AMERICAN UNIVERSITY

WASHINGTON, DC

# Social Network Analysis – Statistical Applications

Anahi Rebatta Sun Han

Advisor: Alexandra Kapatou

December 2012

Technical Report No. 2013-1

https://doi.org/10.17606/bnth-qx21

# Social Network Analysis – Statistical Applications

**Anahi Rebatta Sun Han**

**Department of Mathematics and Statistics**

**American University**

**December 2012**

## Contents

## Abstract

A weighted social network data type is analyzed using two different methods. The first method used is the Exponential Random Graph Models (ERGM), which is a model used on social network analysis (SNA) that includes as parameters the structural characteristics of the network and the network's nodes attributes. However, ERGM does not take into account the weights associated with the network's edges. The second method used is the Cumulative Logistic Regression, which incorporates the weights associated to the network's edges, but it doesn't take into account the network's structural characteristics. Both methods are illustrated using a weighted one-mode data.

Keywords: *Social Network Analysis, ERGM, Logistic Regression*

**List of Figures**

**List of Tables**

# I. Social Network Analysis

## 1. Introduction

Social networks represent the relationships between individuals, groups, families, communities, regions, etc. Therefore, it can be said that a social network is a 'category of actors bound by a process of interaction among themselves'. [1] For example, a social network can represent the way how classmates interact with each other in a classroom, or how colleagues interrelate between them in a determined corporation, or how the different members of a hierarchical tribe socialized with each other.

Social Network Analysis (SNA) is the formal study of social networks, having as main purpose to study the relationships between the individuals who are part of the network, rather than studying the individuals themselves. SNA has been utilized since the mid 1930s in the field of social and behavioral sciences. However, it was not until about 1990 that the interest for social networks started to grow rapidly, and the development of new methodologies began to be explored by scientists of other fields including computer science and mathematics. [2] SNA is not focused exclusively on human networks anymore, and currently it is widely used on different fields like technological, social, biological, informational, etc.[3]

The upcoming sections will explain the structure of the data used on SNA, as well as the quantitative methods utilized to analyze and understand the networks' structural characteristics. In addition, there is an introductory section to statistical models applied to SNA. And finally, on the attempt to apply all the concepts mentioned above, a case study is presented at the end.

## 2. Social Network Data

As mentioned above, SNA studies the relationships between individuals and not the individuals themselves. For this reason, social network data presents unique characteristics and concepts that make it distinguish from data used on analyses that focus on subjects and their behaviors (i.e. data used to analyze the difference between male and female voters political preference). This section will explain some of these characteristics.

### *2.1 Elements of Network Data*

- <u>Actors and Relations</u>
  There are two crucial elements in social networks: actors and relations; and both are indispensable because, when they relate with each other, both form a social network. *Actors* can include individual actors, such as college students graduating this semester, or employees in a specific consulting firm, and collective actors, such as law firms in a particular city, or nations participating in a conference.[4]

[1] Models for Social Networks with statistical applications (3)
[2] Models and Methods in Social Network Analysis (1)
[3] Statistical Analysis of Network Data: Methods and Models (3)
[4] Social Network Analysis (e-book 2.2)

*Relations* are the types of connections between two actors. When such relationship exists, they can either flow in both directions (symmetric) or in one direction (asymmetric).[5]

Figure 1

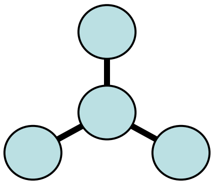


Figure 1 shows a graphic representation of a social network, where the circles represent the actors and the lines connecting them represent the relations.

### *2.2 Type of Networks*

- Egocentric and Dyadic Networks
  An *egocentric network* is formed by one actor (ego) and the other actors (alters) to whom the ego relates with. In a network with *n* actors, an egocentric analysis approach will have *n* units to analyze, and each ego can be described by the different characteristics of its ties with the other actors. A *dyadic network* consists of two actors being paired. In a dyadic network with n actors where the order of a pair is irrelevant, there will be $(n^2 – n) / 2$ units of analysis. However, if the order of a pair matters, there will be $(n^2 – n)$ ordered pairs.[6]

- One-mode and Two-mode Networks
  A *one-mode network* can be defined as a network that has one set of actors associated by one set of relationships.[7]
  A *two-mode network* data has two different set of entities (i.e. actors and events), and a connection that joins the actors with the events. When such data is presented in a matrix form, the actors are located in the rows and the events in the columns.[8]

Figure 2

| | A | B | C | D |
|---|---|---|---|---|
| A | 1 | 1 | 0 | 1 |
| B | 1 | 0 | 0 | 1 |
| C | 1 | 1 | 1 | 0 |
| D | 0 | 1 | 1 | 0 |

Figure 2.a

| | E1 | E2 | E3 |
|---|---|---|---|
| A | 1 | 0 | 1 |
| B | 1 | 1 | 1 |
| C | 0 | 0 | 1 |
| D | 1 | 1 | 0 |
| E | 1 | 0 | 0 |

Figure 2.b

[5] Models for Social Networks with statistical applications (2)
[6] Social Network Analysis (electronic book – section 2.4)
[7] Models and Methods in Social Network Analysis (8)
[8] Models and Methods in Social Network Analysis (63)

Figure 2.a represents a one-mode network; where A, B, C and D is a set of actors linked by a set of 0s and 1s that represent the absence and/or presence of a relationship. Figure 2.b shows a two-mode network, where for example, A, B, C, D and E could represent students from a middle school class, and E1, E2, and E3 could represent the afterschool clubs the students belong to.

- Binary and Weighted Networks
  A social network is said to be a *binary network* when the relationship between actors is presented by '0s' and '1s', meaning the absent and present relationship respectively.[9] See Figure 2.a above for a binary network representation.
  In a *weighted network* the relations between actors or the actors themselves have a 'weight' or strength associated to them.[10] For example, if we are analyzing a network from a financial services company, the actors labeled as "A" and "C", could represent the CEO and Jr. Analyst respectively, and both have different values. Or, if we are analyzing a particular relationship between the employees of a small call center, actor "A" may interact more with actor "B" than with actor "C", and in this case both relationships (A,B) and (A,C) have different weights.

Figure 3

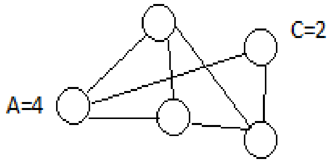
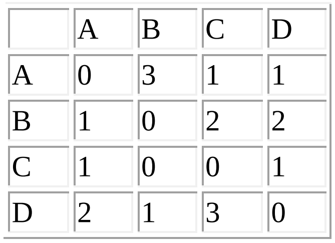


Figure 3.a



| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 3 | 1 | 1 |
| B | 1 | 0 | 2 | 2 |
| C | 1 | 0 | 0 | 1 |
| D | 2 | 1 | 3 | 0 |

Figure 3.b

Figure 3.a shows the graphic version and figure 3.b shows the matrix representation of weighted networks.

### *2.3 Informant Bias*

Sometimes due to the nature of social network data, researchers have to deal with informant bias. Informant bias can be defined as the difference between the behavior reported by the actor and its actual behavior.[11] Since social network data that intent to study the relationship between persons is collected through surveys, it may not be possible to have a precise definition of how strong the ties or relationships are perceived from person to person. For example, a survey asks to define your relation with person A as 'acquaintance' or 'friend'. However, the meaning of acquaintance and friend could differ widely from one respondent to another, and informant bias becomes an issue. Informant bias can be overcome by designing questions with ordinal type answers. For example, if we ask "how often have you met with person A for the past month?', a possible set of answers could be '0: I don't know this person', '1:never', '2: weekly' and so on.[12]

[9] Introduction to social network methods (e-book, 12)
[10] http://toreopsahl.com/tnet/weighted-networks/defining-one-mode-networks/
[11] Social Network Analysis (e-book, section 3.4)
[12] http://toreopsahl.com/tnet/weighted-networks/defining-one-mode-networks/

## 3. Social Network Data Representation

Social network data is commonly presented using graphs and matrices, depending on the type of information the researcher is trying to observe and/or present. This section will explain the two methods briefly; however both concepts have been mentioned previously on Section 2 in order to present some definitions and examples.

### *3.1 Graphs*

- Graphs and digraphs

  In order to visualize – and analyze - how actors relate with each other, social networks can be conceptualized as graphs. This graphic representation is called a *graph* if the relationship doesn't have a direction, and it is called a *digraph* if the relationship has a direction. When presenting the network graphically, actors are called *nodes* or *vertices*; and the relations between a two actors is called *tie*. Note that, a tie with a direction is called an *arc*, and tie without direction is called an *edge*. [13]

  As mentioned previously on Section 1, mathematics is one of the fields that studies also SNA, and graph theory provides definitions and techniques that are used on the graphical representation of networks. On this context, a graph G = (V, E) is a structure that has a set V of nodes, and a set E of edges. In non-directed graphs, the elements of E are unordered pairs, meaning that $\{u,v\} = \{v,u\}$; on the other hand, the elements of E in directed graphs are ordered pairs such that $\{u,v\} \neq \{v,u\}$. [14]

- Families of graphs

  The following are some of the most common families of graphs. [15]

  - *Complete graph*: A graph is said to be complete when every vertex is connected to every other vertex. And a clique is defined as a complete sub-graph A that is contained in graph G.
  - *Triangle*: A triangle is a complete graph of order three (i.e. three vertices).
  - *Tree*: A tree is a connected graph that doesn't have cycles on it.

Figure 4

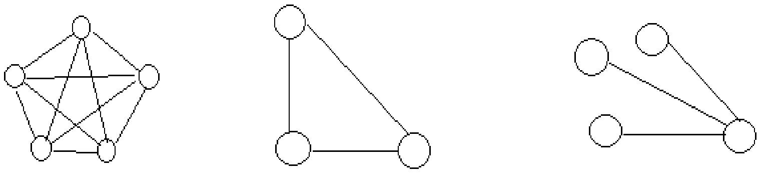


Figure 4.a
Complete graph

Figure 4.b
Triangle

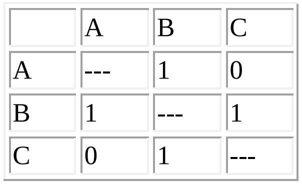
Figure 4.c
Tree

[13] Models for Social Network with Statistical Applications (7)
[14] Statistical Analysis of Network Data: Methods and Models (16)
[15] Statistical Analysis of Network Data: Methods and Models (18)

### *3.2 Matrix*

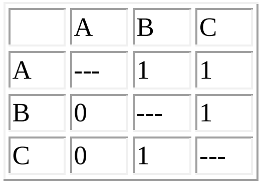
When a network has too many actors and/or relationships, a graph may not be the most appropriate form to present it, since it can be difficult to visualize it; and in such cases, matrix representation comes handy. The a*djacency matrix* is the most common matrix used in SNA, which is simply a square matrix, where the rows and columns are the actors in the network data, and the elements of the matrix represent the ties between each pair of actors. In a binary adjacency matrix, the relationship between two actors is represented by 1, and the non-existence of a relationship is represented by 0.[16] Additionally, depending on the type of relationships between actors, an adjacency matrix can be symmetric or asymmetric.

Figure 5



| | A | B | C |
|---|---|---|---|
| A | --- | 1 | 0 |
| B | 1 | --- | 1 |
| C | 0 | 1 | --- |

Figure 5.a
Symmetric Matrix



| | A | B | C |
|---|---|---|---|
| A | --- | 1 | 1 |
| B | 0 | --- | 1 |
| C | 0 | 1 | --- |

Figure 5.b
Asymmetric Matrix

Figure 5.a shows a symmetric matrix that can represent the relationship between friends. For example, A is friend of B, and B is friend of A. Figure 5.b shows a different scenario, where A considers B as a friend, but B does not feel the same way about A. Note that, in an asymmetric matrix the sender of a tie is located in the rows, and the receiver of the tie is located in the columns.

## 4. Social Network Analysis Quantitative Measures

Graphs and matrices offer straightforward ways to visualize which actors are the most active in a network (i.e. which actors have the higher number of ties). However, in addition to visual techniques, SNA has formal methods to calculate structural properties of the network. Methods that focus on finding information about the network's vertices (i.e. which actors are the most important, active, isolated, etc) are known as vertex *centrality measures*.[17] Also, there are other measures, such as *density*, that are interested on the network cohesion.[18] Some of these measures can be calculated directly from the graphs or matrices; however, some others are calculated by computational algorithms.[19] In this section we will explain the most generally used SNA quantitative measures: *degree, closeness*, *betweenness*, *eigenvector*, and *density*.

- <u>In-degree and Out-degree Centrality</u>

  Degrees are the number of direct connections an actor has, and depending on the type of information (providing/receiving) they can be in-degree or out-degree. In a network, the actors with the most direct connections are considered the most active actors in the network. *In-degree*

[16] Introduction to social network methods (e-book, 55)
[17] Statistical Analysis of Network Data: Methods and Models (80)
[18] Statistical Analysis of Network Data: Methods and Models (94)
[19] Statistical Analysis of Network Data: Methods and Models (23)

centrality measures the number of incoming ties an actor has, and *out-degree* measures the number of outgoing ties an actor has for a given relationship.[20]
When analyzing a network, an actor with a high in-degree, is referred as a 'sink' or 'receiver of information', and this actor may be seen as prestigious or more powerful, but also, it may suffer from overload information. On the other hand, an actor with a high out-degree could be seen as an actor that influences other actors in the network. [21]

- Closeness
  In a network, closeness measures how 'close' an actor is located with respect to many other actors; and for practical reason this measure is normalized in order to have values between 0 and 1, and be able to be compared with other centrality measures. [22] Closeness can be interpreted as follows: the higher an actor's closeness score is, the closer the actor is to the other actors; and this actor could be considered as the most central in the network. Note that, closeness is calculated based on the sense of the shorter geodesic distance (i.e. the length of the shortest path between actors).[23]

- Betweenness
  Betweenness measures how the relationship between a pair of actors that do not have a direct tie is controlled by the other actors that lie in between the pair's geodesic distance. Additionally, betweenness indicates which actors have control over the network relationships, therefore those actors located in between a large number of pairs have more chance to control over the information flow in the network.[24]

- Eigenvector
  Eigenvector centrality measures the status of an actor in a network, and it is calculated by using the method of factor analysis.[25] This measure can be interpreted as 'the more central the neighbors of a vertex are, the more central that vertex itself is'. [26]

- Density
  The density in an egocentric network measures the overall direct connection among actors (i.e. how connected they are among themselves). Let's define D as density, L as the total number of reported dyadic ties, N as the total number of nodes, and ${}_{N}C_{2}$ = N! / {2! * (N-2)!}. Depending on the type of data you are working on, density can be calculated as follow: [27]

  - Non-directed binary data:
    $D = L / {}_{N}C_{2}$

  - For directed binary data
    $D = L / (2 * {}_{N}C_{2})$

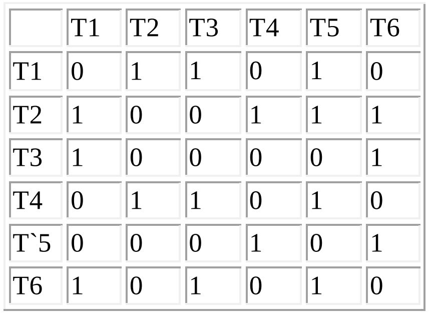
[20] The Hidden Power of Social Network (157)
[21] Introduction to Social Network Methods (97:100)
[22] Statistical Analysis of Network Data: Methods and Models (89)
[23] Social Network Analysis (e-book, section 4.4)
[24] Social Network Analysis (e-book, section 4.4)
[25] Introduction to Social Network Methods (157)
[26] Statistical Analysis of Network Data: Methods and Models (90)
[27] Social Network Analysis (4.3)

- For valued no directed data
  $D = \sum L_w / ({}_N C_2)$, where $\sum L_w$ represents the sum of all weighted ties

- For valued directed data
  $D = \sum L_w / (2 * {}_N C_2)$

***Example***: The following is a 'made up' network that will help illustrate the concepts described above. Let's say that a group of six teenagers (T1, T2,…,T6) were asked whom they considered their closest friends within the group. Figure 6, shows the adjacency matrix (6.a) and digraph (6.b) of the network.

Figure 6



| | T1 | T2 | T3 | T4 | T5 | T6 |
|---|---|---|---|---|---|---|
| T1 | 0 | 1 | 1 | 0 | 1 | 0 |
| T2 | 1 | 0 | 0 | 1 | 1 | 1 |
| T3 | 1 | 0 | 0 | 0 | 0 | 1 |
| T4 | 0 | 1 | 1 | 0 | 1 | 0 |
| T`5 | 0 | 0 | 0 | 1 | 0 | 1 |
| T6 | 1 | 0 | 1 | 0 | 1 | 0 |

Figure 6.a

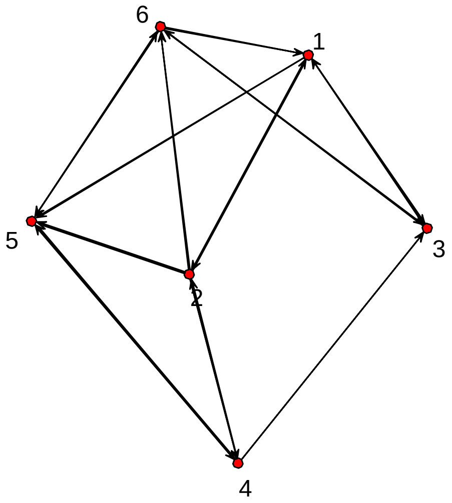


Figure 6.b

At the beginning of this section, it was said that some centrality measures are easy to visualize from a graph or matrix. Figure 6.b shows that T5 is the actor with the most incoming ties and T2 is the actor with the highest outgoing ties; meaning that T5 could be considered as the most popular teenager in the group and T2 could be considered the most outgoing one. Additionally, the density of this network is equal to 17/30 = 0.57, which it can be interpreted as the probability to be chosen as a close friend in this particular group of friends.

Table 1 shows the results of the centrality measures for this network. According to the results, and based on the closeness and eigenvector values, T2 is the most central actor, meaning that teenager 2 could be seen as the most influential teen in the network (i.e. even though T2 is not the most popular among the other teens, he/she can reach others faster than the rest of the teens). T1 has the highest betweenness value, meaning that this teen serves as 'bridge or connector' between the teens that aren't related (i.e. Figure 6.a shows that T3 and T5 are not related that all, and T3 considers T1 as a close friend, as well as, T1 considers T5 as a close friend; then, since both teens, T3 and T5, know T1, there is a chance that T3 and T5 would get connected through T1).

Table 1

| Node | In-degree | Out-degree | Betweenness | Closeness | Eigenvector |
|---|---|---|---|---|---|
| **T1** | 3 | 3 | 3.50 | 0.71 | 0.18 |
| **T2** | 2 | 4 | 2.00 | 0.83 | 0.24 |
| **T3** | 3 | 2 | 1.16 | 0.56 | 0.12 |
| **T4** | 2 | 3 | 1.83 | 0.71 | 0.18 |
| **T5** | 4 | 2 | 2.83 | 0.63 | 0.12 |
| **T6** | 3 | 3 | 2.67 | 0.71 | 0.16 |

## 5. Exponential Random Graph Models *($p^*$)*

Through previous sections we have discussed different methods that are used to measure and describe properties of a network and its nodes; and even though these techniques are useful for understanding the network and its characteristics, they do not allow the researcher to make inferences about the observed network (i.e. does the observed behavior among actors happen by chance?). Therefore, in recent years, the study of suitable statistical models, such as Exponential Random Graph Models (ERMG), for social networks has been of interest among SNA researchers. ERGM, also known as $p^*$ models, are important due to their capacity to represent social networks structural characteristics, and their ability to be simulated through stochastic processes in order to be compared with the observed network. [28] This section will describe briefly the methods and applications of ERGM.

### *5.1 Model* [29]

In conventional statistical methods, it is known that for a discrete variable case, a random vector Z is part of the exponential family when its probability mass function (pmf) can be written in the form:

$$P(Z = z) = \exp \{ \theta^T \ g(z) - \Psi(\theta)\}, \qquad (1)$$

where, $\theta \epsilon R^p$ is a $p$ x 1 vector of parameters, g(.) is a $p$-dimensional functions of z, and $\Psi(\theta)$ is a normalization term, that makes possible that sums of P(.) equals one. The same formula can be applied to a continuous random variable.

Now, let's say that we have a graph G = (V, E) and Y = [$Y_{ij}$] is the random adjacency matrix of G, where $y_{ij}$ is a binary random variable that indicates the presence or absence of a tie between the vertices i and j. Then, the ERGM have the following general form:

$$P(Y = y) = (1/k) \exp \{ \sum_{H} \theta_H \,.\, g_H(y) \}, \qquad (2)$$

where,

- H is defined as a set of possible edges on G, and each H is called a configuration,
- $g_H(y) = \prod_{yij \epsilon H} y_{ij}$, it is equal to one if the configuration H exists, and zero otherwise,

[28] An introductions to exponential random graph (p*) models for social networks (2:3)
[29] Statistical Analysis of Network Data: Methods and Models (180:181)

- $\theta_H$ is the parameter corresponding to the configuration H, a non-zero value of $\theta_H$ means that the $Y_{ij}$ are dependent for all pair of vertices in {i,j} in H, conditional upon de rest of the graph,
- k = k(θ) is a normalization constant that assures (2) is an appropriate probability distribution,

$$k(\theta) = \sum_{Y} \exp \{ \sum_{H} \theta_H \, . \, g_H(y)\} \qquad (3)$$

### *5.2 Model Construction* [30]

When fitting an ERGM for a social network, the following steps are followed:

a) *Ties are assumed to be random:* By assuming that the presence or absence of relationships between the network actors is random, we assume that we do not know under which circumstances such relationships are formed, and therefore we will expect some ‘noise’ in the model. However, this assumption is achieved because we set a stochastic framework for the fixed set V of nodes.
b) *State dependence hypothesis:* This hypothesis illustrates how social relationships are formed. For example, actors that share similar characteristics (i.e. same gender, same age group, etc) tend to relate with each other This tendency of actors to associate with others like themselves is formally know as *homophily*.
c) *The hypothesis stated on part b implies a particular form of the model:* Based on the hypotheses made, then the model represents a particular distribution of random graphs.
d) *Simplification of parameters:* If, when fitting the model, we end up with too many parameters, interpretation can become difficult. Therefore, techniques like imposing homogeneity restrains to the model can alleviate this issue.
e) *Parameters estimation and interpretation*

### *5.3 Bernoulli, Dyadic, and Marcov Random Graphs Models*

In this section, we will explain three random graph models in order to have a better understanding of ERGM. Note that, ERGM is formed by different statistical models for social networks, and they are not limited to these three models since more elaborate models have been proposed beyond Marcov random graphs.

- Bernoulli Random Graphs [31]

Assuming that in a network all edges are independent (dependence hypothesis), then the Bernoulli model is written in the form

$$P(Y=y) = (1/k) \exp \{\sum_{i,j} \theta_{ij} \, y_{ij}\} \qquad (4)$$

Note that, this model is very similar to model (2), except that in this model every single possible edge {i,j} is present and there is a parameter $\theta_{ij}$ for each of the configurations. The probability of an edge being observed can be expressed as

[30] An introductions to exponential random graph (p*) models for social networks (5:6)

[31] Statistical Analysis of Network Data: Methods and Models (182)

$$p_{i,j} = \exp(\theta_{ij}) / [\, 1 + \exp(\theta_{ij})] \qquad (5)$$

However, if we have a network with a large number of nodes, we will end up with a model with N * N parameters and this may not be suitable for model fitting and interpretation. To solve this problem, and as mentioned in the section above, a homogeneity assumption across G can be applied to the model (i.e. one of the many possible ways for imposing homogeneity is defining $\theta_{ij} \equiv \theta$), so the model will be written as

$$P(Y=y) = (1/k) \exp\{\theta\, L(y)\}, \qquad (6)$$

where, $L(y) = \sum_{i,j} y_{ij}$ is the same as the number of edge reported in the graph. Consequently, the probability of an edge being observed is

$$p = \exp(\theta) / [1 + \exp(\theta)]. \qquad (7)$$

A big disadvantage of this model is that assuming that all edges are independent is not a very realistic scenario. In addition, this model does not reflect a lot of the structural properties observed in social networks. The Bernoulli model is considered the null model because it is the simplest model of interest it can be fit with ERGM.

- Dyadic models [32]

  Dyadic models are used on directed networks, and in comparison to Bernoulli models, they are more complex, although still not very realistic. The dependence hypothesis assumes that dyads are independent - instead of the edges, as it is assumed on Bernoulli models -
  Therefore, the model now has two configurations, one for the single edges (i.e. for a specific pair {i,j} the tie $y_{ij}$ is observed, but $y_{ji}$ isn't), and other for the reciprocated edges (i.e. for a specific pair {i,j} both ties $y_{ij}$ and $y_{ji}$ are observed). Also in order to avoid an over-parameterized model, the homogeneity assumption is imposed (i.e. $\theta_{ij} \equiv \theta$). Then, the model is written as

$$P(Y=y) = (1/k) \exp\{\theta \sum y_{ij} + \rho \sum y_{ij}\, y_{ji}\} = (1/k) \exp\{\theta L(y) + \rho M(y)\}, \qquad (8)$$

  where L(y) is the number of single edges on **y**, and M(y) is the number of reciprocated edges on **y**.

- Marcov Random Graphs [33]

  Marcov models' dependency hypothesis states that two possible ties are conditionally dependent whenever they have a common vertex. This means that the ties between pairs {i,j} and {j,k} are conditionally dependent since they both share vertex j. With homogeneity assumption imposed, the model can be written as

$$P(Y=y) = (1/k) \exp\{\sum_{k=1}^{Nv-1} \theta_k\, S_k(y) + \theta_\tau T(y)\} \qquad (9)$$

[32] An introductions to exponential random graph (*p**) models for social networks (10)

[33] Statistical Analysis of Network Data: Methods and Models (182:183)

where, $S_1$(y) is the number of edges, $S_k$ (y) is the number of *k*-stars (or trees), for k ϵ [ 2, $N_{v-1}$ ], and T(y) is the number of triangles observed on the network.

### *5.4 Model Fitting and Goodness of Fit*

- Model Fitting[34]
  In conventional statistical models, the variables are assumed to be independent and identically distributed (*iid*) and model can be fitted through the method of maximum likelihood estimators (MLE), and the estimated parameters θ*hat* have confidence intervals and test statistics. On the contrary, ERGM are still being developed and parameters estimation and testing are not as straightforward as the ones described above.
  The MLE for the vector θ = ($\theta_H$) on the general form of ERGM – see equation (2) – is defined as θ*hat* = arg $\max_\theta$ *l(θ)*, where *l(θ)* is the log-likelihood, and can be express as

  $$l(\theta) = \theta^T \mathrm{g(y)} - \Psi(\theta), \qquad (10)$$

  where, g is the vector of function $g_H$ and $\Psi(\theta)$ = log k(θ).
  On the other hand, if we take the derivatives on each side, and knowing the E [g(Y)] = $\partial\Psi(\theta)/\partial\theta$, the MLE can be written as the solution to the system of equations

  $$E_{\theta hat}\,[\mathrm{h(Y)}] = \mathrm{g(y)} \qquad (11)$$

  Note that, calculation of $\Psi(\theta)$ is non-trivial, since it takes into account to summation in equation (3) over all possible choices of y, for each θ. In order to calculate approximate values for θ*hat*, stochastic processes are used. First, Marcov Chain Monte Carlo (MCMC) maximum likelihood estimation is used to estimate the log-likelihood of equation (10); and a stochastic version of the Newton-Raphson algorithm is used to approximate the solutions of the system of equations in (11).*
  **See Statistical Analysis of Network Data: Methods and Models for detailed calculation.*

- Goodness of Fit (GoF)
  Because ERGMs are still on their early stages of study, GoF methods used in conventional statistical models cannot be applied to ERGM. So far, in order to asses how good our model is, first random graphs are simulated from the fitted model, and then they are compared with the observed network. If the simulated network matches closely the characteristics of the observed network (i.e. degree centrality), then it is said that the proposed model has a good fit.[35]

## 6. Software

There is a wide variety of different software used for the analysis of social networks, such as UCINET, NetMiner, KrackPlot, Mage, Multinet, and some packages in R, just to mention a few of them. However, I will briefly mention the R packages used to present the different definitions and examples through the current project.

[34] Statistical Analysis of Network Data: Methods and Models (185:186)
[35] Statistical Analysis of Network Data: Methods and Models (187)

- R
  R is an open source programming language used for statistical computing. Due to its open source nature, many different contributors around the world had contributed to the development of the current R. In the beginning, the program was written by Robert Gentleman and Ross Ihaka (also known as *R & R*) from the Statistics Department of the University of Auckland.[36]
  The following R packages have been used on this project:
  - *igraph*: A package used for network analysis and graph visualization.[37]
  - *statnet:* This package, in addition to network analysis, allows the user to perform network modeling based on ERGM.[38]
  - *network:* This package is used to create and manipulate network objects.[39]

## 7. Example: Consulting firm

This section will illustrate the applications of the different SNA concepts and ERGMs described on the previous sections. The data set used in this analysis was obtained from Rob Cross website (http://www.robcross.org).

### *7.1 Data Set*

- Data Description
  The observed network 'Advice Network' contains an intra-organizational weighted one- mode network obtained from a consulting company. A total of 46 employees were asked:
  - How often have you turned to this person for information/advice on work-related topics in the last three months?

| | | |
|---|---|---|
| 0 = I do not know this person, | 1 = Never, | 2 = Seldom, |
| 3 = Sometimes, | 4 = Often, | 5 = Very often |

In addition, the data contains attributes about the 46 employees:

- Gender

| | | |
|---|---|---|
| 1 = Male | 2 = Female | |

- Organizational level

| | | |
|---|---|---|
| 1 = Research Assistant | 2 = Junior Consultant | 3 = Senior Consultant |
| 4 = Managing Consultant | 5 =Partner | |

- Office region

| | | |
|---|---|---|
| 1 = Europe | 2 = United States | |

- Office location

| | | |
|---|---|---|
| 1 = Boston | 2 = London | 3 = Paris |
| 4 = Rome | 5 = Madrid | 6 = Oslo |
| 7 = Copenhagen | | |

[36] http://www.r-project.org
[37] http://cran.r-project.org/web/packages/igraph/igraph.pdf
[38] http://statnet.org
[39] http://cran.r-project.org/web/packages/network/network.pdf

- Data Changes/Corrections

  A few changes have been made to the observed network in order to make the analysis feasible. First, it was noticed that two employees (6 and 26) reported that they turned to themselves for advice, which in this case was not appropriate since they were asked to respond about other employees and not themselves. Therefore, these two ties were removed from the network. Additionally, for analysis purposes the observed weighted network was converted into a binary network; see below for details:

    - Instead of recording how often a person goes for advices, this network records if a person goes or doesn't go for advice. Consequently, those employees whose ties values are equal to 0 and 1, will be now have a new tie equal to 0; and those employees whose ties values are equal to 2,3,4,5, will have a new tie equal to 1.

Table 2

| Old values | New value | Description |
|---|---|---|
| 0, 1 | 0 | No, I did not go for advice |
| 2, 3, 4, 5 | 1 | Yes, I went for advice |

### *7.2 Analysis*

**Note that, instead of using the original advice network, its binary version will be use on the analysis section because some of the functions used on the R packages only support binary networks type of data. Additionally, for the complete R script and output see Appendix A.*

The binary 'Advice network' is of interest when studying the flow of seeking advice on topics related to work in this particular consulting firm.

- Exploratory/Preliminary Analysis

  By converting the binary data into an object network, the following information is obtained:

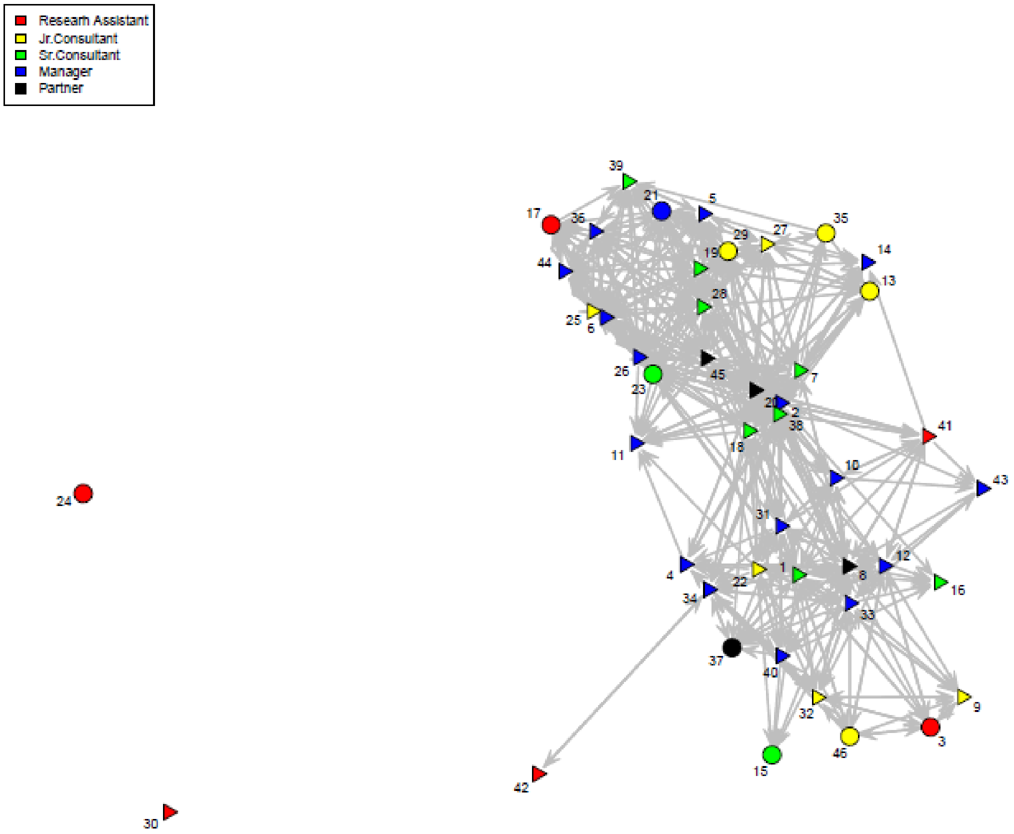
Total of nodes/vertices = 46  
Total edges / dyads = 521  
Directed network = True  
Total number of mutual / reciprocate dyads = 201  
Vertex attributes = gender, organizational level, location, region, and 'id' employees.

The summary above tells us that the network has a total of 46 vertices, and 521 edges out of 2070 possible edges (i.e. 2 * ${}_{46}C_2$ = 2070). In addition, out of the 521 edges, 201 edges are mutual (i.e. (i,j) = (j,i) ), this means that if employee A seeks advice on B, then B seeks advice on A.

Below, Figure 6 shows the graphic representation of the binary advice network. The vertices represent the 46 employees, the vertex shape indicates the employees' gender (triangle for male employees and circles for female employees) and the different colors indicate the employees' positions within the company. Visually, we can identify two isolated vertices, meaning these two Research Assistants do not ask for advice to their co-workers, neither are asked for advice by other employees. Also, note that most of the Jr. Assistants and all the Research Assistants are located in the borders of the graph, meaning that information does not flow as often among them.

In addition, it seems that vertices 45, 20, 2, 18, and 38, who happen to hold higher level positions in the company, are the most central on the network (due to their position on the graph). When looking at gender, it seems that female employees are not as active as their male counterparts, and this could be due to a large difference between the number of female and male employees. Overall, the graph gives an idea on how the information tends to flow in this company based on some of the employees attributes.

Figure 6



In addition to the advantages of visualizing possible behavioral patterns on the plot, we can also create matrixes that reflect the relationship between actors according to the particular attributes. See tables 3, 4 and 5 for more details

Table 3 – Organization Level

| From \ To | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| 1 | 0 | 5 | 7 | 12 | 3 | 27 |
| 2 | 3 | 18 | 24 | 32 | 9 | 86 |
| 3 | 5 | 26 | 30 | 54 | 18 | 133 |
| 4 | 11 | 31 | 51 | 71 | 33 | 197 |
| 5 | 4 | 12 | 19 | 37 | 6 | 78 |
| Total | 23 | 92 | 131 | 206 | 69 | 521 |

1 = Research  Assistant
2 = Junior Consultant
3 = Senior Consultant
4 = Managing Consultant
5 = Partner

Table 3 suggests that employees with lower organization level positions tend to look for advice on those employees that hold higher level positions. This is something somehow expected on real world scenarios, for example a research assistant will ask for advice to those employees with more experience. Also, employees that hold mid and higher level managerial positions tend to interact with employees at their same or higher level position.

Table 4 - Location

| From \ To | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|---|---|---|---|---|---|---|---|---|
| 1 | 314 | 3 | 18 | 8 | 5 | 5 | 0 | 353 |
| 2 | 4 | 0 | 5 | 0 | 0 | 0 | 0 | 9 |
| 3 | 13 | 3 | 37 | 8 | 4 | 10 | 11 | 86 |
| 4 | 8 | 1 | 9 | 1 | 1 | 3 | 0 | 23 |
| 5 | 2 | 0 | 3 | 1 | 2 | 0 | 0 | 8 |
| 6 | 5 | 0 | 10 | 4 | 2 | 4 | 1 | 26 |
| 7 | 0 | 0 | 10 | 0 | 1 | 0 | 5 | 16 |
| Total | 346 | 7 | 92 | 22 | 15 | 22 | 17 | 521 |

1 = Boston
2 = London
3 = Paris
4 = Rome
5 = Madrid
6 = Oslo
7 = Copenhagen

Table 4 shows that employees at the Boston office interrelate the most with their office mates; however, employees across Europe tend to ask for advice to the employees at the Paris office. One can assume that this pattern may happen because the headquarters in Europe are located in Paris. (i.e. one can assume that employees located in the same region, US and Europe, tend to interact more with same region employees)

Table 5 - Gender

| From \ To | 1 | 2 | Total |
|---|---|---|---|
| 1 | 361 (83%) | 75 (17%) | 436 |
| 2 | 70 (82%) | 15 (18%) | 85 |
| Total | 431 | 90 | 521 |

1 = Male
2 = Female

Table 5 shows that male employees are asked for advice more often than their female counterparts. This may happen due the large difference between the number of male employees and female employees.

- Centrality Measures

After calculating all centrality measures (see Appendix A-II for complete tables), it is concluded that employees 20, 2, 45 and 8 have the higher number of in-and-out degrees in the network; this means that they seem to be influential employees in the company since information flows in both directions through them. In addition, and based on their closeness and eigenvector values, employees 2, 20, and 45 are considered the most central actors. Note that these formal results confirm what it was observed on the Figure 6.

Besides that, note that in this particular network, information is received and passed through those employees that hold high rank positions, which it can be assumed to be expected in a work office environment. Additionally, the density of the network is equal to (521 / 2079) = 0.25.

Since all values of centrality measure have been calculated, we can compute correlations between them and determine how closely these measures are related to each other. Table 6 below shows that in and out degrees are highly correlated, possibly meaning that looking for advice within employees on this company could be reciprocal (i.e. if employee A goes for advice to employee B, employee B may tend to go for advice to employee A). Betweenness is correlated with out-degree, this could be interpreted that in this particular company those employees that ask for advice tend to control the flow of information in the company (i.e. they serve as bridges between those employees that are not associated directly). In and out degrees are highly correlated with eigenvector, meaning that those employees that look for advice the most to other employees, and those employees that are largely sought for advice by others, are connected to other highly related employees, and possibly being the most central employees of the company.

Table 6

| | In-degree | Out-degree | Betweenness | Closeness | Eigenvector |
|---|---|---|---|---|---|
| In-degree | 1 | 0.85 | 0.69 | 0.65 | 0.88 |
| Out-degree | 0.85 | 1 | 0.72 | 0.58 | 0.86 |
| Betweenness | 0.69 | 0.72 | 1 | 0.39 | 0.48 |
| Closeness | 0.65 | 0.58 | 0.39 | 1 | 0.52 |
| Eigenvector | 0.88 | 0.86 | 0.48 | 0.52 | 1 |

- Model Fitting and Goodness of Fit

As mentioned early on this section, we are interested on assessing the effects of seeking for advice among the consulting firm employees. Therefore, a few models were fit (see Appendix A - II for complete summaries of all fitted models); however, only two models will be presented on this section.

First, we fit a Bernoulli model (the simplest model). This model only takes into account the total number of edges from the observed network. By using this model we are trying to estimate the probability of looking for advice in the observed network, and according to equation (6), we can expressed the probability as,

$$P(Y=y) = (1/k) \exp \{\theta\ L(y)\} = (1/k) \exp \{\theta_{edges} * 521\}$$

The log-odd of one tie is equal to

$$\text{logit } P(Y_{ij}=1) = \theta_{edges}\,\Delta\,(g(y))_{i,j} = \theta_{edges} * 1$$

where is an observed tie in $Y_{ij}$, and $\Delta\ (g(y))_{i,j}$ is the change in g(y). In our case of a binary network, since the addition of any tie to the network changes the number of ties by 1, then the change $\Delta\ (g(y))_{i,j}$ is equal to 1 for all ties.

And as mentioned on equation (7), the probability of an edge (i.e. looking for advice) being observed is calculated as,

$$\text{prob} = \exp(\theta_{edges}) / [1 + \exp(\theta_{edges})]$$

*Summary of model fit*

```
Maximum Likelihood Results:
           Estimate   Std. Error   MCMC s.e.     p-value
edges      -1.08961      0.05065          NA     <1e-04 ***

AIC: 2337.7    BIC: 2343.4
```

The summary above shows the estimate for the log-odds of a tie, in addition to its standard error and p-value of significance. MCMC s.e. stands for MCMC standard error, which in this case is not applicable. It also includes two measures of model fit (AIC and BIC).
By using the edges estimated parameter, the log-odds of any tie (i.e. asking for advice) occurring in our network is equal to { (-1.09) * 1 } = (-1.09), and its corresponding probability is equal to exp(-1.09) / [1 + exp(-1.09)] ≈ 0.25 - which it happens to be equal to the network's density -. However, it's already known that this model is unrealistic and it does not reflect the network characteristics.

Since, the first model doesn't take into account any of the network's structural characteristics; a better model (i.e. a model that includes parameters that we are interested on) is fitted.
On the preliminary analysis it was observed that employees tend to interact the most with those employees that are located in the same region, and it was also observed that employees tend to ask for advice to employees of their same gender. Also, it was observed that organization level has an effect on the way how employees seek for advice. Besides that, we know that about 38% of the total observed edges are mutual (201 out of 521).

Then, we fit a model that has three network statistics and six attributes statistics. The network statistics are the number of edges, the number of mutual edges, and mixed-two star (i.e. if *i* goes to *j*, and *j* goes to *h*, then there is a potential for i and h to get connected through j). The attributes statistics include the main effects for 'organizational level' (i.e. the level position of an employee tends to affect the way how he/she looks for advice), and the second-order effects for 'region' and 'gender' (i.e. tendency of employees of forming ties with others from the same regions, or of the same gender)

The overall model is written as,

$$P(Y=y|X=x) = (1 / k(\theta,\beta)) \exp \{\theta_{edges} S_1(y) + \theta_{mutual} S_2(y) + \theta_{m2star} S_3(y) + \beta^T g(y,x)\},$$

Where g is the vector for the attributes statistics, and β is the corresponding vector of parameters. Note that, $g(y,x) = \sum_{1 \leq i \leq j \leq N_v} y_{ij} h(x_i, y_j)$, where h is a function of $x_i$ and $x_j$, and $x_i$ is the vector of observed attributes for the *i-th* vertex. In our case, h has two categories, one for main effects and other for second order effects.

Main effects: $h(x_i , y_j) = \text{location}_i + \text{location}_j$
Second order effects: $h(x_i , y_j) = I\{\text{region}_i = \text{region}_j\}$ , $h(x_i , y_j) = I\{\text{gender}_i = \text{gender}_j\}$

<u>*Summary of model fit*</u>

```
MCMC sample of size 10000
Monte Carlo MLE Results:
                     Estimate      Std. Error    MCMC s.e.    p-value
    edges            -6.302960       0.201777      3.325      <1e-04 ***
    mutual            2.822515       0.239312      1.147      <1e-04 ***
    m2star            0.058162       0.006892      0.058      <1e-04 ***
    nodefactor.level.2  0.653985     0.095175      0.743      <1e-04 ***
nodefactor.level.3    0.706072      0.088235    0.488       <1e-04 ***
nodefactor.level.4    0.710419      0.081561    0.727       <1e-04 ***
nodefactor.level.5    1.168843      0.033310    0.000       <1e-04 ***
nodematch.region      1.781939      0.019971    0.004       <1e-04 ***
nodematch.gender      0.346361      0.017694    0.001       <1e-04 ***

AIC: 1487.5    BIC: 1538.2
```

The summary provides the estimates and two sets of standard errors. In addition, the summary shows that all estimates are significant. Note that before interpreting the estimates, we check that the proposed model is a non-degenerate model (see Figure 10 – Appendix A-II). In Figure 1, the plots on the left show the chain as a time series and the plots on the right show the chain in a histogram (for each statistic). In a converge model, the statistics will vary stochastically around the mean, and if the density histograms have a bell shaped curve (or approximate) then there is enough evidence that we have generate a non-degenerate model.

The estimated coefficients of the networks statistics can be interpreted in terms of the log-odds. For example, the odds of observing a tie between two employees that have different attributes and that doesn't belong to a reciprocate pair neither have a node in common is exp(-6.30) = 0.002. However, if the pair described above has a vertex in common, then the odds increases to exp(-6.30 + 0.06) = exp(-6.24) = 0.002. And, the odds of observing a tie that is a reciprocate is exp (-6.30+2.82) = exp(-3.48) = 0.031.

The estimated coefficients of the main effects (organizational level) attributes can also be interpreted as the log-odds ratio (in the sense of *all else being equal).* Note that 'Research Assistant (level 1)' is being considered as the *baseline* and interpretations will be comparing all the other levels vs. level 1. For example, we can say that being a 'senior consultant (level 3)' rather than a 'research assistant (level1) will increase the odds of seeking advice by nearly two times since exp(0.706) ≈ 2.03. Finally, the estimated coefficients of the second order effects are also interpreted as the log-odds ration. For example, being from the same region increases the odds of looking for advice between two employees by a factor of exp(1.782) ≈ 5.94. Similarly, being of the same gender increases the odds of looking for advice by a factor of exp(0.41) ≈ 1.42.

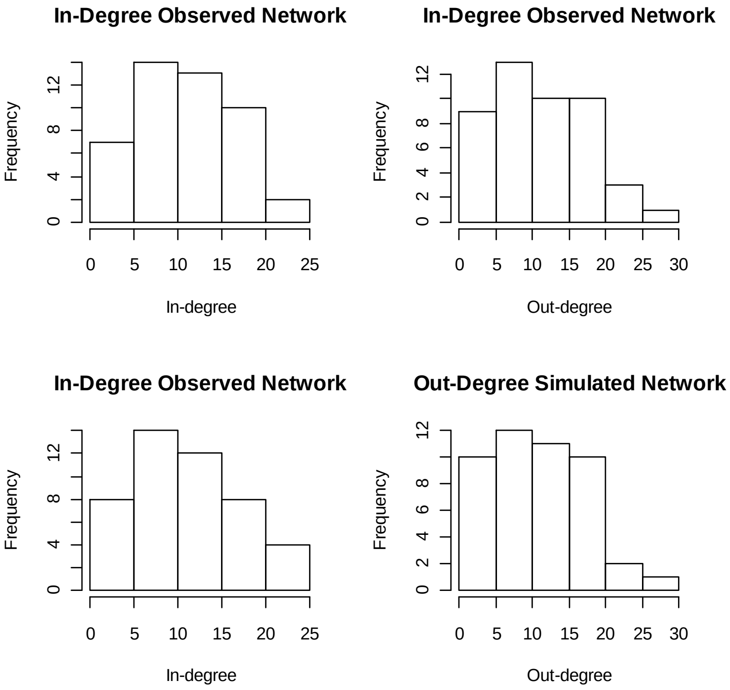
Now, let's check the GoF of that last model. First, random graphs are simulated based on the model and simulated network is obtained. Then, this simulated network was compared with the observed network and we see that some of the structural characteristics we were interested on are very similar on both of them (see below).

Table 7 – Observed vs. Simulated Network

| | Edges | Density | Reciprocate dyads | Two paths | Triangles |
|---|---|---|---|---|---|
| Observed Network | 521 | 0.25 | 201 | 7013 | 5364 |
| Simulated Network | 526 | 0.25 | 197 | 7163 | 4706 |

As mentioned on Section 6, another way of checking GoF is by comparing the degree centrality of the observed and simulated network. Figure 8 shows the histograms of the in-and-out degree distributions of both networks, and histograms of the simulated network are very similar to the observed ones.

Figure 8



The statnet package also includes a built-in GoF function. This GoF function simulates 100 networks based on the fitted model, and plot their averages against the observed network.

Figure 9

Goodness-of-fit diagnostics

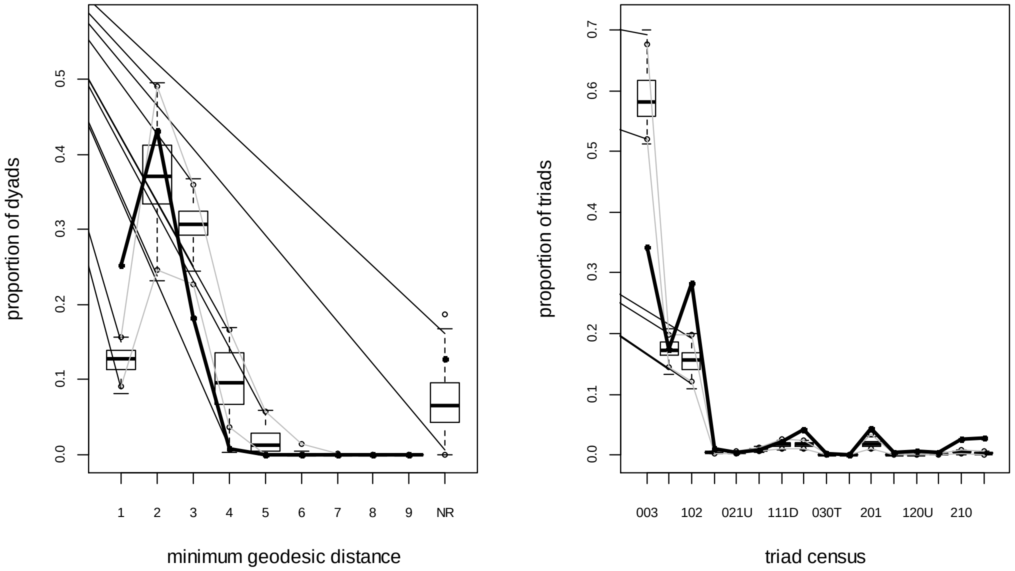


Figure 9 shows the plots from the built-in GoF function on statnet. The black lines represent the observed network, and the dotted lines represent 100 realizations from the model with the distance and triad census (triangles) parameters. Based on the plots of these two structural characteristics, the fit of the model is good in overall.

## II. Ordinal Categorical Data Analysis

### 1. Introduction

As it has been described on Section I, SNA offers useful tools when trying to describe and understand the relationships formed in a network by its actors. However, SNA does lack effectiveness when trying to predict and make inference based on the observed data. More specific, this lack of effectiveness becomes an issue when the network has weighted/valued ties since the proposed models are only applicable on binary networks, and by converting a weighted network into a binary network, the researcher is at risk of possibly loosing important information. This lack of efficiency was the motivation to use some conventional statistical methods to analyze social network data. Although, instead of analyzing the weighted ties formed by the actors, we will analyze the actors themselves.

Social networks described human relationships, and on weighted networks, the values assigned to the actors could describe the hierarchy of an actor in the network, and the values assigned to the ties could describe the frequency of an event. Therefore, since there is an obvious order in the response variable (i.e. how often do you ask for advice? Don't know this person, never, ...., very often) that can be taken into account the model specifications, it was thought that ordinal logistic regression models will be appropriate for this type of data.

### 2. Ordinal Logistic Regression

As mentioned on Section I–2.2, in a weighted network the relationships between actors can have a specific value (and represent an ordered value). If we considered the valued relationship as the response variable, a way to recognize the order of the response is by using cumulative logits,

*Cumulative Logits* [40]

If the response variable has c outcome categories with probabilities $\pi_1, \pi_2, \ldots, \pi_c$ , then the cumulative logits are defined as

$$\begin{aligned}\text{logit}\,[\,P(Y\leq j)\,] &= \log\,\{\,P(Y\leq j)\,/\,[1-p(Y\leq j)\,]\,\} && (12)\\ &= \log\,\{(\pi_1+\ldots+\pi_j)\,/\,(\pi_{j+1}+\ldots+\pi_c)\}, && j=1,\ldots,c\text{-}1.\end{aligned}$$

*Cumulative Logit Models*[41]

For subject i, yi denotes the outcome category for the response variable, and xi denote a column vector of the values of the explanatory variables. And the model uses all c-1 cumulative logits. Then the cumulative logit models can be written as,

[40] Analysis of Ordinal Categorical Data (44)

[41] Analysis of Ordinal Categorical Data (46:47)

$$\text{logit} [ P(Y\leq j) ] = \alpha_j + \beta' x_i = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots, \qquad (13)$$

for j=1, ..., c-1, and column vector β of parameters that describes the effects of the explanatory variables. The expression for the cumulative probability is,

$$P (Y\leq j) = \exp (\alpha_j + \beta' x_i) / \{1 + \exp (\alpha_j + \beta' x_i)\}, \qquad (14)$$

And the probability of a specific outcome is equal to

$$\begin{aligned} P (Y=j) &= \exp (\alpha_j + \beta' x_i) / \{1 + \exp (\alpha_j + \beta' x_i)\} \\ &= \exp (\alpha_{j-1} + \beta' x_i) / \{1 + \exp (\alpha_{j-1} + \beta' x_i)\} \qquad (15) \end{aligned}$$

*Goodness of Fit* [42]

When checking the GoF, we are testing

Ho: The model holds vs. Ha: The model doesn’t hold

The Pearson statistics is

$$X^2 = \sum_i \sum_j [(n_{ij} – u_{ij})^2 / u_{ij}], \qquad (16)$$

The Likelihood Ratio is

$$G^2 = 2 \sum_i \sum_j n_{ij} \log(n_{ij} / u_{ij}). \qquad (17)$$

Note that in both equations, $n_{ij}$ is the observed value, and $u_{ij}$ is the predicted value.

When both statistics, $X^2$ and $G^2$, are significant (i.e. the model doesn’t hold), the *Hosmer-Lemeshow** statistic is used instead of the Pearson and Likelihood ratio test.
**See ‘An Introduction to Generalized Linear models’ (135:137) for additional information about Hosmer-Lemeshow statistic..*

## 3. Case Study: Consulting Firm

The same network data used on the previous section will be used on this example. However, instead of being modeled as a network, the data will be modeled using the cumulative logit regression model. The ‘Advice Network’ data has 46 employees and a total of 877 responses to the question ‘How often have you turned to this person for information/advice on work-related topics in the last three months?’, and the responses were recorded as follow,

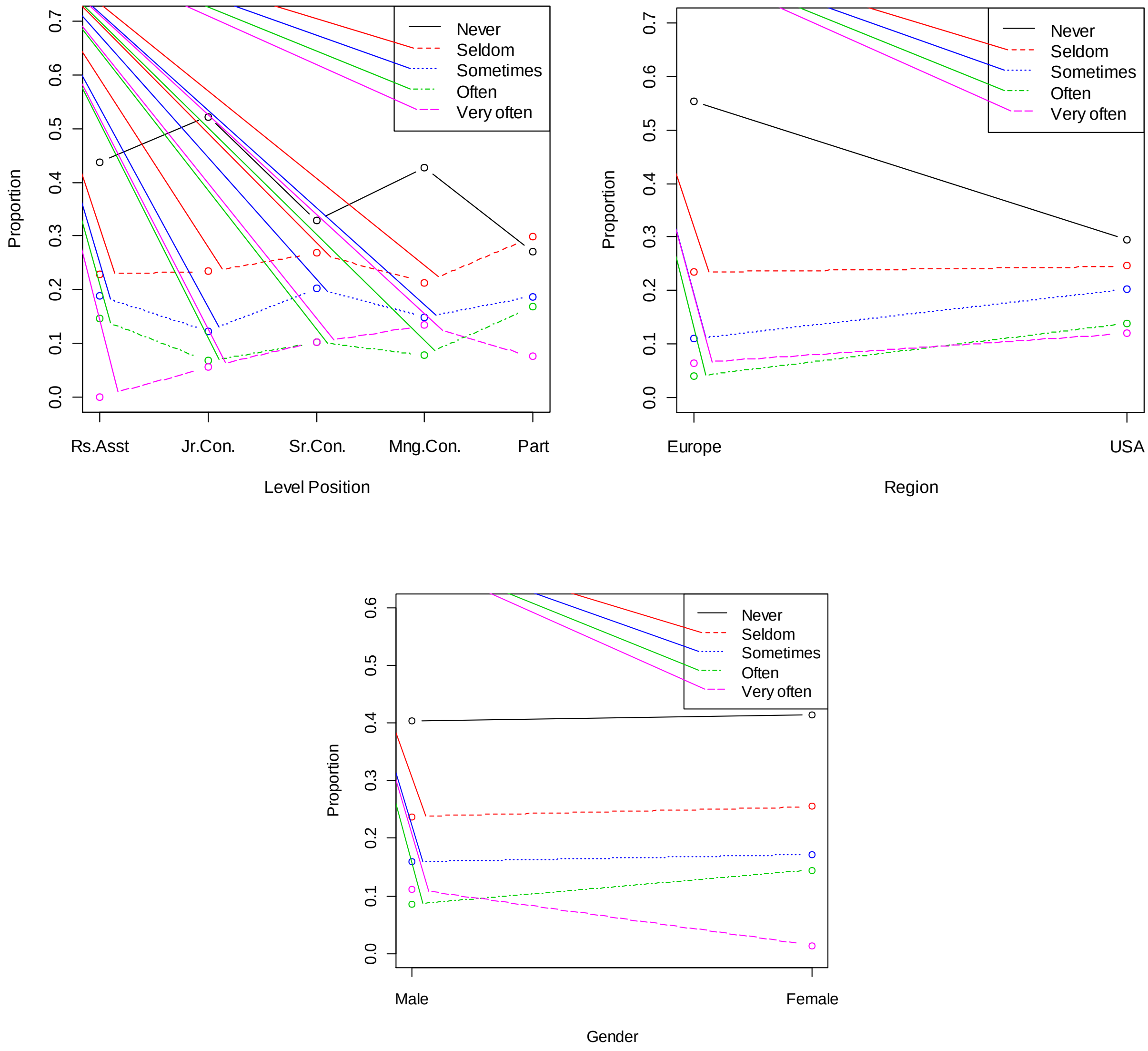
1 = Never, 2 = Seldom, 3 = Sometimes, 4 = Often, 5 = Very often

[42] Analysis of Ordinal Categorical Data (67)

- Exploratory Analysis
  *See Appendix B –III,IV for complete SAS script and output*

The plots below on Figure 10 show the proportions of asking for advice among the consulting firm employees look for advice according to their organizational level position, region office, and gender. The frequency of asking for advice differs depending on the position the employees held at the company; although it seems to be a general trend to ask less frequent for advice among the employees with higher organizational levels. Also, employees that work in the US seem to ask for advice more often then those who work in Europe. In addition, we can see that in overall women employees ask for advice less frequent than male employees.

Figure 10



- Model Fitting and Goodness of Fit: Explaining frequency of advice seeking by each of the explanatory variables.

*Organizational Level / Advice*

Let's apply the cumulative logit model for the following contingency table that treats the employees' answers as the response variable, and the organization level as the explanatory variables. By fitting the cumulative logit model we are trying to see if employees ask for advice less / more frequent according to their level position.

Table 8

'How often have you turned to this person for advice on work-related topics in the last three months?

| | 1= Never | 2 = Seldom | 3=Sometimes | 4=Often | 5=Very often |
|---|---|---|---|---|---|
| 1 = Research Ast | 21 | 11 | 9 | 7 | 0 |
| 2 = Jr. Consultant | 94 | 42 | 22 | 12 | 10 |
| 3 = Sr. Consultant | 65 | 53 | 40 | 20 | 20 |
| 4 = Mng. Consultant | 147 | 73 | 51 | 27 | 46 |
| 5 = Partner | 29 | 32 | 20 | 18 | 8 |

*Summary Organization Level / Advice*

| Parameter | Estimate | St. Error | Wald Chi-Sqr | Pr>ChiSq |
|---|---|---|---|---|
| Intercept 1 | 0.2209 | 0.1991 | 1.2308 | 0.2672 |
| Intercept 2 | 1.2139 | 0.2033 | 35.6483 | <0.0001 |
| Intercept 3 | 2.0558 | 0.2113 | 94.6256 | <0.0001 |
| Intercept 4 | 2.8654 | 0.2259 | 160.9239 | <0.0001 |
| Level | -0.1824 | 0.0565 | 10.4281 | 0.0012 |

The summary shows the results of the model for the contingency table. Note that there are 4 parameters intercepts, since the response variable has a total of 5 ordinal variables, and the explanatory variable level is being considered as a quantitative variable since it represents the order of the hierarchy level organization of the consulting firm. The estimated level parameter shows that in general the cumulative probability decreases as organizational level increases.

For example, for Jr. Consultant level we can estimate its correspondent cumulative probability:

Sometimes: $P(Y\leq 3) = \exp\{(2.0558)+(2)(-0.1824)\} / \{1+ \exp[(2.0558)+(2)(-0.1824)]\} = 0.84$

Often: $P(Y\leq 4) = \exp\{(2.8654)+(2)(-0.1824)\} / \{1+ \exp[(2.8654)+(2)(-0.1824)]\} = 0.92$

If we want to calculate the exact probabilities for each response for Jr. Consultant,

$P(Y = 4) = P(Y\leq 4) - P(Y\leq 3) = 0.08$

$P(Y = 5 ) = 1 - P(Y\leq 4) = 1 - 0.92 = 0.08$

The same calculations can be applied to each different organizational level.

Region/Advice

Then, let's apply the cumulative logit model to the following contingency table that treats the employees' answers as the response variable, and region as the explanatory variables. By fitting the cumulative logit model we are trying to see if employees ask for advice less / more frequent according to their work region. Note that the proportions of employees located in Europe of asking for advice is smaller across all the responses (with the exception of never)

Table 9

‘How often have you turned to this person for advice on work-related topics in the last three months?

| | 1= Never | 2 = Seldom | 3=Sometimes | 4=Often | 5=Very often |
|---|---|---|---|---|---|
| 1 = Europe | 208(55%) | 88(23%) | 41(11%) | 15(4%) | 24(6%) |
| 2 = USA | 148(30%) | 123(25%) | 101(20%) | 69(14%) | 60(12%) |

*Summary Region / Advice*

| Parameter | Estimate | St. Error | Wald Chi-Sqr | Pr>ChiSq |
|---|---|---|---|---|
| Intercept 1 | 1.3031 | 0.2134 | 37.3059 | <0.0001 |
| Intercept 2 | 2.3610 | 0.2243 | 110.8455 | <0.0001 |
| Intercept 3 | 3.2445 | 0.2354 | 189.9404 | <0.0001 |
| Intercept 4 | 4.0716 | 0.2504 | 264.4468 | <0.0001 |
| Region | -1.0876 | 0.1296 | 70.3986 | <0.0001 |

The summary shows the results of the model for the contingency table. Note that the parameter for region is negative, which shows the tendency of employees from US to ask for advice less frequent that their counterparts in Europe, and region1 (Europe) is being considered as the baseline.

For example, the cumulative probability USA employees of asking for advice is,
Sometimes: $P(Y\leq 2) = \exp\{(2.3610)+(2)(-1.0876)\} / \{1+ \exp[(2.3610)+(2)(-1.0876)]\} = 0.55$
Often: $P(Y\leq 3) = \exp\{(3.2445)+ (2)(-1.0876)\} / \{1+ \exp[(3.2445)+ (2)(-1.0876)]\} = 0.74$

If we want to calculate the exact probabilities for each response for USA,

$$P(Y = 3) = P(Y\leq 3) - P(Y\leq 2) = 0.19$$

Gender / Advice

Finally, the cumulative logit model is applied to the following contingency table that treats the employees’ answers as the response variable, and the employees’ gender as the explanatory variables. By fitting the cumulative logit model we are trying to see if employees ask for advice less / more frequent according to their gender. Note that the proportion of women asking for advice is slightly larger across all the responses (with the exception for very often)

Table 10

‘How often have you turned to this person for advice on work-related topics in the last three months?

| | 1= Never | 2 = Seldom | 3=Sometimes | 4=Often | 5=Very often |
|---|---|---|---|---|---|
| 1 = Male | 296 (40%) | 174(24%) | 117(16%) | 63(9%) | 82(11%) |
| 2 = Female | 60 (41%) | 37 (26%) | 25 (17%) | 21(14%) | 2(1%) |

*Summary Gender / Advice*

| Parameter | Estimate | St. Error | Wald Chi-Sqr | Pr>ChiSq |
|---|---|---|---|---|
| Intercept 1 | -0.5508 | 0.2063 | 7.1965 | 0.0073 |
| Intercept 2 | 0.4337 | 0.2050 | 4.4764 | 0.0344 |

| | | | | |
|---|---|---|---|---|
| Intercept 3 | 1.2697 | 0.2101 | 36.5259 | <0.0001 |
| Intercept 4 | 2.0755 | 0.2231 | 86.5269 | <0.0001 |
| Gender | 0.1448 | 0.1658 | 0.7630 | 0.3824 |

The summary shows the results of the model for the contingency table. Note that as on the previous model, there are 4 parameters intercepts, since we are using the same response variable, and in this case the explanatory variable gender is being considered as a qualitative variable, meaning that gender 1 (male) is being used as the baseline.

For example, the cumulative probability of women asking for advice is,
Sometimes: $P(Y \le 2) = \exp\{(0.4337)+(2*0.1448)\} / \{1+ \exp[(0.4337)+(2*0.1448)]\} = 0.67$
Often: $P(Y \le 3) = \exp\{(1.2697)+(2*0.1448)\} / \{1+ \exp[(1.2697)+(2*0.1448)]\} = 0.82$

If we want to calculate the exact probabilities for each response for women,
$P(Y = 3) = P(Y \le 3) - P(Y \le 2) = 0.15$

Now for male employees, the cumulative probability for asking for advice is
Sometimes: $P(Y \le 3) = \exp\{(1.2697)+(0.1448)\} / \{1+ \exp[(1.2697)+(0.1448)]\} = 0.80$
Often: $P(Y \le 4) = \exp\{(2.0755)+(0.1448)\} / \{1+ \exp[(2.0755)+(0.1448)]\} = 0.90$

If we want to calculate the exact probabilities for each response for women,
$P(Y = 5) = 1 - P(Y \le 4) = 1 – 0.90 = 0.10$

Goodness of Fit
Out of the three models fitted above, Gender/Advice is the only model that shows lack of fit (See Appendix B-IV).

- Model Fitting and Goodness of Fit:
  Finally, we would like to build a model using advice as the response variable, and region and level as the explanatory variables.

$$\text{logit}\,[\, P(Y \le j)\,] = \beta_{0j} + \beta_{1j}\, x_1 + \beta_{2j}\, x_2, \qquad j=1,2,...,5.$$

where,
j = 1:Never, 2:Seldom, 3: Sometimes, 4: Often, 5: Very often
$x_1$ = Region = 1: Europe and 2: USA
$x_2$ = Level = 1: Res. Assist 2: Jr. Con. 3: Sr. Con 4: Mng. Con 5: Partner

*Summary Model Advice ~ Region + Level*

| Parameter | Estimate | St. Error | Wald Chi-Sqr | Pr>ChiSq |
|---|---|---|---|---|
| Intercept 1 | 1.9424 | 0.2881 | 45.4539 | <0.0001 |
| Intercept 2 | 3.0097 | 0.2989 | 101.3682 | <0.0001 |
| Intercept 3 | 3.9007 | 0.3093 | 159.0864 | <0.0001 |
| Intercept 4 | 4.7342 | 0.3219 | 216.2586 | <0.0001 |
| Region | -1.0939 | 0.1300 | 70.8181 | <0.0001 |
| Level | -0.1908 | 0.0576 | 10.9595 | 0.0009 |

The summary shows the results of the cumulative logistic regression model. There are 4 intercept parameters because, as on the previous models, the response variable has 5 categories. Note that the negative parameter of organization level suggests that the cumulative probability decreases as the organizational level increases, and the negative parameter of region also suggests that the cumulative probability decreases for US when comparing to Europe.

For example, for a fixed organizational level the estimated odds ratio of seeking advice in the US vs. Europe below any level j is approximately exp(-1.0939) = 0.33. Also, for a fixed region the estimated odds of seeking advice for 'Research Assistants' below any level j is approximately exp(-0.1908) = 0.83. As well, following the same logic, the estimated odds can be calculated for the other organizational levels.

Additionally, the score test for the proportional odds assumptions shows that the model holds, meaning that it fits the data (See Appendix B-IV)

## III. Conclusions

Social Network Analysis provides appropriate methods to model and analyze social network data, and ERGM are only one of the many possible available models that SNA has. We chose ERGM because of their capacity to model data based on the structural characteristics and the nodes attributes, and also because their assumptions and parameters interpretations are similar to the ones used on conventional statistical methods. However, since ERGM are still under development there are not formal methods for goodness of fit. In addition, there are theoretical ERGM methods that allow working with weighted networks; however, appropriate software tools for such data are not yet available. As an alternative, Cumulative Logistic Regression was used to include the weights associated to the network; however, we were not able to include any of the structural characteristics on the model due to the nature of this conventional statistical method.

Both methods offer different tools depending on what the researcher would like to predict, and the results obtained using the two methods were satisfactory. Table 11 shows the differences between ERGM and Cumulative Logistic Regression.

Table 11

| ERGM | Cumulative Logistic Regression |
|---|---|
| • The social network itself is the dependent variable<br>• The model parameters take into account structural networks characteristics and nodes' attributes<br>• Estimates the log-odds of observing an edge (seeking for advice) between a pair of nodes (employees)<br>• Non-formal methods of Goodness of Fit<br>• For example, being from the same region increases the odds of looking for advice between two employees by exp(1.782)=5.9 | ▪ The weight associated with the edges is the dependent variable<br>▪ The model parameters take into account the nodes' attributes only (and structural parameters are not part of the model)<br>▪ Estimates the cumulative log odds of how frequent the employees seek advice<br>▪ Formal methods of Goodness of Fit<br>▪ For example, for a fixed level position the estimated odds ratio of seeking advice in the US vs. Europe below any level *j* is exp(-1.0939)=0.34 |

## Appendix A: SNA Consulting Firm

### ***I. R Script***

```
library(igraph)
# Section 1: Social Network Analysis
# Network 1 = Observed Data
# Original values
# 0=dont know this person
# 1=never
# 2=seldom
# 3=sometimes
# 4=often
# 5=very often
network=read.table("Network1_info(advise).txt")
colnames(network)=c('ego', 'alter', 'advise_tie')
Network1 = graph.data.frame(network, directed=TRUE)
Matrix1=get.adjacency(Network1, attr="advise_tie")

# Network 2 = From observed data, a new matrix is created: Going for advise (Yes/No)
# New values
# {0,1}=0            No / Do not go for advise
# {2,3,4,5}=1        Yes / Go for advise
n2=read.table("Network1_info(advise).txt")
n2$V3[n2$V3<=1]=0          ## Change values
n2$V3[n2$V3>=2]=1          ## Change values
n2 <- graph.data.frame(n2, directed=TRUE)
Matrix2=get.adjacency(n2, attr="V3")
Network2=graph.adjacency(Matrix2)## Network2

# Centrality measures / Descriptive Analysis
indegree2= as.data.frame(degree(Network2, mode='in'))
outdegree2= as.data.frame(degree(Network2, mode='out'))
between2=as.data.frame(betweenness(Network2))
close2=as.data.frame(closeness(Network2))
evector2=data.frame(evcent(Network2))
central_N2=data.frame(c(1:46), indegree2, outdegree2, between2, close2, evector2$vector)
colnames(central_N2)=c('node', 'indegree', 'outdegree', 'betweenness', 'closeness', 'evector')

# Additionally we can check who are the most 'central' actors according to each measure
head(central_N2[order(-central_N2$indegree),] )
head(central_N2[order(-central_N2$outdegree),] )
head(central_N2[order(-central_N2$closeness),] )
head(central_N2[order(-central_N2$evector),] )

# Lets see if the measures are correlated btwn each other
cor(central_N2[2:6])
```

```
library(network)
library(statnet)

# Setting Matrix2 as an object network Net2
Net2=as.network.matrix(Matrix2, directed=TRUE, type='edgelist')
Net2 %v% 'gender' = c(1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 1, 1, 2, 1, 2, 2, 1, 1, 1, 1, 2,
1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2)
Net2 %v% 'level' = c(3, 4, 1, 4, 4, 4, 3, 5, 2, 4,4, 4, 2, 4, 3, 3, 1, 3, 3, 5, 4, 2, 3, 1, 2, 4, 2, 3, 2, 1,
4, 2, 4, 4, 2, 4, 5, 3, 3, 4, 1, 1, 4, 4, 5, 2)
Net2 %v% 'region' = c(1, 2, 1, 1, 2, 2, 2, 1, 1, 1,2, 1, 2, 2, 1, 1, 2, 2, 2, 2,2, 1, 2, 2, 2, 2, 2, 2, 2,
2,1, 1, 1, 1, 2, 2, 1, 2, 2, 1,1, 1, 1, 2, 2, 1)
Net2 %v% 'location' = c(3, 1, 7, 4, 1, 1, 1, 3, 7, 3, 1, 3, 1, 1, 6, 3, 1, 1, 1, 1, 1, 6, 1, 1, 1, 1, 1, 1, 1,
1, 4, 3, 3, 5, 1, 1, 3, 1, 1, 6, 2, 5, 3, 1, 1, 7)
Net2 %e% "myeval" = Matrix2
Net2

# Plot Network 2, and see how this graph relates to centrality measures
pdf('net.pdf', height=10, width=10) ### It creates a pdf file for the plot
lev=Net2 %v% 'level'
gn=Net2 %v% 'gender'
col=c('red', 'yellow', 'green', 'blue', 'black')
set.seed(310)
plot(Net2, displayisolates=TRUE, displaylabels=TRUE, label.cex=0.7, label.col='black',
vertex.col=col[lev], vertex.sides=c(3,15)[gn], edge.col='gray')
legend('topleft', legend=c('R.A', 'Jr.C.', 'Sr.C.', 'Mng', 'Partner'),
fill= col, cex=0.7)
dev.off()

Table_Level=mixingmatrix(Net2, "level")
Table_Region=mixingmatrix(Net2, "region")
Table_Location=mixingmatrix(Net2, "location")
Table_Gender=mixingmatrix(Net2, "gender")

# Fitting ERGM model for Network2
m1=ergm(Net2~edges)
summary(m1)
m2=ergm(Net2~edges + nodefactor('level')+ nodematch('region')+nodematch('gender'),
seed=120)
summary(m2)
m3=ergm(Net2~edges + mutual + nodefactor('level')+ nodematch('region')+
nodematch('gender'), seed=200)
m3=logLik(m3, add=TRUE)
summary(m3)
m4=ergm(Net2~edges + mutual + triangle + m2star +nodefactor('level')+  nodematch('region')+
nodematch('gender'), seed=150)
```

```
m4=logLik(m4, add=TRUE)
summary(m4)
m5=ergm(Net2~edges + mutual + m2star +nodefactor('level')+ nodematch('region')+
nodematch('gender'), seed=115)
m5=logLik(m5, add=TRUE)
summary(m5)
m6=ergm(Net2~edges + mutual + m2star + gwdsp(0.5, fixed=T) +nodefactor('level')+
nodematch('region')+ nodematch('gender'), seed=100)
m6=logLik(m6, add=TRUE)
summary(m6)

pdf('diagnostics_m5.pdf')
mcmc.diagnostics(m5)
dev.off()
save.image()

pdf('diagnostics_m6.pdf')
mcmc.diagnostics(m6)
dev.off()
save.image()

# GoF (2 ways)
Net5.1=simulate(m5, verbose = TRUE)
Net5.2=simulate(m5,verbose = TRUE)
summary(Net2 ~ edges + density + mutual + m2star + triangle )
summary(Net5.1 ~ edges + density + mutual + m2star + triangle )
summary(Net5.2 ~ edges + density + mutual + m2star + triangle )
### Plots / Histograms
in2= degree(Net2, cmode='indegree')
in5.2 = degree(Net5.1, cmode='indegree')
on2 = degree(Net2, cmode='outdegree')
on5.2 = degree(Net5.1, cmode='outdegree')
par(mfrow=c(2,2))
hist(in2, main='In-Degree Observed Network', xlab=('In-degree'))
hist(on2, main='In-Degree Observed Network',xlab=('Out-degree'))
hist(in5.2, main='In-Degree Observed Network',xlab=('In-degree'))
hist(on5.2, main='Out-Degree Simulated Network',xlab=('Out-degree'))
Net5.1
# Using GoF
m5gof=gof(m5, GOF = ~ distance + triadcensus,
verbose = TRUE, interval = 5e+4)
par(mfrow = c(1,2))
plot(m5gof)
```

## *II. R Output*

▪ Actors with higher in-degrees

```
head(central_N2[order(-central_N2$indegree),] )
         node  indegree  outdegree  betweenness   closeness   evector
           20     24        30       250.99317    0.3061224   1.0000000
           2      21        22       157.23989    0.2922078   0.7577223
           45     19        22       90.36334     0.2903226   0.8810880
           8      18        20       137.30687    0.2848101   0.4497294
           19     18        20       35.40246     0.2727273   0.8452790
           23     18        17       82.50270     0.2830189   0.7500075
```

▪ Actors with higher out-degrees

```
head(central_N2[order(-central_N2$outdegree),] )
          node  indegree  outdegree  betweenness   closeness   evector
           20     24        30       250.99317    0.3061224   1.0000000
           2      21        22       157.23989    0.2922078   0.7577223
           28     18        22        48.46603    0.2777778   0.8785815
           45     19        22        90.36334    0.2903226   0.8810880
           8      18        20       137.30687    0.2848101   0.4497294
           19     18        20        35.40246    0.2727273   0.8452790
```

▪ Actors with higer closeness

```
head(central_N2[order(-central_N2$closeness),] )
         node   indegree  outdegree  betweenness   closeness    evector
           20     24        30       250.99317    0.3061224   1.0000000
           2      21        22       157.23989    0.2922078   0.7577223
           38     18        20       104.84418    0.2903226   0.7003435
           45     19        22        90.36334    0.2903226   0.8810880
           18     11        19        64.67870    0.2866242   0.5711277
           8      18        20       137.30687    0.2848101   0.4497294
```

▪ Actors with higher evector values

```
head(central_N2[order(-central_N2$evector),] )
          node  indegree  outdegree  betweenness   closeness   evector
           20     24        30       250.99317    0.3061224   1.0000000
           45     19        22       90.36334     0.2903226   0.8810880
           28     18        22       48.46603     0.2777778   0.8785815
           19     18        20       35.40246     0.2727273   0.8452790
           26     18        17       53.14866     0.2812500   0.7925086
           2      21        22       157.23989    0.2922078   0.7577223
```

▪ Fitting ERGM for Network2

<u>Summary Model 1</u>

```
Formula:   Net2 ~ edges
Newton-Raphson iterations:  5
```

```
Maximum Likelihood Results:
                   Estimate   Std. Error  MCMC s.e.  p-value
          edges   -1.08961    0.05065       NA          <1e-04 ***
AIC: 2337.7    BIC: 2343.4
```

Summary Model 2

```
Formula:   Net2 ~ edges + nodefactor("level") + nodematch("region") + nodematch("gender")
Newton-Raphson iterations:  5
Maximum Likelihood Results:
                           Estimate   Std. Error   MCMC s.e. p-value
edges                        -5.9730    0.3614         NA        <1e-04 ***
nodefactor.level.2        1.2178     0.1886          NA        <1e-04 ***
nodefactor.level.3        1.6152     0.1852          NA        <1e-04 ***
nodefactor.level.4        1.4205     0.1776          NA        <1e-04 ***
nodefactor.level.5        2.4695     0.2187          NA        <1e-04 ***
nodematch.region        2.6174     0.1481          NA        <1e-04 ***
nodematch.gender       0.5956     0.1326          NA         <1e-04 ***
AIC: 1741.2    BIC: 1780.7
# Note, this second model has a smaller AIC than m1 (a sign of better fit). However this model
doesn't represent any structural characteristics
```

Summary Model 3

```
Formula:   Net2 ~ edges + mutual + nodefactor("level") + nodematch("region") +
nodematch("gender")
MCMC sample of size 10000
Monte Carlo MLE Results:
                            Estimate Std. Error MCMC s.e.  p-value
edges                      -5.18974    0.32425     0.014     < 1e-04 ***
mutual                      2.84500    0.25219     0.008     < 1e-04 ***
nodefactor.level.2       0.73197    0.19083     0.013    0.000129 ***
nodefactor.level.3       1.00266    0.19821     0.011     < 1e-04 ***
nodefactor.level.4       0.90136    0.17638     0.011     < 1e-04 ***
nodefactor.level.5       1.56162    0.22637     0.014     < 1e-04 ***
nodematch.region        1.62815    0.15453     0.002     < 1e-04 ***
nodematch.gender       0.39786    0.05903     0.008     < 1e-04 ***
AIC: 1512    BIC: 1557.1
# Note that all parameters are significant, and AIC got smaller, lets add transitivity parameters
```

Summary Model 4

```
Formula:   Net2 ~ edges + mutual + triangle + m2star + nodefactor("level") +
nodematch("region") + nodematch("gender")
MCMC sample of size 10000
Monte Carlo MLE Results:
                            Estimate Std. Error  MCMC s.e.   p-value
edges                      -3.32527    0.87805     0.840       0.000157 ***
mutual                      2.21303    1.02520     4.243       0.030994 *
```

```
triangle              0.12554   0.02991    0.009      < 1e-04 ***
m2star               -0.10017   0.04689    0.057      0.032773 *
nodefactor.level.2    0.43716   0.45167    0.845      0.333213
nodefactor.level.3    0.69044   0.03213    0.558      < 1e-04 ***
nodefactor.level.4    0.57438   0.05582    0.004      < 1e-04 ***
nodefactor.level.5    1.41958   0.05867    0.001      < 1e-04 ***
nodematch.region      0.18247   0.04139    0.002      < 1e-04 ***
nodematch.gender      0.43055   0.03926    0.000      < 1e-04 ***
AIC: 2221.4    BIC: 2277.8
```

#Note that AIC got a lot bigger, and one parameter isn't significant, lets take out the transitivity parameters

Summary Model 5

```
Formula:   Net2 ~ edges + mutual + m2star + nodefactor("level") + nodematch("region") +
nodematch("gender")
MCMC sample of size 10000
Monte Carlo MLE Results:
                      Estimate    Std. Error MCMC s.e. p-value
edges                 -6.302960   0.201777    3.325     <1e-04 ***
mutual                 2.822515   0.239312    1.147     <1e-04 ***
m2star                 0.058162   0.006892    0.058     <1e-04 ***
nodefactor.level.2     0.653985   0.095175    0.743     <1e-04 ***
nodefactor.level.3     0.706072   0.088235    0.488     <1e-04 ***
nodefactor.level.4     0.710419   0.081561    0.727     <1e-04 ***
nodefactor.level.5     1.168843   0.033310    0.000     <1e-04 ***
nodematch.region       1.781939   0.019971    0.004     <1e-04 ***
nodematch.gender       0.346361   0.017694    0.001     <1e-04 ***
 AIC: 1487.5    BIC: 1538.2
```

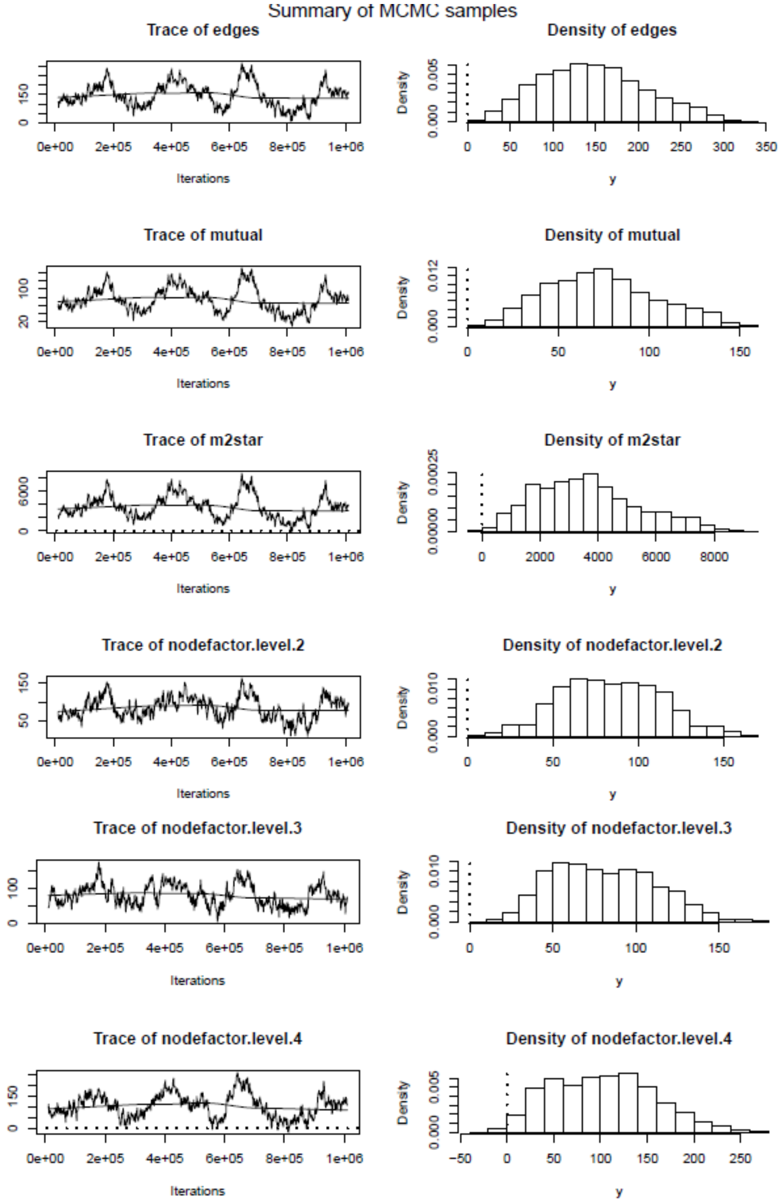
#Note that model improved again, although, lets try to include transitivity with the most robust term for triangle (term gwdsp)

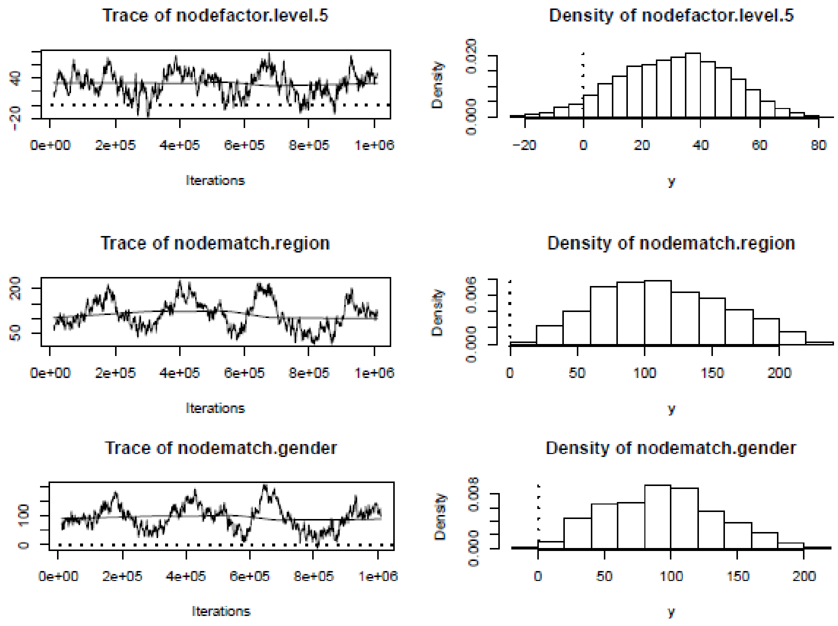
Summary Model 6

```
Formula:   Net2 ~ edges + mutual + m2star + gwdsp(0.5, fixed = T) + nodefactor("level") +
nodematch("region") + nodematch("gender")
MCMC sample of size 10000
Monte Carlo MLE Results:
                      Estimate    Std. Error    MCMC s.e.  p-value
edges                 -3.128491   1.135427    12.309       0.005915 **
mutual                 2.571611   0.666113     9.691       0.000117 ***
m2star                 0.031090   0.034721     0.112       0.370664
gwdsp.fixed.0.5       -0.229674   0.070096     0.762       0.001068 **
nodefactor.level.2     0.165406   0.015157     0.206       < 1e-04 ***
nodefactor.level.3     0.467801   0.018140     0.001       < 1e-04 ***
nodefactor.level.4     0.290019   0.005846     0.003       < 1e-04 ***
nodefactor.level.5     1.352092   0.040972     0.031       < 1e-04 ***
nodematch.region       0.645542   0.012776     0.011       < 1e-04 ***
nodematch.gender     0.367843    0.008031    0.001          < 1e-04 ***
AIC: 2335.2    BIC: 2391.6
#Note that AIC increased again, and not all parameters are significant.
```

We keep Model 5: Even though model 5 doesn't have the smallest AIC, it is the only one that produced a non-degenerate model (which in our case is good!)

Figure 10 – Diagnostic Plots Model 5





Trace of nodefactor.level.5

Density of nodefactor.level.5

Trace of nodematch.region

Density of nodematch.region

Trace of nodematch.gender

Density of nodematch.gender

## Appendix B:  Cumulative Logistic Regression Consulting Firm

### ***I.R Script***

```
level.adviceA=c(21,11,9,7,0,94,42,22,12,10,65,53,40,20,20,147,73,51,27,46,29,32,20,18,8)
level.adviceA = matrix(level.adviceA, nrow=5,ncol=5, byrow=TRUE)
dimnames(level.adviceA)=list(Level=c('R.A', 'Jr.C','Sr.C','Mg.C','Partner'),
Freq.Advice=c('1','2','3','4','5'))
level.adviceA
gender.adviceA= c(296,174,117,63,82,60,37,25,21,2)
gender.adviceA=matrix(gender.adviceA, nrow=2,byrow=TRUE)
dimnames(gender.adviceA)=list(Gender=c('Male','Female'), Freq.Advice=c('1','2','3','4','5'))
gender.adviceA
region.adviceA=c(208,88,41,15,24,148,123,101,69,60)
region.adviceA=matrix(region.adviceA, nrow=2, byrow=TRUE)
dimnames(region.adviceA)=list(Region=c('Europe','US'),Freq.Advice=c('1','2','3','4','5'))
region.adviceA
level.rows=apply(level.adviceA,1,sum)
level.proportions=level.adviceA/level.rows
region.rows=apply(region.adviceA,1,sum)
region.proportions=region.adviceA/region.rows
gender.rows=apply(gender.adviceA, 1, sum)
gender.proportions=gender.adviceA/gender.rows
# Plot Level Proportions
plot( c(1:5),level.proportions[ , 1], type="b", ylim=c(0,0.7), lty=1, xlab="Level Position",
ylab="Proportion")
lines(c(1:5), level.proportions[ , 2],lty=2, type="b", col=2)
lines(c(1:5), level.proportions[, 3], lty=3, type="b", col=4)
lines(c(1:5), level.proportions[, 4], lty=4, type="b", col=3)
lines(c(1:5), level.proportions[, 5], lty=5, type="b", col=14)
legend('topright',legend=c("Never", "Seldom", "Sometimes", "Often", "Very often"),
lty=c(1,2,3,4,5),col=c(1,2,4,3,14))
### Plot Region Proportions
plot(region.proportions[ , 1], type="b", ylim=c(0,0.7), xaxt='n', lty=1, xlab="Region",
ylab="Proportion")
```

```
lines(region.proportions[ , 2],lty=2, type="b", col=2)
lines(region.proportions[, 3], lty=3, type="b", col=4)
lines(region.proportions[, 4], lty=4, type="b", col=3)
lines(region.proportions[, 5], lty=5, type="b", col=14)
legend('topright',legend=c("Never", "Seldom", "Sometimes", "Often", "Very often"),
lty=c(1:5),col=c(1,2,4,3,14))
axis(1, at=1:2, label=c('Europe','USA'))
### Plot Gender Proportions
plot(gender.proportions[ , 1], type="b", ylim=c(0,0.6), xaxt='n',  lty=1, xlab="Gender",
ylab="Proportion")
lines(gender.proportions[ , 2],lty=2, type="b", col=2)
lines(gender.proportions[, 3], lty=3, type="b", col=4)
lines(gender.proportions[, 4], lty=4, type="b", col=3)
lines(gender.proportions[, 5], lty=5, type="b", col=14)
legend('topright',legend=c("Never", "Seldom", "Sometimes", "Often", "Very often"),
lty=c(1:5),col=c(1,2,4,3,14))
axis(1, at=1:2, label=c('Male','Female'))
```

### ***II.R Output***

Tables and plots are shown on Section II – 3

### ***III.SAS Script***

```
*Organizational level and Advice;
data level_advice; input level advice count;
datalines;
1 1 21
1 2 11
1 3 9
1 4 7
1 5 0
2 1 94
2 2 42
2 3 22
2 4 12
2 5 10
3 1 65
3 2 53
3 3 40
3 4 20
3 5 20
4 1 147
4 2 73
4 3 51
4 4 27
4 5 46
5 1 29
```

```
5 2 32
5 3 20
5 4 18
5 5 8
;
run;
*when the number of responses categories excess two, by default the proc ligstic fits the
cumulative model;
proc logistic; weight count;
model advice=level / lackfit; run;


* Gender and advice;
data gender_advice; input gender advice1 count1;
datalines;
1 1 296
1 2 174
1 3 117
1 4 63
1 5 82
2 1 60
2 2 37
2 3 25
2 4 21
2 5 2
;
run;
proc logistic; weight count1;
model advice1=gender/ lackfit; run;

**** Region and advice;
data region_advice; input region advice2 count2;
datalines;
1 1 208
1 2 88
1 3 41
1 4 15
1 5 24
2 1 148
2 2 123
2 3 101
2 4 69
2 5 60
;
run;
proc logistic; weight count2;
```

```
model advice2=region/ lackfit; run;
```

*Cumulative Logistic Regression*

```
data advice; input region level advice count;
datalines;
1 1 1 14
1 1 2 7
1 1 3 3
1 1 4 6
1 1 5 0
1 2 1 50
1 2 2 20
1 2 3 5
1 2 4 2
1 2 5 7
1 3 1 45
1 3 2 4
1 3 3 4
1 3 4 0
1 3 5 1
1 4 1 84
1 4 2 44
1 4 3 23
1 4 4 2
1 4 5 14
1 5 1 15
1 5 2 13
1 5 3 6
1 5 4 5
1 5 5 2
2 1 1 7
2 1 2 4
2 1 3 6
2 1 4 1
2 1 5 0
2 2 1 44
2 2 2 22
2 2 3 17
2 2 4 10
2 2 5 3
2 3 1 20
2 3 2 49
2 3 3 36
2 3 4 20
2 3 5 19
2 4 1 63
```

```
2 4 2 29
2 4 3 28
2 4 4 25
2 4 5 32
2 5 1 14
2 5 2 19
2 5 3 14
2 5 4 13
2 5 5 6
;
run; proc logistic; weight count;
model advice=region level / lackfit; run;
```

***IV.SAS Output***

Goodness of Fit for Level, Region and Gender  tables

*Level / Advice GoF*

**Model Information**

| | |
|---|---|
| **Data Set** | WORK.LEVEL_ADVICE |
| **Response Variable** | Advice |
| **Number of Response Levels** | 5 |
| **Weight Variable** | Count |
| **Model** | cumulative logit |
| **Optimization Technique** | Fisher's scoring |

**Score Test for the Proportional Odds Assumption**

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 1.2421 | 3 | 0.7429 |

*Region / Advice GoF*

**Model Information**

| | |
|---|---|
| **Data Set** | WORK.REGION_ADVICE |
| **Response Variable** | advice2 |
| **Number of Response Levels** | 5 |
| **Weight Variable** | count2 |
| **Model** | cumulative logit |
| **Optimization Technique** | Fisher's scoring |

**Score Test for the Proportional Odds Assumption**

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 6.1538 | 3 | 0.1044 |

*Gender / Advice GoF*

**Model Information**

| | |
|---|---|
| **Data Set** | WORK.GENDER_ADVICE |
| **Response Variable** | advice1 |
| **Number of Response Levels** | 5 |
| **Weight Variable** | count1 |
| **Model** | cumulative logit |
| **Optimization Technique** | Fisher's scoring |

**Score Test for the Proportional Odds Assumption**

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 12.6665 | 3 | 0.0054 |

*Advice ~ Region + Gender, GoF*

**Model Information**

| | |
|---|---|
| **Data Set** | WORK.ADVICE |
| **Response Variable** | Advice |
| **Number of Response Levels** | 5 |
| **Weight Variable** | Count |
| **Model** | cumulative logit |
| **Optimization Technique** | Fisher's scoring |

**Score Test for the Proportional Odds Assumption**

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 8.4404 | 6 | 0.2076 |

## References

Agresti, Alan. *Analysis of Ordinal Categorical Data*. 2nd ed. Hoboken: Wiley, 2010. Print.

Dobson, Annette J., and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. 3rd ed. Boca Raton: CRC Press, 2008. Print.

Bandyopadhyay, Suraj, A R. Rao, and Bikas K. Sinha. *Models for Social Networks with Statistical Applications*. 1st ed. Los Angeles: SAGE Publications, 2011. Print.

Carrington, Peter J., John Scott, and Stanley Wasserman. *Models and Methods in Social Network Analysis*. 1st ed. New York: Cambridge University Press, 2005. Print.

Cross, Rob, and Andrew Parker. *The Hidden Power of Social Networks*. 1st ed. Boston: Harvard Business Press, 2004. Print.

Handcock, Mark S.,David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris (2003). statnet: Software tools for the Statistical Modeling of Network <http://statnetproject.org>

Hanneman, Robert, and Mark Riddle. *Introduction to social network methods*. 2005. Web.

Kolaczyk, Eric D. *Statistical Analysis of Network Data: Methods and Models*. 1st ed. New York: Springer, 2009. Print.

Knoke, David, and Song Yang. *Social Network Analysis*. 2nd ed. Los Angeles, London:SAGE Publications, 2008. eBook.

Opsahl, Tore. "Defining Weighted Networks." N.p., n. d. Web. Web. 9 Dec. 2012. <http://toreopsahl.com/tnet/weighted-networks/defining-one-mode-networks/>.

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. *An introduction to exponential random graph (p*) models for social networks.* Social Networks, Volume 29, Issue 2, May 2007, Pages 173-191 <http://www.sciencedirect.com/science/article/pii/S0378873306000372>