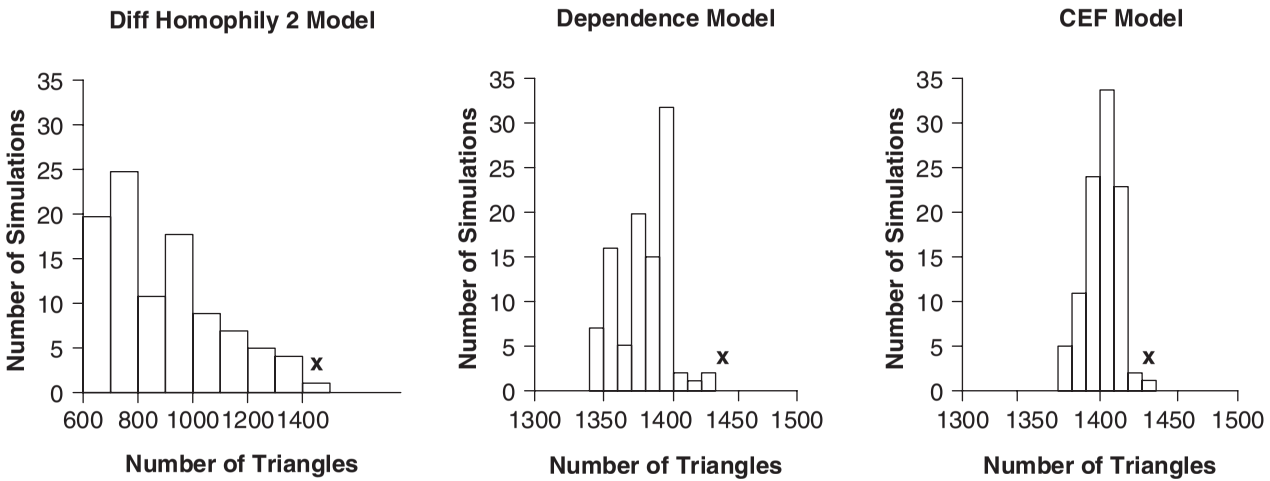
Table 3.14 Network measures for the LHD network and a network simulated from each of the models.

	edges	degree0	degree1	degree2	degree3	degree4	degree5	triangle
LHD	2708	58	117	182	223	226	172	1437
Null	2647	18	97	159	243	276	196	17
Main effects	2660	29	95	166	243	246	202	32
Homophily	2704	48	127	149	234	244	168	1223
Diff homophily	2707	45	125	169	224	231	174	1249
Diff homophily 2	2713	48	112	182	222	233	170	1249
Dependence	2589	26	129	207	254	207	177	1151
CEF model	2652	54	135	218	195	198	150	1306

The network simulations show several differences in model fit, including a lack of triangles in the less complex models. The CEF model appeared to come the closest to capturing triangles, while the three dyadic independent homophily models were best at capturing the total edges in the network, and the second differential homophily model was best at capturing the degree distribution. While these results appear to indicate that there is not a single best model to capture all characteristics of the network, this is just one simulation from each model and should not be the only measure of fit examined.

In 100 simulations, the number of triangles is underestimated in networks based on the differential homophily, dependence, and CEF models where most of the simulated networks fall to the left of the X that points to the observed number of triangles (Command 39; Figure 3.12). However, the simulations from the dependence and CEF models were closer to capturing the number of triangles in the observed network, with the CEF model appearing to be the most consistently close to the observed

Figure 3.12 Histograms showing the distribution of triangles per network in 100 simulations for the differential homophily model and the two dependence models. X shows observed number of triangles in the LHD network.



value. Of the 100 simulated networks, 0 networks from the dependence model and CEF model simulations had more triangles than the observed data; just one from the CEF model had exactly 1,437 triangles. Neither of these models is doing a perfect job at capturing the transitivity in the LHD network, although both appear better than the dyadic independence models.

Additional plots of degree, distance, ESP, and DSP demonstrate fit for the dependence model and the CEF model (Command 40; Figure 3.13). Both appear to be reasonably well fitting, with the CEF perhaps capturing degree more closely.

Model Selection

A comparison of multiple statistical and graphic fit indicators for the seven models (null, main effects, homophily, differential homophily, second differential homophily, dependence, and CEF) has demonstrated that the CEF model has the best fit. The biggest increases in fit during the modeling process came from adding the *homophily terms* to account for LHDs with similar characteristics (state, programming) being connected and adding the *dependence terms* to account for degree distribution and transitivity. In addition to examining the fit statistics and graphics from all models, there may be some utility in examining the changes in model fit represented visually in the network through the stages of model building for smaller networks; for networks as large as the LHD network, it is difficult to discern much difference once the homophily effects were added (Command 41; Figure 3.14).

Although the homophily and CEF model simulation networks appear similar, note the similarity in clustering patterns by program in the CEF model simulation and in the LHD network. The CEF model would therefore likely be selected as the final model for reporting purposes, given the lower AIC and BIC and the demonstration of fit in the simulation procedures. A figure like Figure 3.14 or a table like Table 3.15 displaying the model-building process might be used to demonstrate model development and as a rationale for the selection of a final model.

Interpreting the Results for the Dependence Model

The dependence model yielded significant homophily terms for HIV screening programs, nutrition programs, and state. Connections were 1.23 times as likely in dyads where both LHDs were conducting HIV screening compared with other dyads. Being in the same state was significantly related to ties (OR = 137.4; 95% CI = 114.3–165.2). Programming homophily was significant for nutrition programming,

Figure 3.13 Goodness-of-fit plots for the dependence (top) and CEF (bottom) models.

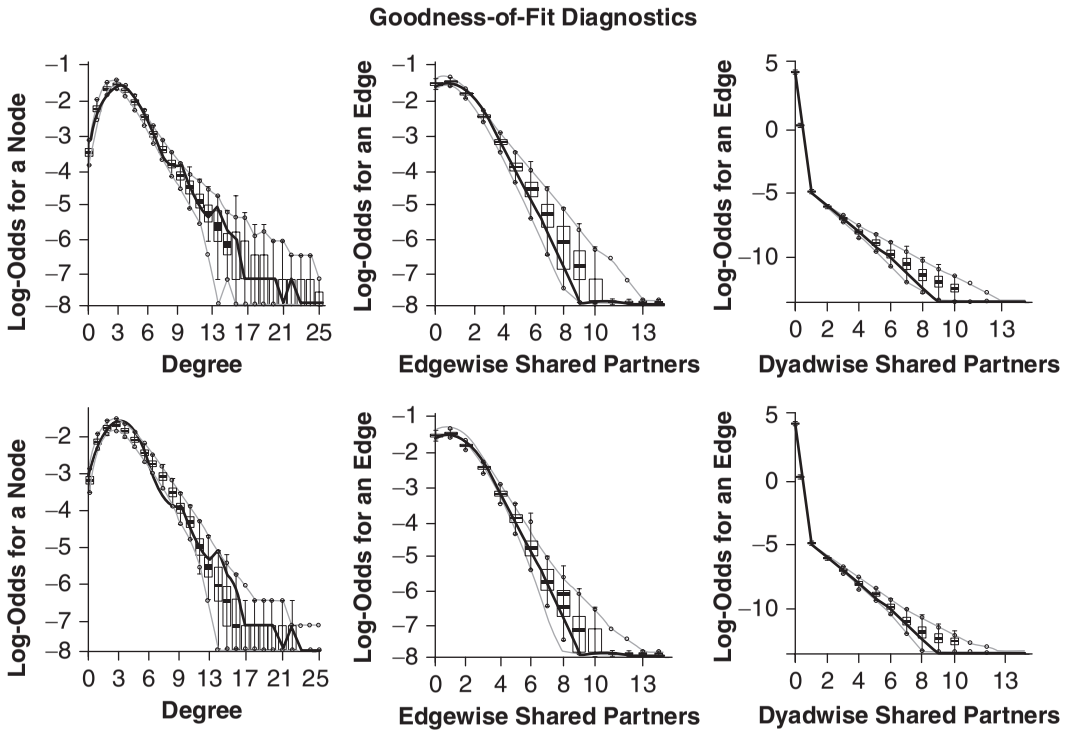
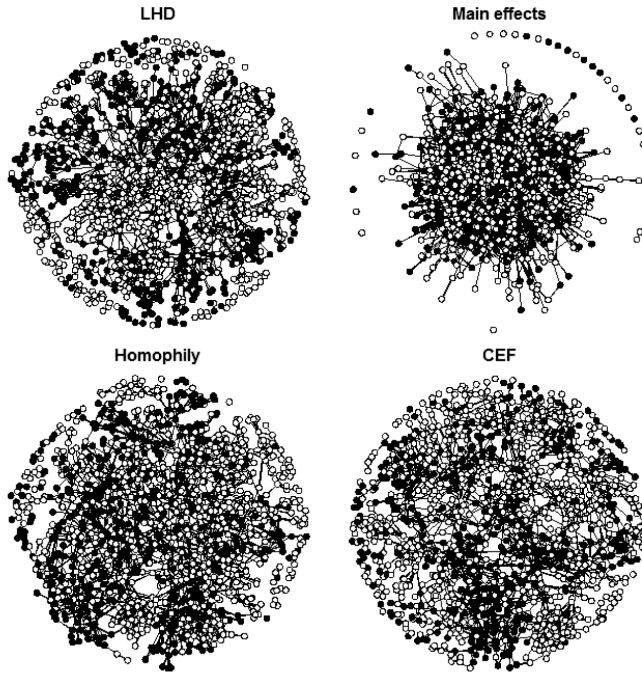


Figure 3.14 The observed LHD network and simulated networks from three of the models.



where two LHDs conducting nutrition programming were 1.21 times (95% CI = 1.11–1.32) more likely to form a connection than dyads where both LHDs were not conducting nutrition programming. Table 3.15 summarizes model development, showing coefficient estimates and standard errors for four of the models; the same table could be reported with odds ratios and confidence intervals depending on which is more useful for the audience.

Once GW terms are added to the model, predicting the probability of a tie between any two network members becomes more complex due to the calculations and interpretation challenges of the change statistic for each geometric term. In the case of the GWD term, if a single edge were added to the network, the $D_i(y)$ and $D_{i+1}(y)$ terms added as part of the summation of the weighted degrees on the right-hand side of Equation 9 would be

Table 3.15 Summary of model results for the null, main effects, differential homophily 2, and CEF models.

	*Estimate (*SE*)*			
	Null Model	*Main Effects*	*Differential Homophily 2*	*CEF*
Edges (constant)	–5.71 (.02)	–6.23 (.06)	–9.56 (.11)	–9.12 (.77)
Main effects				
Population (millions)		.20 (.01)	.33 (.02)	.23 (.03)
Years experience				
1–2		Reference	Reference	Reference
3–5		.14 (.05)	.18 (.05)	.13 (.04)
6–10		.28 (.04)	.32 (.04)	.24 (.04)
11+		.34 (.04)	.35 (.04)	.28 (.04)
Homophily				
State			6.31 (.08)	4.92 (.09)
Conducts nutrition program			.25 (.05)	.19 (.04)
Conducts HIV screening			.46 (.04)	.21 (.04)
Structural terms				
GWD				1.11 (.18)
GWESP				.97 (.03)
GWDSP				–.08 (.08)
Fit				
AIC	36,367	36,176	19,473	17,015
BIC	36,379	36,234	19,566	17,178

replaced by $D_i(y) - 1$ and $D_{i+1}(y) + 1$. To examine how the addition of this edge influences the likelihood of a graph (all else held constant), the change in odds for the graph can be examined by substituting the old and new degrees into Equation 8 (Hunter, 2007). Readers interested in additional detail in demonstrating how the change statistics for GWD follow from the overall model through this substitution can consult Hunter (2007) in deriving the following:

$$\frac{P\left(Y_{ij}=1\right)_{after}}{P\left(Y_{ij}=1\right)_{before}} = \exp\left\{\theta\left(1-e^{-\alpha}\right)^{i}\right\} \tag{16}$$

Note that $P\left(Y_{ij}=1\right)_{before}$ is shorthand for $P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)_{before}$.

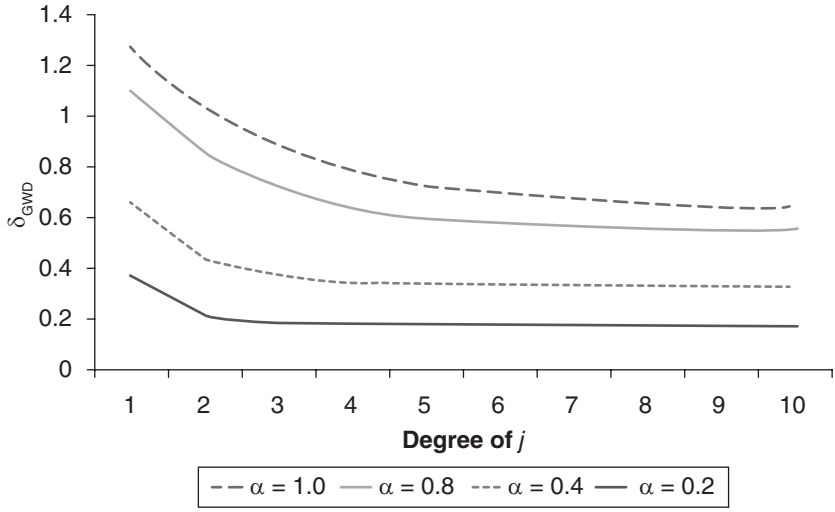
Since the addition of a new edge would increase the degree of both nodes it attached to, the increase would instead be $\theta[\left(1-e^{-\alpha}\right)^{i}+\left(1-e^{-\alpha}\right)^{j}]$ (Hunter, 2007). The change statistic for the GWD term is therefore defined as

$$\delta_{\text{GWD}} = \left(1-e^{-\alpha}\right)^{i}+\left(1-e^{-\alpha}\right)^{j} \tag{17}$$

Remember, as degree increases, $\left(1-e^{-\alpha}\right)^{i}$ decreases geometrically. So, with a positive and significant θ_{GWD}, the log-odds of a tie increases for all degree values of *i* and *j,* but this increase would be of smaller magnitude when *i* and *j* already have higher degree. At some point, the increase levels off and stays constant at any higher degree level for *j.* For a smaller α, this leveling off happens more quickly (Figure 3.15). The GWD network statistic weights higher degrees more, resulting in a larger value of the statistic for networks with more high-degree nodes. The corresponding change statistic demonstrates a preference for adding edges that is strongest in low-degree nodes.

There is more complexity in how overall network structures change for GWESP and GWDSP, since adding a tie will change the number of ESP and DSP not just for the two incident nodes but also for other nodes across the network. Readers interested in how the change statistics for GWESP

Figure 3.15 Change in log-odds of a tie between *i* and *j* when *i* has one link (degree = 1) and *j* has different degree values for several levels of α.



and GWDSP follow from the overall model can consult Hunter (2007) in deriving the following:

$$\delta_{\text{GWESP}} = \left(1 - e^{-\alpha}\right)^{ij_{ESP}} \tag{18}$$

$$\delta_{\text{GWDSP}} = \left(1 - e^{-\alpha}\right)^{ij_{DSP}} \tag{19}$$

In the LHD dependence model with α = 1.0, the calculation

$$1 - e^{-\alpha} = 1 - e^{-1.0} = .63$$

can be substituted into Equations 17 to 19 to determine the change statistics for the three terms:

$$\delta_{\text{GWD}} = .63^{i_d} + .63^{j_d}$$

$$\delta_{\text{GWESP}} = .63^{ij_{ESP}}$$

$$\delta_{\text{GWDSP}} = .63^{ij_{DSP}}$$

δ_{GWD} indicates the log-odds of a tie would increase the most when adding an edge between *i* and *j* when *i* and *j* both have 0 degrees; the increase in log-odds of a tie would shrink as *i* and *j* have more connections (Hunter, 2007). The same general pattern holds for predicting the log-odds of a tie between two specific network members (all else held constant) using the GWESP and GWDSP change statistics. The decreased log-odds with increases in degree, ESP, and DSP in the GW terms could be considered *antipreferential attachment* (Hunter, 2007).

Generally speaking, interpretations of the GWD, GWDSP, and GWESP coefficients are consistent with interpretations of other model coefficients. A positive and significant coefficient for a geometric term indicates that the likelihood of adding a tie between any given *i* and *j* is greater than would happen by chance, all else held constant. Likewise, a negative and significant coefficient indicates that the likelihood of adding a tie between any given *i* and *j* is less than would happen by chance, and a nonsignificant coefficient would be interpreted as no significant difference from chance in the probability of adding a tie between *i* and *j,* all else held constant.

Although the coefficients appear straightforward, the change statistics can complicate deeper interpretation. As described above, the change statistics aim to capture the change in the value of the *network statistic* (see Equations 9–11 in Chapter 2) if a tie were added to the network between nodes *i* and *j* (Hunter, Goodreau, et al., 2008). Given the influence of a single tie on the shared partner distribution across the network, caution should be taken in overinterpreting the coefficients for the GWESP and GWDSP terms in particular. Following Hunter (2007, Section 5), it should be made clear that interpretations of the GWESP and GWDSP coefficients are made "assuming nothing else changes and *all other model effects* have been accounted for" (p. 227, emphasis added).

Finally, because DSP measures share partners for each dyad, connected or not, and ESP measures share partners for only connected dyads, it is important to consider the geometric terms accounting for these distributions added to the model alone and together. If GWDSP is added to the model without GWESP, the resulting coefficient may be driven by the distribution of shared partners across connected *and* unconnected pairs. If GWESP is added to the model without GWDSP, it will account only for the

distribution of shared partners in connected dyads. When GWESP is added to the model with GWDSP, GWESP accounts (or controls) for the distribution of shared partners in connected dyads, allowing the GWDSP to account for the distribution of shared partners for unconnected dyads.

The addition of GW terms to the model increases the amount of information needed to use the model for prediction. For example, consider the predictions previously made for the two LHDs: One LHD has a leader with 1 year of experience (years = 0), has 100,000 constituents (popmil = .1), is not conducting HIV screening (hivscreen = 0), but is doing nutrition programming (nutrition = 1). The other LHD has a leader with 7 years of experience (years = 2), has 2 million constituents (popmil = 2), and is conducting HIV screening (hivscreen = 1) and nutrition programming (nutrition = 1). The main effects model predicted the likelihood of a tie between these leaders to be .0023, and the differential homophily model predicted the likelihood of adding a tie between these two to be .033. To predict the probability of a tie based on the dependence model, we now need to know not only the attributes of the two network members but also the degree for each person in the dyad and the number of edgewise and dyadwise shared partnerships for the dyad. The change statistics for GWD, GWESP, and GWDSP as shown above are substituted into the model along with coefficients and change statistics for the attributes:

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}(-10.07{*}\delta_{\text{edges}}+.20{*}\delta_{\text{popmil}}+ .14{*}\delta_{3-5\text{years}}+.25{*}\delta_{6-10\text{years}}+.30{*}\delta_{>10\,years}+.19{*}\delta_{\text{HIVHom}}+ .18{*}\delta_{\text{nutritionHom}}+5.02{*}\delta_{\text{stateHom}}+.19{*}\delta_{\text{GWD}}+ .96{*}\delta_{\text{GWESP}}-.04{*}\delta_{\text{GWDSP}}$$

Values of predictors can be substituted into the model to predict the probability of adding a tie for some specific cases. Because the model now has a large number of terms, only those terms used to estimate this probability are shown in the calculations below. Case 1 revisits the likelihood of a tie between the two LHDs predicted to be .0023 or .23% by the main effects model and .033 or 3.3% by the differential homophily model. Because the dependence model now includes structural terms for degree, ESP, and DSP, we now must specify these characteristics for the two LHDs in the calculations.

Case 1: One Missouri LHD with a leader with 1 year of experience, 100,000 constituents, no HIV screening, and nutrition programming and the other Missouri LHD with a leader with 7 years of experience, 2 million

constituents, HIV screening, and nutrition programming, with degrees of 3 and 4, respectively, and 0 ESP and 3 DSP.

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}(-10.07*1+.20*2.1+.25*1+ .18*1+5.02*1+.19*(.63^{3}+.63^{4})+.96*.63^{0}-.04*.63^{3})$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}\left(-3.17\right)$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=.040$$

Case 2: Two Oregon LHD leaders, both with more than 10 years of experience, both with 25,000 constituents, both conducting HIV and nutrition programs, with degrees of 2 and 4, 1 ESP, and 2 DSP.

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}(-10.07*1+.20*.05+.30*2+.19*1+ .18*1+5.02*1+.19*(.63^{2}+.63^{4})+.96*.63^{1}-.04*.63^{2})$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}\left(-3.38\right)$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=.033$$

Case 3: Two California LHD leaders, both with more than 10 years of experience, both with 2 million constituents, both conducting HIV and nutrition programs, with degrees of 2 and 4, 1 ESP, and 2 DSP.

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}(-10.07*1+.20*4+.30*2+.19*1+ .18*1+5.02*1+.19*(.63^{2}+.63^{4})+.96*.63^{1}-.04*.63^{2})$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}(-2.59)$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=.070$$

These cases provide some additional insight into connections in the LHD network. For example, the magnitude of the coefficient for popmil does not stand out on its own; however, when it is multiplied by the combined populations of the two LHDs in a dyad, it can make a big difference in the probability of connection. Case 2 and Case 3 only differ by the population size of the two LHD jurisdictions, and the probability of a connection increases dramatically from 3.3% to 7.0%.

Predicting probabilities for CEF models is even more complex since each geometric term estimates its own value for α. In this case, the CEF model produced α values of .838 for GWD, .9451 for GWESP, and 1.822 for GWDSP. Using these values, we can calculate the change statistics for each GW term by first calculating the base value for each:

$$1-e^{-\alpha}=1-e^{-.838}=.57$$

$$1-e^{-\alpha}=1-e^{-.9451}=.61$$

$$1-e^{-\alpha}=1-e^{-1.822}=.84$$

Each value is then substituted into the corresponding equation to determine the change statistics for the three terms:

$$\delta_{\text{GWD}}=.57^{i_d}+57^{j_d}$$

$$\delta_{\text{GWESP}}=.61^{ij_{ESP}}$$

$$\delta_{\text{GWDSP}}=.84^{ij_{DSP}}$$

The full CEF model with coefficients would then be written:

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=$$
$$\text{logistic}(-9.12*\delta_{\text{edges}}+.23*\delta_{\text{popmil}}+.13*\delta_{3-5\text{years}}+.24*\delta_{6-10\text{years}}+$$
$$.28*\delta_{>10\text{years}}+.21*\delta_{\text{HIVscreenHom}}+.19*\delta_{\text{nutritionHom}}$$
$$+4.92*\delta_{\text{stateHom}}+.11*\delta_{\text{GWD}}+.97*\delta_{\text{GWESP}}-.08*\delta_{\text{GWDSP}})$$

And the three predicted probabilities for the three cases shown above would be calculated:

Case 1: One Missouri LHD with a leader with 1 year of experience, 100,000 constituents, no HIV screening, and nutrition programming and the other Missouri LHD with a leader with 7 years of experience, 2 million constituents, HIV screening, and nutrition programming, with degrees of 3 and 4, respectively, and 0 ESP and 3 DSP.

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}\begin{pmatrix}-9.12*1+.23*2.1+.24*1+.19*1+4.92*1+\\ .11*\left(.57^{3}+.57^{4}\right)+.97*.61^{0}-.08*.84^{3}\end{pmatrix}$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}\left(-2.33\right)$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=.088$$

Case 2: Two Oregon LHD leaders, both with more than 10 years of experience, both with 25,000 constituents, both conducting HIV and nutrition programs, with degrees of 2 and 4, 1 ESP, and 2 DSP.

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}(-9.12*1+.23*.05+.28*2+.21*1+.19*1+ 4.92*1+.11*(.57^{2}+.57^{4})+.97*.61^{1}-.08*.84^{2})$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}\left(-2.65\right)$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=.066$$

Case 3: Two California LHD leaders, both with more than 10 years of experience, both with 2 million constituents, both conducting HIV and nutrition programs, with degrees of 2 and 4, 1 ESP, and 2 DSP.

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}\begin{pmatrix}-9.12*1+.23*4+.28*2+.21*1+.19*1+4.92*1+\\ .11*\left(.57^{2}+.57^{4}\right)+.97*.61^{1}-.08*.84^{2}\end{pmatrix}$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=\text{logistic}\left(-1.74\right)$$

$$P\left(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}\right)=.150$$

Refining the Model Using Constraints

Until now, model building has focused primarily on the types of terms to include in the model and some recommended settings for estimation. Some research questions might benefit from constraining the space of possible networks that are considered during the estimation process for any given model. For example, participants in some network surveys are limited in the number of connections they can name. In a situation like this, it may be useful to constrain the possible networks estimated to a maximum degree for each node. Constraints are available to limit the maximum or minimum degree of a node, to preserve the exact degree of each node or the entire degree distribution, and to preserve the number of edges in the network. These constraints are described in more detail by Morris and colleagues (2008) and are defined in the R-ergm help documentation accessible from the R prompt. Since we know little about the social forces at work in any given network, it is not usually advisable to constrain the pool of realizable networks; however, if necessary, this function is available.