3-2007

Examining Clandestine Social Networks for the Presence of Non-Random Structure

Joshua S. Seder

EXAMINING CLANDESTINE SOCIAL NETWORKS FOR THE PRESENCE OF NON-RANDOM STRUCTURE

THESIS

Joshua S. Seder, Captain, USAF

AFIT/GOR/ENS/07-24

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY
Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GOR/ENS/07-24

EXAMINING CLANDESTINE SOCIAL NETWORKS FOR THE PRESENCE OF NON-RANDOM STRUCTURE

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Joshua S. Seder, B.A.

Captain, United States Air Force

March 2007

AFIT/GOR/ENS/07-24

EXAMINING CLANDESTINE SOCIAL NETWORKS FOR THE PRESENCE OF NON-RANDOM STRUCTURE

Joshua S. Seder, BA
Captain, United States Air Force

Approved:

Dr. Marcus B. Perry, PhD (Chairman) Assistant Professor of Operations Research	Date
Dr. Richard F. Deckro, DBA (Member) Professor of Operations Research	Date

Abstract

This thesis develops a tractable, statistically sound hypothesis testing framework for the detection, characterization, and estimation of non-random structure in clandestine social networks. Network structure is studied via an observed adjacency matrix, which is assumed to be subject to sampling variability. The vertex set of the network is partitioned into k mutually exclusive and collectively exhaustive subsets, based on available exogenous nodal attribute information. The proposed hypothesis testing framework is employed to statistically quantify a given partition's relativity in explaining the variability in the observed adjacency matrix relative to what can be explained by chance. As a result, valuable insight into the true structure of the network can be obtained. Those partitions that are found to be statistically significant are then used as a basis for estimating the probability that a relationship tie exists between any two vertices in the complete vertex set of the network. The proposed methodology aids in the reduction of the amount of data required for a given network, focusing analyses on those attributes that are most promising. Ample effort is given to both model demonstration and application, including an example using open-source data, illustrating the potential use for the defense community and others.

For my Wife and Daughter

Acknowledgments

I would like to thank my advisor, Dr. Marcus Perry, for providing me with an interesting topic and guiding me along the path to completion. Thanks also to my reader, Dr. Richard Deckro, for sharing his vast knowledge of information operations with me.

Thanks to both Captain Michael Burns and James Morris for sharing an analyst point-of-view, allowing me to apply this research to real-world applications. To Captain Jason Williams, for showing me that "yes, you do need to understand how the program works to properly debug it". To David Tesdal, who, although he does not know it, motivated me to persevere when I would have rather thrown in the towel.

Finally, and most importantly, I want to thank my family for their constant support and understanding. I worked long hours and many weekends, but they were always in my thoughts and prayers.

Joshua Seder

Table of Contents

List of Figures

List of Tables

EXAMINING CLANDESTINE SOCIAL NETWORKS FOR THE PRESENCE OF NON-RANDOM STRUCTURE

1. Introduction

1.1. Background

In the Global War on Terror, United States forces are pitted against an entrenched enemy practicing guerilla tactics to wage an asymmetric war. While much is known about the enemy, much still remains hidden. The more information which is available, the better US forces are able to combat the hidden enemy. Certainly, fighting against foot soldiers is not the only primary goal of US efforts, but rather striking a blow to the heart of the terrorist network responsible for the recruitment and employment of enemy troops is a central element of strategy. As more information is uncovered, the task which remains is to gauge the quality of the data. Spending man-hours on analyzing extraneous information is a waste of both time and effort, while man-hours spent analyzing salient information is surely the most beneficial. To this end, research was conducted where the main goal was a framework capable of sifting data to reveal the most promising routes to pursue.

1.2. Overview

Clandestine social networks are comprised of individuals, some or all of whom are, attempting to operate in secret. Because of the desire to remain undetected, players in a clandestine social network practice operational security (OPSEC) and military deception (MILDEC). Any observation of a clandestine network is influenced by

OPSEC and MILDEC measures. These measures help the network to remain hidden and mislead the observer's perception of the network layout. Because of this, the observations of a clandestine social network can be likened to spotting an iceberg. As with an iceberg, the majority of the network is often hidden and the total size can only be estimated. The part that is observed may represent only a fraction of the total network. Due to this hidden nature, any observations of the players and links in a clandestine network must be inspected to ensure accuracy and that observations were not influenced by deceptive actions.

The goal of this thesis is to examine clandestine social networks for the presence of non-random structure by employing a hypothesis test. As constructed, the hypothesis test will be tractable and capable of estimating undirected dyad probabilities between network nodes. The observed network dyads are stored in an adjacency matrix, but, due to OPSEC and MILDEC measures, it is assumed that some level of noise / error is introduced in the process of gathering network data.

The developed test uses nodal attributes to partition the observed network into levels which are mutually exclusive and collectively exhaustive. The network structure variability found in the adjacency matrix is tested to see if it is explained by the partition under investigation. If the attribute partition does explain the observed network structure variability, it is investigated in an attempt to yield insight into the true network's structure. If the attribute partition does not explain the observed network structure variability, it is discarded from further analysis. One major benefit of this is to help analysts sift through the mountains of data on hand and focus on the salient network partitions.

1.3. Problem Statement and Research Objectives / Focus

The goal of this research is to:

1) Test observed clandestine social networks for the presence of non-random structure based on nodal attribute partitions.
2) Identify the attributes explaining adjacency matrix variability.
3) Estimate the probability of the existence of arcs.
4) Perform social network analyses based on these findings.

Attribute partitions appearing to explain network structure are used to further analyze the observed network. If a particular partition explains the variability in the observed adjacency matrix more than another partition, it must be weighted to reflect its contribution to the network's structure. Given this weighted potential, the probability of the existence of arcs can be calculated by taking into account all the partitions explaining adjacency matrix variability.

One of the major goals of this research was to develop a hypothesis testing framework capable of explaining the true structure of the network found in the adjacency matrix. Certainly, due to the members of the network practicing OPSEC and MILDEC, there may exist arcs that are not observed. Through study of the network attribute partitions, the probability of dyad formation can be estimated. These arcs are based upon the network attribute partitions explaining adjacency matrix variability and are weighted according to each attribute's contribution to the formation of the observed network. This will be extremely useful when a new organization arises where nothing is known about either structure or layout.

1.4. Assumptions

A brief list of model assumptions is presented here. While these assumptions are elaborated upon in the Methodology section, presenting them here gives the reader a quick reference to the scope of this model.

1) The arcs of the observed network are undirected.
2) A complete set of attribute data (used for the creation of the partitions) is available for each network node.
3) Partitions of the network are mutually exclusive and collectively exhaustive.

1.5. Implications

This research presents a model which can objectively test observed social networks for structure based on attribute partitions. The ultimate goal of clandestine social network study is to understand the structure so that analysts can see past OPSEC and MILDEC measures, yielding the ability to disrupt the network. Since social networks are dynamic by nature, the same is also true for clandestine social networks. The network will evolve to adapt to disruption, resulting in a new network where ties are modified in order to carry on the work of the previous network. It is this current network to which the model needs to be applied next. Different partitions might now better explain the variability of the observed adjacency matrix than during previous network iterations. Identifying these attribute partitions is possible using the network model. While focused on clandestine networks of interest to national defense forces, the approach may be applied to traditional sociological social networks.

1.6. Preview

While clandestine networks may be large, smaller operational networks are used in this study for the sake of brevity. These networks are partitioned based upon the attributes of the individuals comprising the network. These partitions are examined to see if they adequately explain the variability found in the observed adjacency matrix. Focusing on the partitions explaining network variability, further network measures and tools are applied. The results of the model demonstrate the potential use it has for the United States Air Force and wider Department of Defense community.

2.1. Introduction

In the study of social networks, a variety of tools and measures are available to aid analysis. While some were developed specifically to study social networks, many have been gleaned from various other technical disciplines. For instance, statistical network models have been developed to aid the study of dyad existence. The first section of this chapter focuses on the network specific measure of centrality, which seeks to find the key individuals in a social network. The second section provides a general overview of the key pertinent models. In many cases, different models motivated subsequent development, as researchers endeavored to improve the models currently in use. Finally, the statistical method of clustering is detailed. Clustering, which groups data points based on analyst specified attributes, is applicable to social networks where the data points are comprised of the individuals in the network. While taken from the technical field of statistics, clustering is a method with direct application to social networks.

2.2. Centrality Measures

Network modeling focuses on abstractly representing a construct used to transmit some sort of material or information (Wasserman and Faust, 1994: 4). Two everyday examples of physical networks are the postal system transporting mail and the airplane transportation system moving people. In either case, the point of both networks is to transport some kind of material from one location to another, but the material transported need not be physical. An example of non-physical material transportation is information passed between two people such as baseball game scores or personal reviews of a film.

This information can pass from one person to the next, creating a non-tangible network to transmit this information. Social network analysis (SNA) focuses on modeling a network of people interacting together. The issue of centrality has motivated much of the research done in the field of SNA. A core goal of SNA centrality is to pinpoint the most important individual in a social network based on their position in the network (Frank, 2002: 385).

Finding the central node, in this case an individual person, yields a different result depending on the network being modeled. In the case of a newly observed terrorist network, the central or well connected node might be the best guess as to the leader or possibly the individual failing to follow OPSEC practices. In the case of a group of baseball fans, the results would be wholly different. Finding the most central node might pinpoint the individual who attended the game the previous night, the person whose cable subscription happens to carry the game, or perhaps just the most avid fan of the group. The issue of centrality must be considered in light of the network being observed. This applies to both physical and non-physical networks.

In order to further explore this issue, five measures of social network centrality are reviewed: Degree Centrality, Betweenness Centrality, Closeness Centrality, Informational Centrality, and Eigenvector Centrality. All these measures of centrality are explored in respect to networks containing undirected arcs. While representing the more popular measures, this list should not be viewed as all inclusive, as SNA centrality measures are the focus of much research, both past and present.

2.2.1. Degree Centrality

Degree Centrality simply counts the number of arcs attached to each node. The higher the number of arcs, the more central the node is structurally in the network. With this measure, the node's centrality index can range from [1, ∞), assuming that any node not connected to the network is not represented in the model. Rather than having large numbers attached to each node, it is useful to normalize the degree centrality value. This is done by simply dividing the total number of arcs attached to the node by the total number of arcs in the network (Wasserman and Faust, 1994: 178) and yields a degree centrality statistic in the range of [0, 1], where the higher the statistic value, the more central the node.

An example of degree centrality can be seen in a social network modeling a lieutenant and his flight (see Figure 2-1). Every node (person) in the flight is under the command of the "lieutenant" node. In fact, there is a direct arc from the lieutenant to every flight member. A graph of this model would look like a hub and spoke, where the lieutenant node is in the middle and all the flight members are gathered around him. The lieutenant, able to command the entire flight, is the most central node with a degree centrality measure of one.

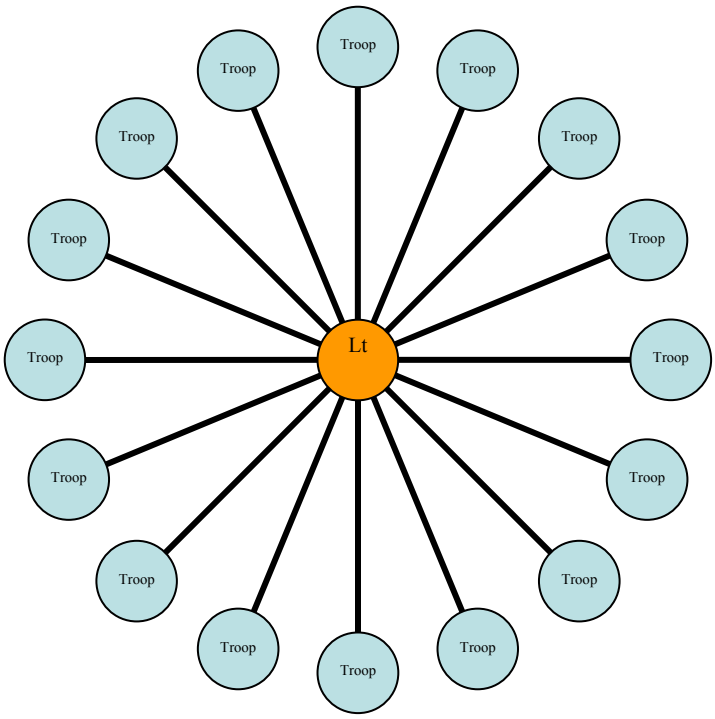

Figure 2-1: Example of Degree Centrality

2.2.2. Betweenness Centrality

Betweenness Centrality can be thought of as dealing with the "middleman" (Borgatti, 2005: 60). The middleman is the node somewhere in the chain between the two endpoints that cannot be avoided. According to the betweenness centrality measure, a node that cannot be avoided is the central node in the network. In order to understand the concept of betweenness, the idea of a "geodesic" is introduced. A geodesic is the shortest path from one point in a network to another (Wasserman and Faust, 1994: 110). That is, if two paths connect node A to node B where one path is three arcs in length and the other is five arcs in length, then the path with three arcs is the geodesic. The case of

more than one shortest path with the same number of arcs is referred to as multiple geodesics.

To find the normalized betweenness centrality statistic of a network, the total number of times a particular node is traversed for all geodesics of the network is enumerated. Next, the total nodes in the network minus one is multiplied by the total nodes in the network minus two and then divided by two. This value is used to divide the total number of times the network was traversed (Wasserman and Faust, 1994: 188). Like the other centrality statistics, the normalized betweenness statistic ranges from [0, 1].

To illustrate betweenness, one can consider an Air Force Scientific and Technical Information (STINFO) office (see Figure 2-2). The job of a STINFO office is to review all published material before dissemination to ensure that the material meets the proper classification level of the receiving organization, in this case a contractor. In this way, all published material must pass through the chokepoint of the STINFO office. The example has the STINFO node positioned between the AF organization and the contractor organization. On one side of the STINFO node lies the AF organization's network, while on the other side of the contractor node is the contractor organization's network. According to the betweenness centrality measure, the STINFO office is most central, as all published material must pass through it. The STINFO office is "between" the AF organization and all other organizations with a betweenness centrality measure of .7556.

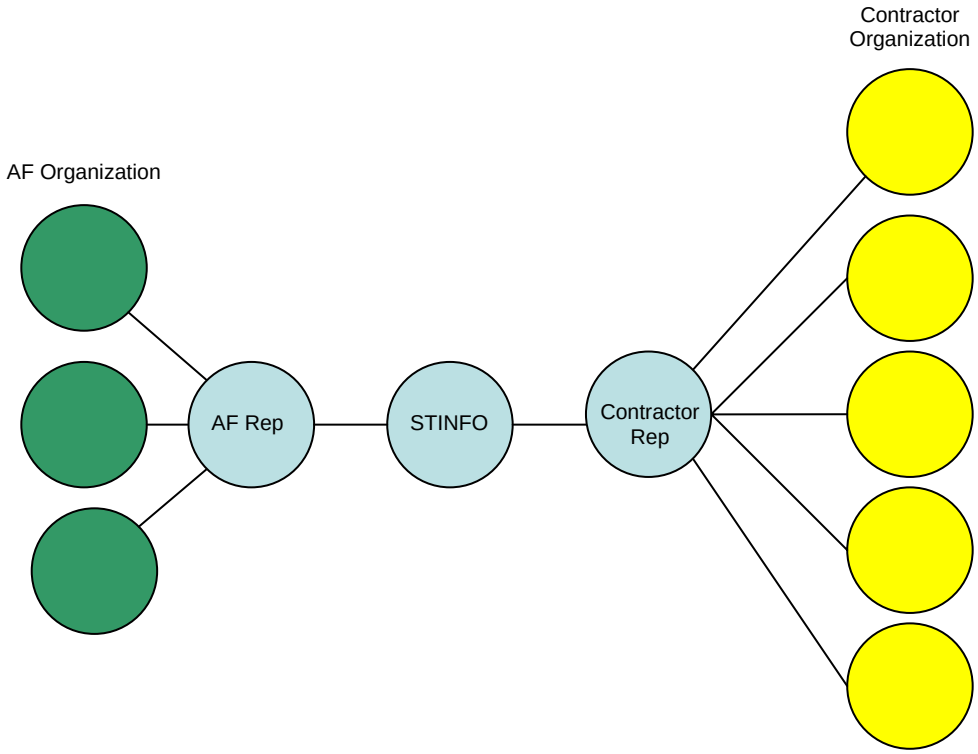

Figure 2-2: Example of Betweenness

2.2.3. Closeness Centrality

Like betweenness, Closeness Centrality also makes use of geodesics, measuring how far one node is from all the other nodes in the social network. The node with the highest measure of closeness centrality is the one able to reach the rest of the nodes in the network over the minimum amount of arcs. In order to find the closeness statistic, the number of geodesics for each node must be calculated. Next, the total number of nodes minus one is divided by the total number of geodesics for each individual node (Borgatti, 2005: 59). The normalized closeness centrality statistic, like the betweenness statistic, is in the range of [0, 1]. The node having the highest closeness statistic is considered the most central node. In the case of small networks, this is fairly simple, but, as the network

grows, so does the difficulty of identifying the geodesics of each node, motivating the analyst to make use of network programs and tools.

To illustrate the closeness centrality measure, think of a secretary servicing an office in a business organization (see Figure 2-3). Not only does the secretary service the entire office, but the secretary also serves as the most direct link between that office and others in the company organization. In servicing just the office, the secretary can be viewed as being the most central node by both the degree and closeness centrality measures. Due to the direct link to the other offices in the organization, the secretary is viewed as the most central with a closeness centrality measure of one.

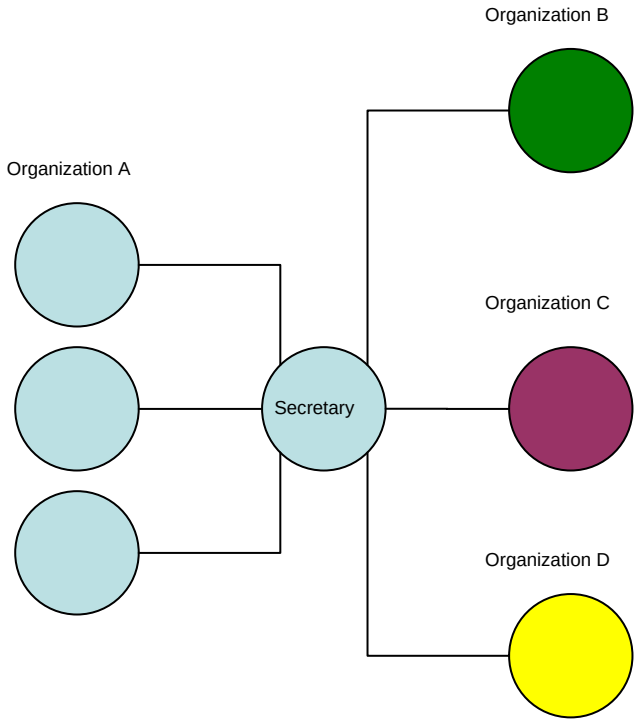

Figure 2-3: Example of Closeness

2.2.4. Informational Centrality

Unlike the centrality measures of degree, closeness, and betweenness, the measure of Informational Centrality relates the amount of information each node (person) transmits (Wasserman and Faust, 1994: 192). In fact, not only does the normalized informational centrality statistic lie in the range of [0, 1], the sum of all the nodal statistics must equal to one. This means the measure assumes that the more information one node has the less all the other nodes have. Should one node have a statistic of one, the rest of the nodes have a statistic of zero, indicating that only one person in the network has the information necessary for the network to operate. Should that node be eliminated, the network would crumble. In this manner, informational centrality identifies the central node.

Calculation of the informational centrality statistics is done partly though linear algebra to manipulate a matrix. The matrix used for the manipulation is a sociomatrix and contains all the numeric arc lengths of the network (Wasserman and Faust, 1994: 70). Once the manipulation of the sociomatrix is complete, its diagonal elements are used in the final calculation of the normalized informational centrality statistic.

Since informational centrality measures the amount of information emanating from a node, an example of this is a network model representing the information flow between the President of the United States and Lt Bagofdonuts (see Figure 2-4). In this model, the President node and the Lt node are directly connected. Theoretically, information flows in both directions. Obviously, the informational centrality statistic for the President will be close to, if not exactly, one. Because of this, the Lt's informational centrality statistic is close to, if not exactly, zero. While a network model of just the

President and the Lt is an oversimplified example, it clearly illustrates the idea of informational centrality.

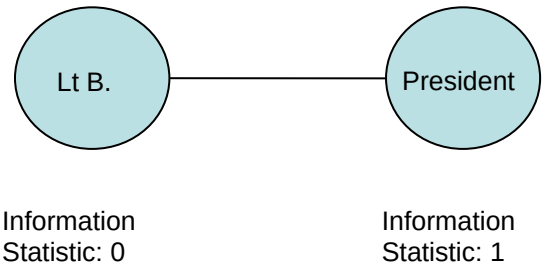

Figure 2-4: Example of Informational Centrality

2.2.5. Eigenvector Centrality

Eigenvector Centrality, like informational centrality, also involves linear algebra. Eigenvector centrality requires that a "correlation" matrix be constructed. Similar to an adjacency matrix, the correlation matrix is an N x N matrix containing the details of the arcs and nodes constructing the social network where the columns and rows correspond to the nodes of the network (Bonacich, 1972: 113). The elements of the correlation matrix are either 1 or 0 depending on whether an arc exists between nodes or does not, respectively. For the sake of illustration, call the N x N correlation matrix Q. If an arc exists between nodes A and B, element $Q_{AB} = 1$; otherwise, $Q_{AB} = 0$. Using this test, all the elements of the correlation matrix are populated. Note that in a social network of undirected arcs, $Q_{AB} = Q_{BA}$. Because of this, the correlation matrix is symmetric with the diagonal row being zero, as there is no arc from a node to itself. Eigenvector centrality requires a symmetric matrix. Should the social structure have directed arcs

such that an arc is not reciprocated, a non-symmetric correlation matrix would occur and the measure of eigenvector centrality cannot be used.

To find the eigenvector centrality statistic, the eigenvalues of the correlation matrix are calculated. The eigenvalues are then rank ordered from largest to smallest. The eigenvector corresponding to the largest eigenvalue is the most central node of the network (Bonacich, 1972: 114). Conversely, the eigenvector corresponding to the smallest eigenvalue is the least central node. With the use of computer programs to calculate the eigenvectors and eigenvalues of matrices, the most difficult part of eigenvector centrality often resides in correctly modeling the observed social network.

The problem with using the centrality measures of degree, betweenness, and closeness is that the analyst must decide ahead of time on which centrality measure to focus. Because of this, these three measures are descriptive rather than prescriptive. While it is useful to know which nodes are central, predicting which nodes have the potential to become central is also valuable. Informational centrality is also descriptive and relies heavily on expert opinion when assigning values to the sociomatrix. These assigned values are then used to find the most central nodes.

In an attempt to find a prescriptive measure, eigenvector centrality is a positive step. It bases the centrality measure solely upon the correlation matrix and does not look for a particular network relationship. Unfortunately, eigenvector centrality falls short of being purely prescriptive, as it only uses the correlation matrix to determine the most central nodes. Since the correlation matrix only takes into account existing arcs, it is unable to weight the values of the arcs. Weighting could be based on a number of

personal social aspects ranging from marital status to criminal background to level of education.

Applying eigenvector centrality to the informational centrality network example, an arc observed between Lt Bagofdonuts and the President carries the same weight, regardless of the obvious fact of the President being much more central than the Lieutenant. Eigenvector centrality does not take into consideration any possible weighting between nodes. Accounting for social attributes could improve the usefulness of eigenvector centrality, possibly yielding a predictive measure rather than just a descriptive measure.

2.3. Statistical Models

Social networks are constructed to facilitate the study of interactions within a social structure. Through SNA, relationships between people are studied for meaning (Wasserman and Faust, 1994: 3). Statistical models have been developed to objectively test the significance of these social network relationships. It is important to understand the development of current SNA statistical modeling techniques in order to better grasp the results of these statistical tests.

Before detailing the statistical models used for social networks, however, a few basic concepts must be reviewed. First, in social networks nodes represent people and arcs represent a relationship between two people. An arc connecting two nodes is called a "dyad" (Wasserman and Faust, 1994: 18). In the Greek language, dyad means two, while in sociology, dyad refers to a pairing. For a network containing directed arcs (see Figure 2-5) dyads can take on three possible states: null, mutual, and asymmetric (Monge

and Noshir, 2003: 117). The null state is when there is no arc connecting the pair of nodes. The mutual state is when two nodes have arcs pointing to each other. The asymmetric state is when one node has an arc pointing to another node, but there is no arc from the receiving node back to the originating node. In dealing with a social network containing undirected arcs (see Figure 2-6) there are only two possible states for the dyads: null and mutual.

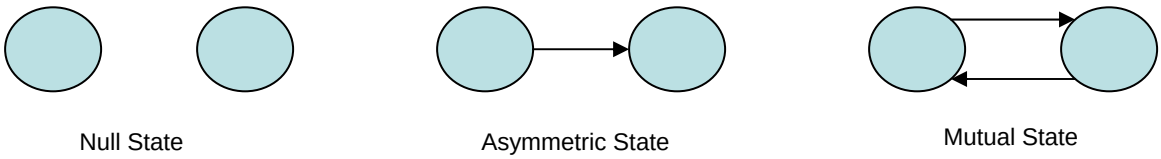

Figure 2-5: Three States of Directed Arcs

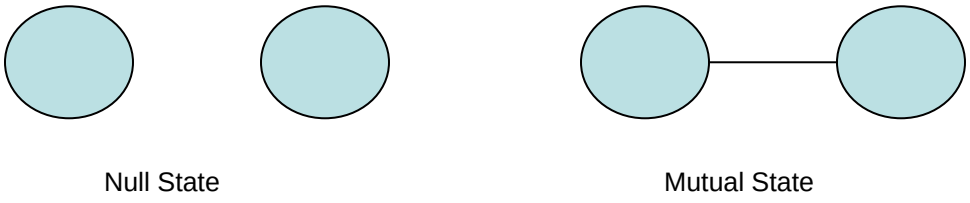

Figure 2-6: Two States of Undirected Arcs

When studying dyads, an N x N matrix is used to track links between nodes. This is called an "adjacency matrix", as it shows which nodes are connected (adjacent) to each other through the presence of dyads. The elements of an adjacency matrix are one if a dyad exists and zero otherwise. In the case of undirected arcs, the matrix will be symmetric with zeros down the diagonal, as ties are reciprocated between nodes and a node cannot have a tie to itself in a social network.

Similar to an adjacency matrix is a "sociomatrix" which represents dyads presence where some kind of binary attribute is tested. For example, say that people at a sports bar are broken up into two subgroups based on gender. Next, focusing on just the female subgroup, an N x N matrix is constructed of the dyads present due to friendships. Again, the elements of this matrix are one when a dyad exists and zero otherwise. This N x N matrix is a sociomatrix of female friends at the sports bar. In essence, sociomatrices employ the probabilistic operation of conditioning on an adjacency matrix to test for dyads found due to a specific binary trait. Like adjacency matrices, the sociomatrices are symmetric with zeros down the diagonal due to reciprocated arcs and the inability for a node to have an arc to itself.

By calculating the row sums for both adjacency matrices and sociomatrices, the total number of arcs out of each node can be found (Fienberg, Meyer, and Wasserman, 1985: 52). Similarly, the column sums yield the total number of arcs into each node (Fienberg, Meyer, and Wasserman, 1985: 52).

2.3.1. p_1 Model

Statistical modeling of social networks yields the ability to test hypotheses about both specific links and linked groups of people contained in the network (Wasserman and Faust, 1994: 606). Statistical tests are conducted by testing an existing hypothesis (often called a "null") compared to an alternative hypothesis. In order to test the statistical significance of a dyad, Paul Holland and Samuel Leinhardt developed the "p_1" model (Holland and Leinhardt, 1981: 33). Originally presented at the "Advanced Research

Symposium on Stochastic Process – Models of Social Structure" in 1977, a paper detailing their model was published in 1981 (Holland and Leinhardt, 1981: 33).

The p_1 model makes use of an adjacency matrix to calculate the number of "in-degrees" and "out-degrees" of each node. In-degrees and out-degrees are each node's corresponding column sum (arcs in) and row sum (arcs out), respectively (Holland and Leinhardt, 1981: 35). Next, the total number of arcs for each node is found by adding the in-degrees and out-degrees together (Holland and Leinhardt, 1981: 35). Finally, the total number of arcs in the network is found by summing the total arcs for each node and dividing by two (Holland and Leinhardt, 1981: 35). The in-degrees, out-degrees, and total number of arcs are used in the construction of the p_1 model, but other parameters are also added, allowing the desired network probability distribution to be included (Holland and Leinhardt, 1981: 37). Including in-degrees and out-degrees takes into account how connected each node is (Fienberg, Meyer, and Wasserman, 1985: 54). To complete the p_1 model, it is divided by a scaling factor to normalize the total probabilities of the model and ensure that they sum to one (Holland and Leinhardt, 1981: 36).

In developing the p_1 model, Holland and Leinhardt successfully found a way to test dyadic ties for probability of existence. Unfortunately, the p_1 model was limited to only being able to test one dyad. In addition, while the scaling factor can be calculated for smaller networks, as networks grow, so does the difficulty of determining the normalizing constant. The difficulty of calculating the scaling factor was a primary motivation of a majority of research aimed at improving the p_1 model.

2.3.2. Markov Graphs

Building on the p_1 model, Frank and Strauss developed a way to probabilistically model arcs between nodes using Markov Graphs (Frank and Strauss, 1986: 832). The Markov property, incorporated into their graphs, is that when transitioning between states, only the current state influences what the next state will be (Kulkarni, 1995: 16). Making use of this property, Markov graphs model dyads whose existence only depends on the originating node (Frank and Strauss, 1986: 832). In the case of undirected arcs, this means that $\text{Prob}(i \rightarrow j \mid i) = \text{Prob}(j \rightarrow i \mid j)$.

To simplify their model, Frank and Strauss assumed that all arcs are equally probable (Frank and Strauss, 1986: 836). Like the p_1 model, the Markov graph model uses adjacency matrices to calculate the probability of the existence of most likely or least likely dyads or subgraphs (Frank and Strauss, 1986: 841).

Because of the ability to statistically test the probability of subgraph existence, the Markov graph model was the next step in the development of statistical network models. In fact, through setting certain parameters equal to zero, the Markov graph model simplifies down to the p_1 model from which it was developed (Frank and Strauss, 1986: 836). A drawback of the Markov graph model is that it does not factor in the arc weights of the dyads, resulting in equal probability for all dyads. This means that all arcs have the same probability of existence, regardless of the likelihood of arc existence.

2.3.3. Markov Chain Monte Carlo

Further investigation of the Markov property reveals that it is present in a fair amount of statistical models through the implementation of Markov Chain Monte Carlo

(MCMC). Many models require division by a normalizing factor to ensure that all the probabilities sum to one. This normalizing factor is often difficult to calculate. MCMC permits this process by generating an estimate which is used as the scaling factor, thus eliminating the need for arduous calculation of the actual scaling factor (Gilks, 1996: 3).

The total probability of an event occurring is simply the desired outcome(s) divided by the total number of possible outcomes. In the discrete case, the total number of outcomes can be enumerated. Unfortunately, total enumeration is not an option for the continuous case. In order to calculate the total number of outcomes, integration must be employed to "enumerate" all the possible outcomes. Both forms of enumeration essentially yield the expected value (Gilks, 1996: 3). Unfortunately, there are many instances where integration yields a value of infinity, which is not useful to the analyst. This is where Monte Carlo random number drawing comes into play.

Since Monte Carlo methods draw random numbers in order to estimate a finite number of outcomes, Monte Carlo integration is simply the average of all the outcomes (Gilks, 1996: 4). According to the law of large numbers, if enough sample observations are obtained, the average generated by those samples should approximate the average of the total population (Wackerly, Mendenhall, and Scheaffer, 2002: 423). This applies even if the population is infinite. Therefore, the expected value yielded by the Monte Carlo integration can be used as the normalizing constant.

The issue that arises from using Monte Carlo integration is the need for stable observations. Since the outcome space is infinite, the outcomes observed can vary drastically. This is where another aspect of the Markov property comes into play. The presence of the Markov property ensures that a steady-state exists for the system

(Kulkarni, 1995: 107). Finding the steady-state of the Markov Chain yields the observations used for Monte Carlo integration to calculate the final normalizing constant.

To find the normalizing constant, the Metropolis-Hastings algorithm is often employed (Gilks, 1996: 7). This algorithm generates the Markov chain and finds the steady-state distribution values used for Monte Carlo integration. Gibbs sampling is an application of the Metropolis-Hastings algorithm where Markov chain transitions are calculated purely through conditioning (Gilks, 1996: 7). Due to the application of purely conditional transitioning, Gibbs sampling is incorporated into many of the MCMC techniques currently in use (Gilks, 1996: 7).

While MCMC techniques eliminate the need to explicitly calculate the actual normalizing constant, calculations are still required to find the estimated normalizing constant. While the estimation of the normalizing constant is much simpler than explicit calculation, eliminating the need to even find a normalizing constant would be simpler yet.

2.3.4. Stochastic and p_1 Blockmodels

In an effort to deconstruct adjacency matrices, Stochastic Blockmodels make use of "blocks" to subgroup dyads due to binary attributes (Wang and Wong, 1987: 9). The blocks count the number of dyads present in the subgroup and calculate a statistic representing the possibility of falling into that block. This statistic is simply the number of dyads in the block divided by the total possible number of dyads (Wang and Wong, 1987: 9). Similar to the hypothesis test proposed by this study, the blocks can be viewed as "partitions" of the adjacency matrix.

Using the p_1 model on the blocks of the adjacency matrix yields the "p_1 blockmodel" (Wang and Wong, 1987: 11). p_1 blockmodels have the ability to conditionally test dyads based upon the binary attributes of the blocks, factoring in the block probability statistics. The drawback is that the block breakdown can possibly skew the statistical results depending on how the dyads are sorted (Wang and Wong, 1987: 11). The methodology proposed herein seeks to counter this problem of skewing statistical results while maintaining the ability to conditionally test dyads based on partition assignment.

2.3.5. p* and Logit p* Models

Combining the p_1 and Markov Random Graph models, the p* model is not limited to just a single dyad (Wasserman and Pattison, 1996: 406). In fact, the p* model is able to test the probability of the existence of any subgraph, even if the "subgraph" includes the entire network. Where the p_1 model could be generalized to a subgraph pertaining to two nodes (p_2 model), the p* model is applicable to the entire network if necessary.

The p* model necessitates the modification of a sociomatrix into three sociomatrices based upon a specific binary trait (Wasserman and Pattison, 1996: 406):

1) The sociomatrix of all the dyads present due to that binary trait
2) The sociomatrix of all the dyads not present due to that binary trait
3) The "compliment" of the initial sociomatrix

Using these three sociomatrices, the p* model is calculated like the p_1 model, which still requires the computation of a scaling factor (Wasserman and Pattison, 1996: 406).

Due to the difficulty of computing the p* model's scaling factor, the "logit p* model" was developed in order to eliminate the need to calculate any scaling factors (Anderson, Wasserman, and Crouch, 2002: 46). This was done through the employment of the "odds ratio". The odds ratio is simply the chance or "odds" of the event(s) happening divided by the chance or "odds" of the event(s) not happening (Montgomery, Peck, and Vining, 2001: 446). In the case of the p* model, the odds ratio is calculated based upon the binary conditioning trait (Wasserman and Pattison, 1996: 407). Taking the (natural) log of the odds ratio is called the "logit" (Wasserman and Pattison, 1996: 407) and makes the odds ratio easier to manipulate mathematically. To this end, the log of the odds ratio of the p* model is taken, resulting in the logit p* model (Wasserman and Pattison, 1996: 407). Without the need for a scaling factor, the logit p* model uses binary traits to test the probability of the existence of subgraphs (Anderson, Wasserman, and Crouch, 2002: 48).

2.4. Clustering Techniques

The goal of clustering is to group data in such a way that data points containing similar traits are gathered together. In the field of SNA, the data points studied are the network nodes where the clusters, often referred to as subgroups, depend on the traits and positions of these nodes. There are a myriad of clustering techniques available when creating subgroups. While each technique operates differently, they all make use of some kind of algorithm to group the data points, where the rules of the algorithm determine how the clusters are built. An overview of these techniques is now be presented.

At the top level, clustering techniques can be broken down into two broad categories: hierarchical clustering and partitioning (Dillon, 1984: 167). The major difference between these two categories is how they treat data points when incorporated into clusters. With hierarchical clustering, once a data point is incorporated into one of the major clusters, it stays in that cluster, regardless of whether it might fit better in another cluster formed at a later time (Dillon, 1984: 168). Partitioning allows data points to change clusters, if doing so will yield a better set of overall clusters (Dillon, 1984: 186).

2.4.1. Hierarchical Clustering

The two most basic types of hierarchical techniques are agglomeration and division (Manly, 2005: 125). Agglomeration starts by considering every data point to be an individual cluster (Jain, 1999: 274). Next, a measurement is computed to find which data points are close together based on a user specified tolerance. Generally, this measurement is the Euclidean distance between points (Manly, 2005: 130). Once the points that are close together are determined, they are grouped into a new cluster. Again, the distance from the single points to the new clusters are computed. Single points found to be close to the centroid of existing clusters are assimilated into those clusters. This is done until all the points are included in a cluster. If two clusters are found to be close enough to each other to satisfy the set tolerance, those two clusters are combined into a single cluster (Manly, 2005: 127).

The hierarchical technique of division works in the exact opposite way as agglomeration. The divisive technique starts with all the data points together in one large

cluster that is then broken down into smaller clusters (Jain, 1999: 274). First, the centroid of the cluster is found. Next, the location of each data point is measured against this average. Again, Euclidean distance is useful in computing the location of the data point versus the average for the cluster (Manly, 2005: 130). Data points that are found to have a distance larger than a set tolerance are split off into a new cluster. This divisive process is then applied to the new clusters. Once the existing clusters all fall within the set tolerance, the clusters of the divisive technique are set (Manly, 2005: 128).

There are a several drawbacks to these techniques. One is that they require measuring the distance between data points. While Euclidean distance is often the simplest method, other methods exist which can be utilized (Manly, 2005: 62). Regardless of the distance method chosen, enough must be known about the data that a suitable method can be used with a specified tolerance. This requires prior knowledge of the data in order to specify both clustering and distance methods, and is a general shortcoming which arises when nonparametric distance-based methods are used.

Another drawback is that they exclude a data point once a decision is made about it. For instance, in the agglomeration method, once a data point is included in a cluster, it cannot be removed from that cluster and, therefore, is not allowed to be moved to another cluster. Even if it is evident that it does not belong to the cluster it is located in, it stays there. The same is true for the divisive method. Once a cluster is split, the points contained in the new cluster are not considered again. Even if after creating more clusters, a data point should be included in a new or different cluster, it stays in the cluster to which it was originally assigned. The issue of switching clusters is allowed through the practice of partitioning, which will be covered later. The benefit of

hierarchical clustering is that once the data point is assigned to a cluster, it is no longer included in future assignment calculations. This results in hierarchical methods generally being faster than partitioning methods.

2.4.2. Monothetic and Polythetic Clustering

Another form of clustering is to break data points down due to a binary trait. This is called monothetic clustering (Jain, 1999: 274). With binary traits, a data point either has the desired element, or does not. Based on this binary trait, the data points are clustered into two categories. These clusters need not be geographically located near each other as in the agglomeration and division methods.

Polythetic clustering is an extension of the monothetic clustering technique where more than one binary attribute is taken into account (Jain, 1999: 274). When using the polythetic clustering method, the initial clustering of the data is done the same way as monothetic clustering. With the data points divided into two clusters, all the data points are tested for another binary attribute. Based upon the attribute, each of the existing clusters is broken down into two more clusters, resulting in a total of four clusters. This breakdown of all the clusters continues for every binary attribute.

The use of the polythetic clustering technique results in the number of clusters growing quickly; for k binary attributes, 2^k clusters are created (Jain, 1999: 274). Using polythetic clustering to test for k = 1 binary attribute is the same as monothetic clustering. This is evident, as $2^1 = 2$ clusters, the same number found using monothetic clustering. While the polythetic method rapidly creates clusters of the data, the side effect is that the

number of clusters quickly becomes unruly. Testing for too many traits can easily result in too many clusters, making analysis difficult.

2.4.3. Hard vs. Fuzzy Clustering

Agglomerative, divisive, monothetic, and polythetic clustering techniques are all examples of clustering that is categorized as being "hard". Hard clustering is any technique that only allows data points to be part of one cluster (Jain, 1999: 274). Partitioning also shares this trait of only allowing each data point to be included in one cluster. In order to allow data points to be included in more than one cluster, the method of "fuzzy" clustering must be employed.

To allow inclusion in multiple clusters, each data point is given a probability of being contained in each cluster (Jain, 1999: 281). Like all probabilities, these values lie in the range of [0, 1]. The closer to one the probability is the more chance of being included in that cluster (Jain, 1999: 281). The closer to zero, the less chance the data point is included (Jain, 1999: 281).

Using this method, fuzzy clustering allows data points to be included in more than one cluster. Provided there is more than one cluster, there exist data points that will not be included in all the clusters. A data point not included in a cluster has a probability of zero. This is represented in two different ways. Clusters can explicitly state that the data point has no chance of being included in the cluster, or it can just be left out of the set (Jain, 1999: 281). Both approaches have their benefits.

Explicitly listing the probabilities of all the data points allows them to be rank ordered based upon the probability values. This ranking can be used to quickly identify

which data points belong at the center of the cluster, which ones belong in the outskirts of the cluster, and which ones do not belong in the cluster at all. On the other hand, not listing the points outside the cluster will result in a shorter cluster list to study. The benefit is a list where all the data points are rank ordered without the extraneous points cluttering up the list. The downside is that to know which points are not included in the cluster requires the use of a master list containing all the data points. Without this list, each cluster must be individually found on the lists of the other clusters. Both approaches have their benefits and should be used depending on the analyst's needs.

2.4.4. Partitioning

A partition can be viewed as a dynamic cluster whose edges grow and shrink to find the optimal cluster grouping. As mentioned previously, partitioning provides more flexibility than hierarchical clustering, allowing data points to switch clusters, while, with hierarchical clustering, once a cluster incorporates a data point, it is set as being part of that cluster. If, due to some calculation, it is found that a data point does not fit the current partition, partitioning removes the data point from the current cluster and places it in another cluster.

One of the most widely used nonparametric distance based partitioning method is the k-means clustering technique. The "k" in k-means is the total number of clusters desired (Dillon, 1984: 186). The desired number of total clusters is specified at the start of executing the k-means partitioning method. This can pose a problem, as it requires additional knowledge of the data prior to running the k-means algorithm.

Once the value of k is determined, the data points are broken up into k partitions (Dillon, 1984: 186). Next, the center of each partition is found. With the center of each partition known, the distances of all the partition's data points to the center are calculated for all the partitions. Euclidean distance can once again be used, as in the agglomerative and divisive techniques (Dillon, 1984: 186). Next, the mean partition distance is calculated by taking the arithmetic mean of the total distance of all the data points from the center (Dillon, 1984: 186).

With these means calculated, an error term for each partition is created using the squared-error method (Dillon, 1984: 187). The squared-error method calculates the error of each partition for all the data points in that partition using the following formula:

$$\sum_{i=1}^{k} (\text{partition mean} - \text{partition location of node i})^2$$

Once each partition's error term is found, these terms are summed into one overall error term (Dillon, 1984: 187).

The goal of k-means partitioning is to minimize the overall error term (OET). This is done using an iterative technique to move data points from one cluster to another (Jain, 1999: 279). Upon moving data points to another partition, the OET is recalculated. If it shrinks, the partition change was beneficial and is kept. If it grows, the partition change was negative and the algorithm reverts to the previous partition set. At the start of this process, a desired OET tolerance decrease must be specified. Each iteration's OET is compared to the OET of the prior partition set. If the OET does not improve

enough to satisfy the threshold of the specified tolerance, the current partition pattern is accepted as the optimal set of clusters.

Just like with k-means clustering, many other partitioning techniques also require the total number of clusters to be specified before running the algorithm. This is the major weakness of nonparametric distance-based partitioning methods. Often times, a partitioning technique is run multiple times on the same data set where a different number of total clusters are specified for each run (Jain, 1999: 278). It is then up to the analyst to determine which specified number of clusters is the preferred choice for the data being studied.

Clustering has a direct application to social networks, using network nodes as the data points for the clusters. A few ways to form network clusters are by focusing on nodal attributes, network location, or similar ties to other nodes. Creating network clusters could help to identify groups and organizations embedded in an observed network, and also provide a way to quickly categorize a newly observed member of the network. Through the use of clustering, the leadership of an organization can be isolated for study. Once these individuals are pinpointed, the effects of removing them from the network can be explored, perhaps resulting in the identification of the network's future leaders.

Since most people usually belong to more than one social group, network clusters must allow nodes to belong to more than one cluster. This means that network clustering should make better use of fuzzy methods than hard methods. Use of fuzzy methods ensures a dynamic network model where individuals have the potential to be members of all the clusters. Such a model will allow for changing alliances, friendships, and making

new connections. Keeping these aspects in mind, clustering techniques have great potential when applied to SNA.

2.5. Summary

A great deal of work has taken place in the field of social network analysis. While much of it has been conducted solely for application to social networks, some tools and methods have been taken from other disciplines. Regardless of origin, the main goal was to gain greater insight into the formation and interactions taking place within a social network. With the knowledge of the development existing tools and their application, this research proposes a new methodology with which to investigate non-random structure in social networks based on a hypothesis testing framework. The main goal of this effort is a tool for use by the Department of Defense community.

3. Methodology

3.1. Introduction

This chapter details the methodology of developing a hypothesis test capable of investigating a social network for non-random structure. The methodology developed herein partitions the observed network based on the social attributes of the nodes, providing the ability to test each attribute partition to see if it explains the variability found in the adjacency matrix. To this end, the first part of the chapter focuses on model development and is followed by examples showing the application of the model and corresponding hypothesis test.

3.2. Variables

In order to develop the model used to conduct the hypothesis test, network variables must be specified to allow parameterization of a probability mass function. These variables are:

$n \equiv$ Number of Nodes Observed in the Network

$N \equiv$ Total Possible Ways Dyads Can Form in the Observed Network

$n_h \equiv$ Number of Nodes Observed in Level h

$N_h \equiv$ Total Possible Ways Dyads Can Form in Level h

$n_{ij} \equiv$ Number of Nodes Observed in Level i or Level j

$N_{ij} \equiv$ Total Possible Ways Dyads Can Form from Level i to Level j

$d \equiv$ Number of Observed Dyads in the Network

$d_h \equiv$ Number of Observed Dyads in Level h

$d_{ij} \equiv$ Number of Observed Dyads from Level i to Level j

$p_h \equiv$ Probability of an Arc Existing Between any Two Nodes Contained in Level h

$p_{ij} \equiv$ Probability of an Arc Existing Between a Node in Level i and a Node in Level j

$p_0 \equiv$ Probability of an Arc Existing Between any two Nodes in the Observed Network

$z(k) \equiv$ Exhaustive k Level Partition

3.3. Model Development

Using the network variables, a probability mass function (PMF) can be constructed capable of assigning a probability to dyad existence conditional on the k level partition. That is, the binomial distribution is parameterized in terms of the k level network partition as follows:

$$f\left(\mathbf{d} \mid \mathbf{p},\, z(k)\right) = \prod_{h=1}^{k} \binom{N_h}{d_h} p_h^{\,d_h} (1-p_h)^{N_h - d_h} \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} \binom{N_{ij}}{d_{ij}} p_{ij}^{\,d_{ij}} (1-p_{ij})^{N_{ij} - d_{ij}} \qquad (3.1)$$

Both p_h and p_{ij} are in the interval [0, 1]. In addition, $d_h \in \{0, 1, \ldots, N_h\}$ and $d_{ij} \in \{0, 1, \ldots, N_{ij}\}$ where $N_h = (½)(n_h)(n_h - 1)$ and $N_{ij} = (½)(n_{ij})(n_{ij} - 1) - N_i - N_j$. With the variables d_h, d_{ij}, N_h, and N_{ij} calculated, they can be used to find d and N:

$$d = \sum_{h=1}^{k} d_h + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d_{ij} \tag{3.2}$$

$$N = \sum_{h=1}^{k} N_h + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_{ij} \tag{3.3}$$

Notice that the PMF in (3.1) is constructed by assuming conditional independence based on the exhaustive k level partition. Letting D_h and D_{ij} denote the events of observing d_h dyads in level h and d_{ij} dyads from level i to level j, respectively, the initial conditioning is

$$p\left(D_1 \cap D_2 \cap D_3 \cap \cdots \cap D_k \cap D_{1,2} \cap D_{1,3} \cap \cdots \cap D_{1,k} \cap D_{2,k} \cap \cdots \cap D_{k-1,k} \mid z(k)\right) \tag{3.4}$$

and by the conditional independence assumption, (3.4) can be written as

$$\begin{aligned} &p\left(D_1 \mid z(k)\right) p\left(D_2 \mid z(k)\right) p\left(D_3 \mid z(k)\right) \cdots p\left(D_k \mid z(k)\right) p\left(D_{1,2} \mid z(k)\right) \\ &p\left(D_{1,3} \mid z(k)\right) \cdots p\left(D_{1,k} \mid z(k)\right) p\left(D_{2,k} \mid z(k)\right) \cdots p\left(D_{k-1,k} \mid z(k)\right) \end{aligned} \tag{3.5}$$

The PMF in (3.1) is defined to all partition levels $k \in \{1, 2, \ldots, n\}$. The upper and lower bounds of k present two special cases for the PMF. At $k = 1$, (3.1) becomes:

$$f\left(\mathbf{d} \mid \mathbf{p},\, z(1)\right) = \binom{N}{d} p^d (1-p)^{N-d} \tag{3.6}$$

In (3.6), there is no partition and, as such, only one parameter is required.

At $k = n$, (3.1) becomes:

$$f\left(\mathbf{d}\,|\,\mathbf{p},\ z\left(k\right)\right) = \prod_{i=1}^{k-1}\prod_{j=i+1}^{k}\binom{N_{ij}}{d_{ij}} p_{ij}^{\ d_{ij}} \left(1-p_{ij}\right)^{N_{ij}-d_{ij}} \tag{3.7}$$

Using (3.7), each network node is placed in its own level. Since n levels are present, the number of parameters needed is $(½)(n)(n-1)$.

While (3.6) presents a case where no dyads occur between levels, (3.7) presents a case where no dyads occur within a level. In order to explore dyad formation both within and across levels, meaning that at least two nodes are assigned to each level, bounds must be placed on k such that

$$2 \le k \le \left\lfloor \frac{n}{2} \right\rfloor \tag{3.8}$$

where $\lfloor\ \rfloor$ denotes the floor function. While the bounds of (3.8) are necessary, they are not sufficient for ensuring that every partition has at least two nodes because it is still possible for partitions with only one node within a level to exist. In the cases where each partition contains at least two nodes, the number of parameters required for k levels is

$$k + \binom{k}{2} \quad \text{or} \quad \left(\frac{1}{2}\right)k(k+1) \tag{3.9}$$

Further, the total number of ways to partition the observed network into u levels resulting in all u levels having at least two nodes is

$$\left(\begin{matrix} \left\lfloor \frac{n}{2} \right\rfloor \\ u \end{matrix} \right); \quad u = 2,3,...,\left\lceil \frac{n}{2} \right\rceil \tag{3.10}$$

where $\lceil \ \rceil$ denotes the ceiling function. In (3.10), n remains the total number of observed nodes and u is the level of the partition under consideration. This provides the analyst with a count of the total number of possible partitions resulting in at least two nodes per level. To illustrate the use of (3.10), consider a $k = 3$ level partition applied to a network of $n = 20$ nodes. There are 120 unique partitions such that each level contains at least two nodes.

Since parameters p_h and p_{ij} ($h = 1, 2, \ldots, k$; $i < j = 2, 3, \ldots, k$) are not known, they must be estimated from the observed adjacency matrix. The likelihood function proportional to (3.1) is:

$$L\left(\mathbf{p} \mid \mathbf{d},\, z\left(k\right)\right) = \prod_{h=1}^{k} p_h^{\,d_h} (1-p_h)^{N_h - d_h} \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} p_{ij}^{\,d_{ij}} (1-p_{ij})^{N_{ij} - d_{ij}} \tag{3.11}$$

The natural log of the likelihood function is preferred over the likelihood function due to ease of manipulation. As such, the natural log of (3.11) is:

$$l\left(\mathbf{p}\,|\,\mathbf{d}, z\left(k\right)\right) = \sum_{h=1}^{k}\left[d_h \log\left(p_h\right) + \left(N_h - d_h\right)\log(1-p_h)\right] + \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left[d_{ij}\log\left(p_{ij}\right) + \left(N_{ij} - d_{ij}\right)\log(1-p_{ij})\right] \tag{3.12}$$

With the log-likelihood function specified, the maximum likelihood estimates for p_h and p_{ij} ($h = 1, 2, \ldots, k$; $i < j = 2, 3, \ldots, k$) can be shown to be (Casella and Berger, 2002: 316):

$$\hat{p}_h = \frac{d_h}{N_h}$$
$$\hat{p}_{ij} = \frac{d_{ij}}{N_{ij}} \tag{3.13}$$

(An explicit demonstration of these estimates for the $k = 3$ level partition is shown in Appendix A.)

Since the estimates yielded by (3.13) are maximum likelihood estimates (MLEs), their asymptotic properties can be exploited to estimate an approximate 100(1 – α)% confidence interval (CI) for each estimated parameter. The benefit of such a calculation is that it provides a method for analysts to gauge the quality of variable estimation. Small CIs indicate better quality of estimation while large CIs indicate lower quality of estimation. Confidence bounds provide lower and upper limits for the CI, yielding a range for $\hat{p}_h$ and $\hat{p}_{ij}$.

It is well known that a MLE of θ, say $\hat{\theta}$, is asymptotically distributed as approximately multivariate normal with $\mathrm{E}\left[\hat{\theta}\right]=\theta$ and $\mathrm{Var}\left[\hat{\theta}\right]=\left[\mathbf{I}\left(\theta\right)\right]^{-1}$ where $\mathbf{I}(\theta)$ is the Fisher information matrix with element $I_{p,q}$ given by

$$I_{p,q}=\mathrm{E}\left[\frac{dl\left(\mathbf{x}\,|\,\theta\right)}{d\theta_p}\bullet\frac{dl\left(\mathbf{x}\,|\,\theta\right)}{d\theta_q}\right] \tag{3.14}$$

Here, l denotes the log-likelihood function and the expectation is taken over the random variables $\mathbf{x}$. For any unbiased estimator, the diagonal elements of the inverse of $\mathbf{I}(\theta)$ are then the Cramer-Rao lower bounds of the variance of the parameter estimates. Thus, since MLEs are asymptotically unbiased, $\mathbf{I}\left(\theta\right)^{-1}$ can be viewed as the asymptotic variance-covariance matrix of the estimator $\hat{\theta}$.

Under the assumed model in (3.1), the asymptotic variance-covariance matrix of $\hat{\mathbf{p}}=\left[\hat{p}_1,\hat{p}_2,\ldots,\hat{p}_{k-1},\hat{p}_k,\hat{p}_{1,2},\ldots,\hat{p}_{1,k},\ldots,\hat{p}_{k-1,k}\right]$, denoted by $\mathrm{Var}\left(\hat{\mathbf{p}}\right)$, is then given by

$$\begin{pmatrix}
\frac{p_1(1-p_1)}{N_1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & \frac{p_2(1-p_2)}{N_2} & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{p_{k-1}(1-p_{k-1})}{N_{k-1}} & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & \frac{p_k(1-p_k)}{N_k} & 0 & \cdots & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & \frac{p_{1,2}(1-p_{1,2})}{N_{1,2}} & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \frac{p_{1,k}(1-p_{1,k})}{N_{1,k}} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & \frac{p_{k-1,k}(1-p_{k-1,k})}{N_{k-1,k}}
\end{pmatrix} \tag{3.15}$$

Thus, the asymptotic distribution of $\hat{\mathbf{p}}$ is approximately multivariate normal with mean vector $\mathbf{p}$ and variance-covariance matrix $\mathrm{Var}(\hat{\mathbf{p}})$. Note that since $\mathrm{Var}(\hat{\mathbf{p}})$ is a diagonal matrix, this suggests that the covariances between the parameter estimates are zero for any given partition $z(k)$ of the observed network.

Using the likelihood function of (3.11), a hypothesis test can be derived for testing the parameters p_h and p_{ij} ($h = 1, 2, \ldots, k$; $i < j = 2, 3, \ldots, k$) for equality. While the null hypothesis proposes all parameters are equal, the alternative hypothesis proposes that at least one parameter differs. That is,

$$H_0 : p_1 = p_2 = p_3 = \ldots = p_k = p_{1,2} = p_{1,3} = \ldots = p_{1,k} = p_{2,k} = \ldots = p_{k-1,k} = p_0$$
$$H_A : p_h \neq p_0 \cup p_{ij} \neq p_0 \text{ for at least one } h \text{ or } i, j \ (i < j) \tag{3.16}$$

The alternative hypothesis of (3.16) can be specified simply or as a composite. The form used is at the discretion of the analyst seeking to explore the variability of the observed adjacency matrix. Through the use of the proposed hypothesis test, adjacency matrix variability can be explored in a quantifiable way to determine which attribute partitions best provide insight into these observations.

Under the null hypothesis, the likelihood function is not conditioned on the partition and takes on the form

$$L_0\left(d \mid p_0, z(1)\right) = L_0\left(d \mid p_0\right) = {p_0}^d \left(1 - p_0\right)^{N-d} \tag{3.17}$$

proposing that network partitions do not explain the variability in the observed adjacency matrix. In fact, (3.17) implies that chance better explains observed variability rather than any of the partitions. For this reason, $z(1)$ is eliminated from the likelihood function, as (3.11) reduces to (3.17) under the null hypothesis. (Proof of this reduction for the $k = 3$ level partition is shown in Appendix A.) As with the other probability parameters, p_0 is contained in the interval [0, 1] and d and N are defined by (3.2) and (3.3), respectively.

The hypothesis test is carried out through the creation of a likelihood ratio statistic, denoted R, and is simply the likelihood function specified under the alternative hypothesis divided by the likelihood function specified under the null hypothesis. In equation form it is:

$$R\left(d,\mathbf{p}\mid z\left(k\right)\right)=\frac{L_1\left(\mathbf{d}\mid\mathbf{p},z(k)\right)}{L_0\left(d\mid p_0\right)}=\frac{\prod_{h=1}^{k}p_h^{\,d_h}\left(1-p_h\right)^{N_h-d_h}\prod_{i=1}^{k-1}\prod_{j=i+1}^{k}p_{ij}^{\,d_{ij}}\left(1-p_{ij}\right)^{N_{ij}-d_{ij}}}{p_0^{\,d}\left(1-p_0\right)^{N-d}} \tag{3.18}$$

and is quite similar to the odds ratio discussed in section 2.3.5. In essence, it is the "odds" that the network formed based upon the partition structure versus forming completely at random.

Working with the natural log of *R* produces the log-likelihood ratio statistic which is easier to manipulate mathematically. Doing so changes (3.18) to

$$\begin{aligned} r\left(d,\mathbf{p}\mid z\left(k\right)\right)&=\log\left(R\left(d,\mathbf{p}\mid z\left(k\right)\right)\right)\\ &=\sum_{h=1}^{k}\left[d_h\log\left(p_h\right)+\left(N_h-d_h\right)\log(1-p_h)\right]\\ &+\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left[d_{ij}\log\left(p_{ij}\right)+\left(N_{ij}-d_{ij}\right)\log(1-p_{ij})\right]\\ &-\left[d\log\left(p_0\right)+\left(N-d\right)\log\left(1-p_0\right)\right], \end{aligned} \tag{3.19}$$

where small values of *r* suggest that the network attribute partition does not significantly explain the variability in the observed adjacency matrix relative to what can be explained by chance. Conversely, large values of *r* indicate that the network attribute partition does significantly explain the variability in the observed adjacency matrix relative to what can be explained by chance.

Unfortunately, the true partition probabilities are not known. This was the reason that partition MLEs were found using (3.13). With these parameters estimated, the only

parameter left to estimate is p_0. Similar to (3.13), it can be shown that the value of p_0 that maximizes (3.19) is given by

$$\hat{p}_0 = \frac{d}{N} \tag{3.20}$$

Plugging all the parameter MLEs into (3.19) produces:

$$\begin{aligned}\hat{r}\left(d,\mathbf{p}\,|\,z(k)\right) = &\sum_{h=1}^{k}\left[d_h \log\left(\hat{p}_h\right)+\left(N_h-d_h\right)\log(1-\hat{p}_h)\right] \\ &+\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left[d_{ij}\log\left(\hat{p}_{ij}\right)+\left(N_{ij}-d_{ij}\right)\log(1-\hat{p}_{ij})\right] \\ &-\left[d\log\left(\hat{p}_0\right)+\left(N-d\right)\log\left(1-\hat{p}_0\right)\right]\end{aligned} \tag{3.21}$$

As calculated by (3.21), $\hat{r}$ is the test statistic used to explore observed adjacency matrix variability. Small values of $\hat{r}$ suggest that the network partition does not significantly explain the variability in the observed adjacency matrix relative to what can be explained by chance and large values of $\hat{r}$ indicate that the network attribute partition does significantly explain the variability in the observed adjacency matrix relative to what can be explained by chance.

Now that the test statistic has been derived, a tractable method for quantifying the significance level of the test must be developed. Since the proposition of the null hypothesis is that no structure is present in the observed adjacency matrix and that network formation is totally random, Monte Carlo simulation is used for the generation of

random networks to compare to the observed network. The random networks have the same number of nodes as the observed network, but the number of dyads is a random variable. Dyads form with a probability of $\hat{p}_0$ and fail to form with a probability of $1 - \hat{p}_0$.

As in most statistical tests, computing the "attained significance level" is necessary for deciding whether or not the null hypothesis stands as stated or is rejected in favor of the alternative hypothesis. The statistic numerically representing the attained significance level is the *p*-value (Wackerly, Mendenhall, and Scheaffer, 2002: 482). Smaller *p*-values imply higher statistical significance. Conversely, larger *p*-values imply lower statistical significance, resulting in failure to reject the null hypothesis.

For s randomly generated networks, log-likelihood ratio statistics will be calculated and sorted in descending order where r_i denotes the log-likelihood ratio statistic of the i^{th} randomly generated network. By comparison to the list corresponding to the randomly generated networks, the observed network's log-likelihood ratio statistic is assigned the ranking r_{obs}. In this way, the estimated *p*-value is obtained by

$$\hat{p}\text{-}value = \frac{rank\left(r_{obs}\right)}{s+1} \tag{3.22}$$

where $rank(r_{obs})$ denotes the rank assigned to the log-likelihood ratio statistic of the observed network as compared to the list of log-likelihood ratio statistics corresponding to the s randomly generated networks. For example, if $s = 4$, $r_{obs} = 13$, $r_1 = 3$, $r_2 = 5$, $r_3 = 7$, and $r_4 = 11$, so $rank(r_{obs}) = 1$ and $\hat{p}$ -value = .2.

With the $\hat{p}$-value found, the attained significance level of the observed network's log-likelihood ratio statistic is compared to α, the statistical significance threshold. If $\hat{p}$-value ≤ α, the test statistic is determined to be statistically significant, thus resulting in rejection of the null hypothesis in favor of the alternative hypothesis. If that is the case, the network partition being investigated does indeed explain the variability in the observed adjacency matrix. The larger the value for *s*, the more significant the $\hat{p}$-value becomes. For this research effort, s = 999 random networks are generated, which is a sufficiently large enough number to result in a quality $\hat{p}$-value estimation.

The results of this test are extremely beneficial as they allow network analysts to objectively characterize the partitions identified as statistically significant and aid in explaining adjacency matrix variability. The network partitions identified as not statistically significant have little or no relativity in explaining the variability in the observed adjacency matrix. Thus, those attributes that serve as a basis for the non-significant partitions can be excluded from future consideration, focusing efforts on attributes with the most potential.

3.4. Model and Methodology Assumptions

The model presented in section 3.3 can handle cases where k = 1, 2, …, n. This research effort focuses on cases involving two to five attribute partition levels, where each level contains at least two nodes.

Full datasets are available upon which to base network partitions. That is, all the nodal attributes from which the partitions are constructed are known.

The network nodes are partitioned into mutually exclusive and collectively exhaustive subsets. This means that for any given partition, a node can only be contained in one level at a time. While different attributes exist, each attribute is focused on individually to create the partition. For example, take the case of an attribute representing whether or not an individual is married. The individual can either be married or single, but not both, and therefore is restricted to one of the two possible levels. Should another attribute correspond to the same node, for instance level of education, another partition is constructed on the basis of "education level".

3.5. Model Demonstration

With the hypothesis testing framework now laid out, it will be applied to three different examples. These examples will detail the methods and application of the hypothesis test used to detect, characterize, and estimate non-random structure in a network, making use of both simulated and real-world network data.

3.5.1. $k = 5$ Level Partition

For the first example, a network with $n = 25$ nodes and a $k = 5$ level partition is constructed. The level assignments are: Level 1 ≡ {1, 3, 5}, Level 2 ≡ {20, 21, 22, 23, 24, 25}, Level 3 ≡ {15, 16, 17, 18, 19}, Level 4 ≡ {2, 4, 6, 7, 8, 9, 10}, and Level 5 ≡ {11, 12, 13, 14}. According to (3.9), fifteen parameters must be estimated, denoted p_1, p_2, p_3, p_4, p_5, p_{12}, p_{13}, p_{14}, p_{15}, p_{23}, p_{24}, p_{25}, p_{34}, p_{35}, and p_{45}. The specified values of these parameters are $p_1 = .6667$, $p_2 = .4667$, $p_3 = .7$, $p_4 = .2857$, $p_5 = .3333$, $p_{14} = .5238$, $p_{23} = .3$, $p_{24} = .0476$, $p_{25} = .1667$, $p_{45} = .0714$, and $p_{12} = p_{13} = p_{15} = p_{34} = p_{35} = 0$. Given the true relationship structure corresponding to these specifications is shown in the

symmetric "dyad probability matrix" of Figure 3-1. One realization of the adjacency matrix based on Figure 3-1 is shown in Figure 3-2 and visually represented in Figure 3-3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	**0**	0.5238	0.6667	0.5238	0.6667	0.5238	0.5238	0.5238	0.5238	0.5238	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2		**0**	0.5238	0.2857	0.5238	0.2857	0.2857	0.2857	0.2857	0.2857	0.0714	0.0714	0.0714	0.07	0	0	0	0	0	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
3			**0**	0.5238	0.6667	0.5238	0.5238	0.5238	0.5238	0.5238	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4				**0**	0.5238	0.2857	0.2857	0.2857	0.2857	0.2857	0.0714	0.0714	0.0714	0.07	0	0	0	0	0	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
5					**0**	0.5238	0.5238	0.5238	0.5238	0.5238	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6						**0**	0.2857	0.2857	0.2857	0.2857	0.0714	0.0714	0.0714	0.07	0	0	0	0	0	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
7							**0**	0.2857	0.2857	0.2857	0.0714	0.0714	0.0714	0.07	0	0	0	0	0	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
8								**0**	0.2857	0.2857	0.0714	0.0714	0.0714	0.07	0	0	0	0	0	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
9									**0**	0.2857	0.0714	0.0714	0.0714	0.07	0	0	0	0	0	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
10										**0**	0.0714	0.0714	0.0714	0.07	0	0	0	0	0	0.0476	0.0476	0.0476	0.0476	0.0476	0.0476
11											**0**	0.3333	0.3333	0.33	0	0	0	0	0	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
12												**0**	0.3333	0.33	0	0	0	0	0	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
13													**0**	0.33	0	0	0	0	0	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
14														**0**	0	0	0	0	0	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
15															**0**	0.7	0.7	0.7	0.7	0.3	0.3	0.3	0.3	0.3	0.3
16																**0**	0.7	0.7	0.7	0.3	0.3	0.3	0.3	0.3	0.3
17																	**0**	0.7	0.7	0.3	0.3	0.3	0.3	0.3	0.3
18																		**0**	0.7	0.3	0.3	0.3	0.3	0.3	0.3
19																			**0**	0.3	0.3	0.3	0.3	0.3	0.3
20																				**0**	0.4667	0.4667	0.4667	0.4667	0.4667
21																					**0**	0.4667	0.4667	0.4667	0.4667
22																						**0**	0.4667	0.4667	0.4667
23																							**0**	0.4667	0.4667
24																								**0**	0.4667
25																									**0**

Figure 3-1: True Network Relationship Structure

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	**0**	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	**0**	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
3	0	0	**0**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	**0**	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	**0**	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	**0**	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	**0**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	**0**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	**0**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	**0**	1	0	0	0	0	0	0	0	0	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	1	0
13	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	0	0	0	0	1	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	1	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	1	1	1	0	1	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	1	1	1	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	1	1	1	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	1	1	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	1	1	1	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	1	1
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Figure 3-2: Simulated Realization Based on Figure 3-1

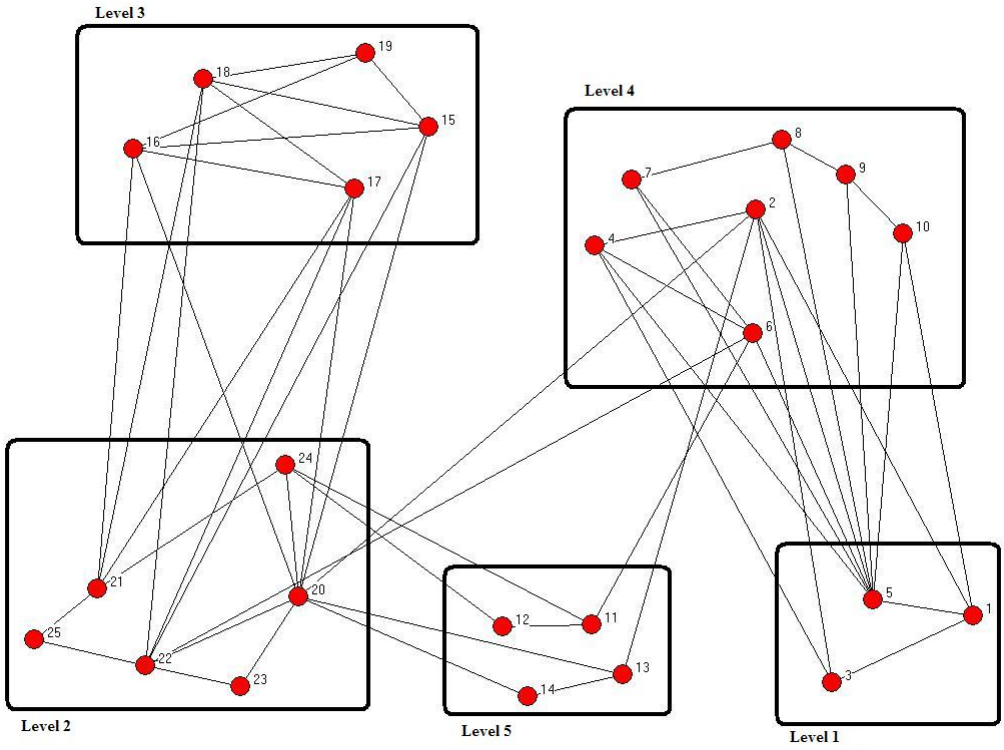

Figure 3-3: Visual Representation of Simulated Realization in Figure 3-2

The PageRank Algorithm (PRA) used by Google is designed to objectively examine the nodes of a network for connectivity (Page and Brin, 2006: n. pag.). In the application used by Google, the nodes of the network are webpages and the overall network is the internet. When applied to dyad probability matrices, the PRA indicates which nodes belong to which groups and also gauges an individual node's overall connectivity to all the other nodes in the network. This is achieved through finding the steady-state of the dyad probability matrix and assigning a "PRA Score" to each node. The PRA Score is a measure of how connected a node is to the rest of the nodes in the network. Nodes with similar connectivity are grouped together. The results of applying the PRA to the dyad probability matrix of Figure 3-1 are shown in Table 3-1.

Table 3-1: PRA Results

Score	Node	
0.0927	11	**Least Connected Nodes**
0.0927	12	
0.0927	13	**Level 5**
0.0927	14	
0.1209	9	
0.1209	10	**Level 4**
0.1209	2	
0.1209	4	
0.1209	6	
0.1209	7	
0.1209	8	
0.1441	1	**Level 1**
0.1441	3	
0.1441	5	
0.2557	20	**Level 2**
0.2557	21	
0.2557	22	
0.2557	23	
0.2557	24	
0.2557	25	
0.2859	15	**Level 3**
0.2859	16	
0.2859	17	
0.2859	18	
0.2859	19	**Most Connected Nodes**

Table 3-1 indicates that the PRA accurately identifies the levels of the network based on the dyad probability matrix and identifies level three as the most connected level. Referring to dyads contained within a level as "level dyads" and dyads occurring across levels as "cross-level dyads", it is evident that level three contains a large number of level and cross-level dyads. This, coupled with the high probability of dyad

occurrence (p_3 = .7), makes level three the most connected level. The PRA results also indicated that level five is the least connected level. Compared to level three, the probability of dyad occurrence in level five is low (p_5 = .3333) and level five has few level and cross-level dyads. Despite the fact that level three only has cross-level dyads with level two while level five has cross-level dyads with both levels two and four, level three remains the most connected level due to high probability of dyad occurrence and a large number of total dyads.

The benefit to social network analysts is that PRA results provide a prioritized list of target nodes. In this case, the most connected target set is level three while the least connected target set is level five. Due to being a high value target, the members of level three might not be easy to attack. If this is the situation, the PRA results also provide the next desirable target set, being level two.

3.5.2. Building Confidence Intervals (CIs)

For an illustration of how to construct CIs, a network comprised of n = 20 nodes will be broken down into a k = 2 level partition based upon a single binary attribute. Level One ≡ {1 – 5, 11 – 15} and Level Two ≡ {6 – 10, 16 – 20}. While this is a simplistic breakdown, it will fully illustrate the methods used to place bounds on parameter estimates. The arbitrary partition probabilities assigned are p_1 = .6, p_2 = .9, and p_{12} = .15. The symmetric dyad probability matrix corresponding to the specifications given for this example is illustrated by Figure 3-4.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	**0**	0.6	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15	0.6	0.6	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15
2		**0**	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15	0.6	0.6	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15
3			**0**	0.6	0.6	0.15	0.15	0.15	0.15	0.15	0.6	0.6	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15
4				**0**	0.6	0.15	0.15	0.15	0.15	0.15	0.6	0.6	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15
5					**0**	0.15	0.15	0.15	0.15	0.15	0.6	0.6	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15
6						**0**	0.9	0.9	0.9	0.9	0.15	0.15	0.15	0.15	0.15	0.9	0.9	0.9	0.9	0.9
7							**0**	0.9	0.9	0.9	0.15	0.15	0.15	0.15	0.15	0.9	0.9	0.9	0.9	0.9
8								**0**	0.9	0.9	0.15	0.15	0.15	0.15	0.15	0.9	0.9	0.9	0.9	0.9
9									**0**	0.9	0.15	0.15	0.15	0.15	0.15	0.9	0.9	0.9	0.9	0.9
10										**0**	0.15	0.15	0.15	0.15	0.15	0.9	0.9	0.9	0.9	0.9
11											**0**	0.6	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15
12												**0**	0.6	0.6	0.6	0.15	0.15	0.15	0.15	0.15
13													**0**	0.6	0.6	0.15	0.15	0.15	0.15	0.15
14														**0**	0.6	0.15	0.15	0.15	0.15	0.15
15															**0**	0.15	0.15	0.15	0.15	0.15
16																**0**	0.9	0.9	0.9	0.9
17																	**0**	0.9	0.9	0.9
18																		**0**	0.9	0.9
19																			**0**	0.9
20																				**0**

Figure 3-4: Dyad Probability Matrix Corresponding to the Two Level Partition of the Twenty Node Example

To construct a 95% CI on each of the parameter estimates, one can make use of the asymptotic variances discussed in section 3.3. With $\alpha = .05$, finding the lower and upper bounds simply requires finding the value associated with points .025 and .975 on the normal curve corresponding to the mean and asymptotic variance of $\hat{p}_i$ (or $\hat{p}_{ij}$) discussed earlier. Table 3-2 summarizes the input values and the results of the calculations for each parameter estimate.

Table 3-2: 95% CI for Twenty Node Example

	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_{12}$
Variable Estimate	0.6	0.9	0.15
Nodes in Partition	10	10	20
Asymptotic Variance	0.0053	0.0020	0.0013
95% Lower Bound	0.5895	0.8961	0.1475
95% Upper Bound	0.6105	0.9039	0.1525

Notice that the asymptotic variance for $\hat{p}_{12}$ is lower than the asymptotic variance for either $\hat{p}_1$ or $\hat{p}_2$. The lower asymptotic variance is the result of having more observations upon which to base the estimate. Twice as many observations are available for the cross-level estimate, thus resulting in improved parameter estimation. In addition, notice that the asymptotic variance of $\hat{p}_2$ is lower than the asymptotic variance of $\hat{p}_1$. This is due to higher probability of dyads in level two. In general, as N_h (or N_{ij}) gets small, the asymptotic variance of p_h (or p_{ij}) gets large, while the asymptotic variance of $\hat{p}_h$ (or $\hat{p}_{ij}$) is minimized at $\hat{p}_h = 0$ or 1.

Corresponding to the specifications of Table 3-2, a box plot is constructed and presented in Figure 3-5. The benefit of a box plot is that it provides a visual tool upon which to evaluate the quality of parameter estimates. If the CI is large, parameter estimation is poor, while, if the CI is small, parameter estimation is of higher quality. For this reason, small CIs are desired. Since box plots present CIs visually, the tighter the bands, the better the quality of the estimates. In the case of Figure 3-5, $\hat{p}_{12}$ is the best quality estimate as shown by the tight confidence interval, while $\hat{p}_1$ is the lowest quality estimate as shown by the wider confidence interval. It should be noted that at the 95% confidence level, all three estimates are of high quality.

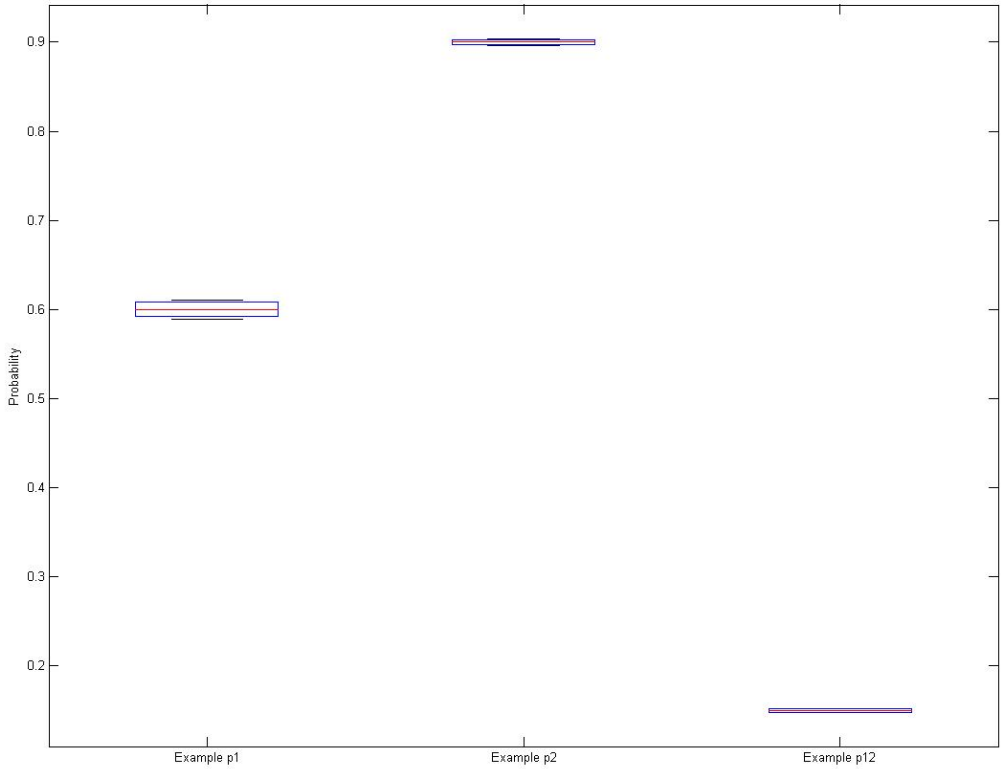

Figure 3-5: 95% Confidence Interval for Probability Estimates of the Twenty Node Network Example

3.5.3. Example Based on Real-World Data

The previous two examples were based on simulated data and only illustrate single partitions. Using the open source, real-world data compiled by Marc Sageman on the Al Qaeda terrorist network (Sageman, 2006: n. pag.), multiple $k = 2$ level partitions will now be explored. Focusing on the first one hundred nodes in the Sageman dataset, the attribute partitions of the "Friendship" network are tested to see if they significantly explain the variability in the observed adjacency matrix. The dyads of the "Friendship" network represent observed friendships between network actors.

As in the previous example, the nodal attributes were broken down in a binary fashion. Despite this breakdown, what is presented here can be expanded to encompass the case of any $k > 2$ level partition. Table 3-3 presents the sixteen binary attributes explored. The methods presented in this thesis require a full dataset and, unfortunately, data was missing from the Sageman dataset. For this research effort, nodes with missing attribute data were given a default value. Table 3-4 details the assumptions required to populate missing data, generally assigning either the lowest level or most common attribute. These assumptions were made to illustrate the example. Exploring appropriate ways of dealing with missing data is suggested as an area of future research.

Table 3-3: Binary Attributes

Attribute	Binary Breakdown
Age Joined the Jihad	1 if Joined the Jihad at 25 or Older; 0 OW
Clump	1 if in Arab Clump (Core or Maghreb); 0 OW
Criminal Background	1 if Any Criminal Background; 0 OW
Date of Birth	1 if Born Before 1970; 0 OW
Fate	1 if Alive; 0 OW
Kids	1 if Has Kids; 0 OW
Level of Education	1 if Some College Education or Greater; 0 OW
Married	1 if Married; 0 OW
Occupation Type	1 if Professional or Semi-Prof Occupation; 0 OW
Place Joined the Jihad	1 if Joined the Jihad in Native Country; 0 OW
Religious Background	1 if Muslim Religious Background; 0 OW
School Type	1 if Attended a Madrassa (Muslim School); 0 OW
Socio-Economic Status	1 if Upper Class; 0 OW
Type of Education	1 if Type of Educ is Scientific (Social / Tech / Natural); 0 OW
Year Joined the Jihad	1 if Joined the Jihad before 1995; 0 OW
Youth National Status	1 if Youth National Status is "Native"; 0 OW

Table 3-4: Assumptions for Missing Data

Attribute	Assumption Made if Missing Data
Age Joined the Jihad	Younger than 25
Clump	Full Data Set
Criminal Background	No Criminal Background
Date of Birth	After 1970
Fate	Full Data Set
Kids	No Kids
Level of Education	No College
Married	Not Married
Occupation Type	Not Professional or Semi-Prof
Place Joined the Jihad	Native Country
Religious Background	Not Muslim
School Type	Madrassa (Muslim School)
Socio-Economic Status	Not Upper Class
Type of Education	Not Scientific
Year Joined the Jihad	After 1995
Youth National Status	Not Native

A "weighted dyad probability matrix" can be constructed using the statistically significant partitions. In order to do this, the log-likelihood ratio statistics of the partitions explaining adjacency matrix variability must be normalized to find a weight for each partition where w_i is the "relativity" weight corresponding to the i^{th} partition. That is,

$$w_i = \frac{\hat{r}\left(\mathbf{e},\mathbf{d} \mid z_i(k)\right)}{\sum_m \hat{r}\left(\mathbf{e},\mathbf{d} \mid z_j(k)\right)} \tag{3.23}$$

where i indicates the log-likelihood ratio statistic for the i^{th} partition and the summation is conducted over all the statistically significant partitions. Since the sum of all w_i are one and w_i is in the range [0, 1], another way to compute $\hat{p}_h$ is

$$\hat{p}_h = \sum_m w_m \left(\hat{p}_h \mid z_m \right) \tag{3.24}$$

where the summation is conducted over all *m* statistically significant partitions.

With a "full" dataset at hand, the ability to find relativity weights, the attributes broken up in a binary fashion, and a statistical significance level threshold of α = .05, the Friendship network is ready for the hypothesis testing framework. Table 3-5 details the results of the hypothesis tests.

Table 3-5: Significant Partitions of the Friendship Network Hypothesis Tests

Attribute	$\hat{p}$ -value	$\hat{r}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_{12}$	Weight
Age Joined	0.020	5.021	0.0108	0.0222	0.0085	0.0243
Clump	0.001	39.3741	0.0333	0.0188	0.0000	0.1902
Criminal Background	0.001	17.3311	0.0130	0.0553	0.0040	0.0837
Date of Birth	0.006	6.3059	0.0145	0.0204	0.0068	0.0305
Fate	0.006	6.6748	0.0301	0.0124	0.0075	0.0323
Kids	0.002	7.9796	0.0158	0.0216	0.0059	0.0386
Level of Ed	0.001	23.6071	0.0095	0.0238	0.0022	0.1141
Occupation Type	0.001	24.9879	0.0101	0.0236	0.0018	0.1207
Place Joined	0.001	9.8076	0.0160	0.0216	0.0035	0.0474
School Type	0.001	8.6639	0.0138	0.0000	0.0000	0.0419
Type of Ed	0.001	12.7106	0.0081	0.0293	0.0073	0.0614
Year Joined	0.001	36.858	0.0317	0.0187	0.0004	0.1781
Youth Nat'l Status	0.003	7.6422	0.0161	0.0170	0.0052	0.0369

From inspection of Table 3-5, it is evident that the "Clump" partition, detailing each individual's assignment within Al Qaeda, has the largest log-likelihood ratio statistic (39.3741) with a relativity weight of .1902. According to the estimated probability of friendship ties, there is no probability of members of either Arab clump being friends

with any of the members of the Southeast Asian or Central Staff clumps. That is, $\hat{p}_{12} = 0$. It is estimated that a 3.3% chance exists for a friendship to form within the Arab clumps ($\hat{p}_1 = .0333$) and an estimated 1.9% chance of a friendship tie forming in the Southeast Asian or Central Staff clumps ($\hat{p}_2 = .0188$). This shows a large divide in the clump partition of the Friendship network. Clearly, friendships do not form between Arab clump individuals and those in the Southeast Asian or Central Staff clumps if the only factor is clump affiliation. This is the sort of knowledge which may be extremely useful to social network analysts seeking ways in which to disrupt terrorist networks by eliminating individual nodes. The 95% confidence intervals corresponding to the estimated clump parameters are presented in Figure 3-6. With $n_1 = 36$, $n_2 = 64$, and $n_{12} = 100$, parameter estimation is of high quality. As in section 3.5.2, the confidence intervals are again computed using the asymptotic variances.

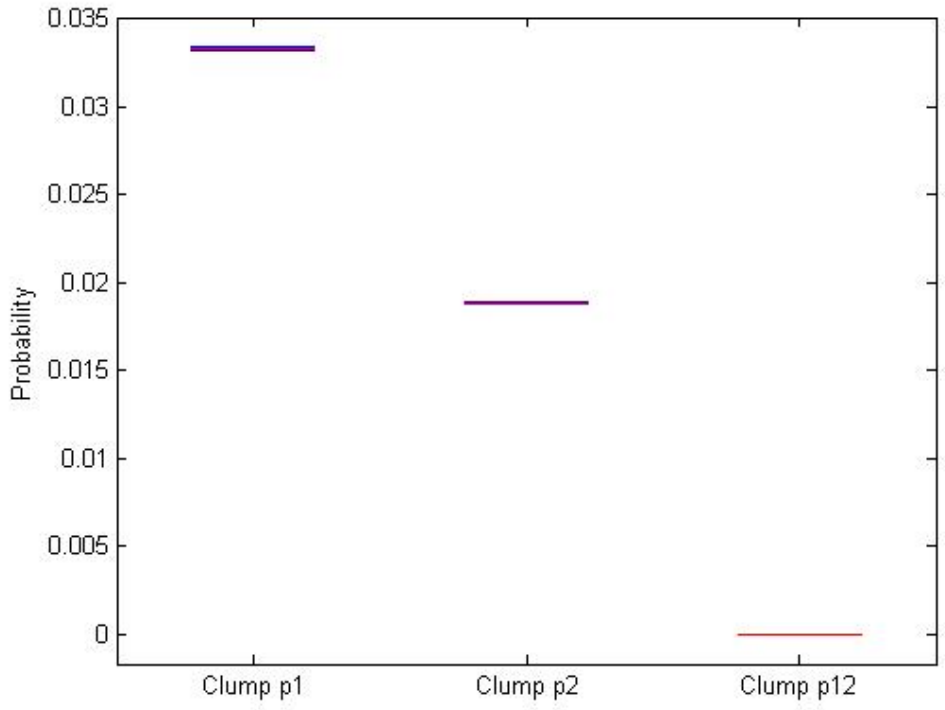

Figure 3-6: 95% Confidence Interval Based on the Clump Partition

Table 3-5 provides a great deal of insight into the relationship structure of individuals in the friendship network. While the Clump partition had the largest relativity weight, the "Year Joined" partition had the second largest relativity weight of .1781. This suggests that whether or not individuals joined the network before 1995 greatly explains with whom they are friends. The third largest relativity weight is held by the "Occupation Type" partition (.1207), suggesting that individuals with similar occupations will likely be friends, a commonplace occurrence in any working environment.

The largest probability of dyad occurrence is located in the "Criminal Background" partition. Partitioning the network based on whether or not an individual has a criminal background, an estimated 5.53% probability of dyad occurrence exists for

individuals with a criminal background to be friends. Similar to the Clump partition, there is no estimated probability of dyads forming between levels in the "School Type" partition. This suggests that individuals attending a Madrassa (Muslim school) are not likely to make friends with individuals attending either Christian or secular schools. With the known rift between Muslim and Christian fundamentals, this certainly makes intuitive sense.

Using the weights and dyad probability matrices of the significant partitions, an overall weighted dyad probability matrix can be constructed by

$$\hat{P} = \sum_{m} w_m \left(\text{Dyad Probability Matrix } m \right) \tag{3.25}$$

where the summation is conducted over all m statistically significant partitions. $\hat{P}$ is a 100 x 100 matrix for this example and, due to size, is not displayed. Despite this issue, the PRA is used to summarize $\hat{P}$. Table 3-6 shows the results of this method. Again, a rank ordered priority target list is produced.

Table 3-6: PRA Results for the Friendship Network

PRA Score	Node(s)	
0.074	94, 95, 96	**Least Connected Nodes**
0.076	52	
0.077	45, 55, 56, 69	
0.078	37, 71, 75, 77, 82, 88, 92, 93, 97	
0.079	20, 43, 46, 48, 70, 98	
0.080	47, 84, 87	
0.082	22, 34, 89	
0.083	9	
0.084	100	
0.088	99	
0.091	91	
0.092	7, 38, 58	
0.094	16, 19	
0.096	85	
0.099	51	
0.100	36	
0.101	44	
0.103	74, 78, 80, 81, 83	
0.104	54, 73	
0.105	28, 53, 59	
0.106	18, 23, 42, 64, 68	
0.107	5	
0.108	41, 62, 72, 76, 90	
0.109	50, 57, 86	
0.110	39, 40, 49, 60, 61, 67	
0.111	15, 65, 66, 79	
0.112	4, 63	
0.113	8, 24, 25, 29	
0.114	14, 26	
0.115	1	
0.116	3	
0.117	2, 6, 10, 11, 12, 13, 17, 21, 27, 30, 31, 32, 33, 35	**Most Connected Nodes**

Node one corresponds to Osama bin Laden and node two corresponds to Ayman al-Zawahiri, bin Laden's personal doctor. The PRA list, based on the weighted dyad probability matrix constructed from the findings of the hypothesis test, shows that both men are highly ranked and nodes of interest. According to the results, al-Zawahiri is the more valued node. While neutralizing bin Laden may strike a major blow to al Qaeda,

al-Zawahiri has more inter-relations which may be useful in aiding in the capture of other operatives, possibly to include bin Laden. By considering the assumption that higher ranked nodes might be more heavily protected, the PRA results provide a gauge of the "reachability" of nodes. In this case "reachability" refers to the number of potential paths to the person of interest.

According to element [1, 2] of $\hat{P}$, representing the estimated probability of dyad occurrence between Osama bin Laden and Ayman al-Zawahiri, the dyad probability estimate is .0256. This is the same value achieved by taking the weighted sum in (3.24) based on the results of Table 3-5. While it is known that a tie exists between bin Laden and al-Zawahiri, the hypothesis testing framework suggests a .0256 estimated probability of dyad existence based upon network partitions. Although this seems like a small probability, the largest weighted probability generated by $\hat{P}$ is .026. Clearly, the estimated weighted dyad probabilities of $\hat{P}$ are small, but still provide a proxy strength-of-tie measure. In light of this fact, the .0256 estimated probability of a tie between bin Laden and al-Zawahiri is one of the larger dyads estimated by $\hat{P}$. Scaling the bin Laden / al-Zawahiri dyad is accomplished through division by the largest dyad probability and results in a 98.46% chance of dyad formation based on the results of $\hat{P}$.

The highest estimated dyad probabilities (.026) correspond to the links between each node of set {2, 10, 12, 17, 32, 33}. Notice that all six nodes are identified by the PRA as being in the set containing the most connected nodes. In addition, Ayman al-Zawahiri is a member of this group, once again giving credence to the possibility that he is more highly connected than Osama bin Laden.

3.6. Summary

This chapter laid out a hypothesis testing framework capable of examining an observed adjacency matrix for variability. Three different examples were provided, showing the use and ability of the hypothesis test when applied to both simulated and real-world networks. When using open source data for the Al Qaeda terrorist network, the hypothesis test identified the "Arab Clump" partition as best explaining the formation of friendship ties while identifying the "Married", "Socio-Economic Status", and "Religious Background" partitions as not explaining the formation of friendship ties. With insights like this yielded by the hypothesis testing framework, social network analysts are better able to analyze network members. A potential benefit of this analysis is a reduction in the amount of data being explored by the social network analysts as the hypothesis testing framework causes the salient partitions to rise to the surface. These results must be viewed in the light of the analysis being conducted.

4.1. Introduction

As constructed, the hypothesis test evaluates observed networks to determine if attribute partitions are capable of explaining the variability in the observed adjacency matrix. In order to measure the full strength of the hypothesis test, it must be evaluated for accuracy. The best way to evaluate accuracy is to investigate the type I and type II errors of the hypothesis test. To this end, a test network was constructed where the underlying attribute partitions were formed arbitrarily and level and cross-level dyad parameters were assigned. Based on the dyad probabilities of the constructed network, simulated networks were generated for hypothesis test evaluation. Since the partitions and corresponding parameters were arbitrarily assigned, the partitions truly explaining observed adjacency matrix variability are known. Given this knowledge, the hypothesis test was evaluated to see if it could correctly identify partitions explaining adjacency matrix variability.

4.2. Constructing the Test Network

For the purpose of evaluation, the partitions explaining adjacency matrix variability were specified before evaluation of the proposed methodology took place. As such, the correct results were known. That is, if the methodology correctly identified the attribute partitions explaining adjacency matrix variability, it was known. It was also known if the methodology failed to identify the attribute partitions explaining adjacency matrix variability.

With the partitions explaining adjacency matrix variability known before evaluation began, the level and cross-level parameters were arbitrarily assigned. With this information specified, the "True" Dyad Probability Matrix (TDPM) was constructed where the elements of the TDPM represent the actual probability of dyad occurrence based on the known partition assignment and corresponding parameters. It was this TDPM which was used to generate random networks for the evaluation of the proposed methodology.

4.3. Underlying Structure of the Test Network

To evaluate the proposed methodology, $n = 20$ network nodes were arbitrarily assigned to two $k = 2$ level partitions. The level assignments for partition one were: Level 1 ≡ {1 – 4, 14 – 16} and Level 2 ≡ {5 – 13, 17 – 20}. The level assignments for partition two were: Level 1 ≡ {1 – 2, 7 – 20} and Level 2 ≡ {3 – 6}.

These levels are simple to represent in column form. If a node is assigned to level one, it is given a value of zero, while if a node is assigned to level two, it is given a value of one. For the remainder of this evaluation, the attribute partitions are referred to as "test columns". The reasoning behind this notation is that a level is simply a subset of the nodes contained in the observed network as broken down by attribute. Each subset is stored as a column vector containing the list of nodes and their corresponding level assignment. For the evaluation of the proposed methodology, three test columns are used. Test columns one and two completely explain the variability in the observed adjacency matrix, while test column three does not. Table 4-1 details the test columns used for evaluation.

Table 4-1: Test Columns

Node	Test Column 1	Test Column 2	Test Column 3
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	1	0
6	1	1	1
7	1	0	0
8	1	0	1
9	1	0	0
10	1	0	1
11	1	0	0
12	1	0	1
13	1	0	0
14	0	0	1
15	0	0	0
16	0	0	1
17	1	0	0
18	1	0	1
19	1	0	0
20	1	0	1

4.4. Evaluating the Type I and Type II Errors of the Hypothesis Test

Monte Carlo simulation was used to evaluate the hypothesis test through the generation of simulated adjacency matrices based on the TDPM. The test columns were used to inspect the variability of these simulated adjacency matrices, reporting whether or not the test columns explained the observed variability. As with any statistical analysis, it is desirable to control type I error, the probability of false alarm or finding that a partition explains variability in the observed adjacency matrix when, in reality, the partition does not. The issue that arises is that when two or more partitions are explored, type I error becomes difficult to control. The general reason for this is due to the presence of partitions which are not mutually orthogonal. In the case where partitions are mutually

orthogonal, they can be viewed as independent. Assuming independence, the experimentwide error rate, denoted α, is

$$\alpha = 1 - (1 - \alpha_0)^t \tag{4.1}$$

where t denotes the total number of test columns being evaluated and α_0 denotes the $\hat{p}$-value required for a test column to be found statistically significant. In the case of mutually orthogonal partitions, the experimentwide error rate will equal α. In the more common case where partitions are not mutually orthogonal, the experimentwide error rate will be less than or equal to α. In other words, α is the upper bound on error, ensuring that with enough repetitions, error will not exceed this specified threshold (Wu and Hamada, 2000: 133).

To find an "across the board" statistical significance threshold for each column being tested, (4.1) must be solved for α_0:

$$\alpha_0 = 1 - \exp\left(\frac{\log(1-\alpha)}{t}\right) \tag{4.2}$$

In this evaluation of the hypothesis testing framework, α = .05 and t = 3. Solving (4.2) for α_0 yields a required statistical significance threshold of .017. A $\hat{p}$-value $\leq$ .017 will ensure that the type I error rate is less than or equal to .05.

The evaluation centered on whether or not the hypothesis test detected network structure both when it was and was not present. That is, how often did the hypothesis test fail to detect the presence of structure when no structure was truly present? Similarly, how often did the hypothesis test detect structure when some level of structure truly existed? Finding no structure when none is present is desirable, while finding structure when none is present is type I error. It should be mentioned that the estimated type I error will vary with each simulation run conducted, but, over a sufficient number of runs, the estimated error will approach, if not drop below, the specified error rate. With this in mind, one thousand simulation runs were conducted.

Finding structure when it is truly present is desirable, while failing to find structure when it is truly present is type II error. As with type I error, it is desirable to control type II error. In fact, from the estimated type II error, the "power" of the hypothesis test can be estimated (Montgomery, 2001: 34). Power calculation is: 1 – Prob(Type II Error). The higher the estimated power of the hypothesis test, the more reliable the test is.

There are two aspects of type II error evaluation which are explored. The first aspect is when the magnitude of difference between p_1, p_2, and p_{12} is large, that is, $p_2 >> p_1$. The second aspect is when the magnitude of difference between p_1, p_2, and p_{12} is small, that is, $p_1 > p_2$. In both aspects, the probability of type II error was explored, as well as the probability of identifying all test columns responsible for structure and identifying at least one test column responsible for structure. It would stand to reason that the greater the magnitude of difference between partition parameters, the better the evaluation will perform.

4.5. Verifying Hypothesis Test Accuracy

In the case where no network structure was present, all dyads formed with a probability of .2, thus $p_0 = .2$. From the results of the simulated experiments, the estimated type I error is .0153, well below the upper bound of .05. As such, the estimated probability of correct detection is .9847. This shows that the hypothesis test performs well by failing to reject the null hypothesis when no structure is present.

Table 4-2 shows the case when structure is present in the network and the magnitude of the difference between partition parameters is large.

Table 4-2: Structure Present, Large Magnitude Between Partition Parameters

	p_1	p_2	$p_{1,2}$
Partition 1	0.25	0.75	0.10
Partition 2	0.25	0.75	0.10

With these parameter values, the estimated probability of detecting structure when structure is present is one. That is, the estimated power of the test is one and type II error is estimated to be zero. The probability that all truly significant test columns are deemed significant for each simulation run is one, as is the probability that at least one truly significant test column is deemed significant for each simulation run. These are good results for the hypothesis test, and show that, when applied to level parameters with a large magnitude of difference, the hypothesis testing framework performs accurately.

Table 4-3 shows the case when structure is present in the network and the magnitude of the difference between partition parameters is small.

Table 4-3: Structure Present, Small Magnitude Between Partition Parameters

	p_1	p_2	$p_{1,2}$
Partition 1	0.80	0.75	0.30
Partition 2	0.80	0.75	0.30

With these parameter values, the estimated probability of detecting structure when structure is present is .884. With the estimated power of the hypothesis being .884, the estimated type II error is .116. The probability that all truly significant test columns are deemed significant for each simulation run is .994, and the probability that at least one truly significant test column is deemed significant for each simulation run is one. These are good results for the hypothesis test, and show that, when applied to level parameters with a small magnitude of difference, the hypothesis testing framework performs accurately.

4.6. Summary

Based on the results of section 4.5, it can be safely said that the hypothesis test, more often than not, correctly rejects the null hypothesis when structure is present and fails to reject the null hypothesis when structure is not present. In the cases where structure was present, the hypothesis testing framework accurately identified structure, although, as would be expected, it identified structure better in the case where there was a large magnitude of difference between partition parameters.

The result of this evaluation is the verification of the accuracy of a robust hypothesis test able to determine which partitions explain observed adjacency matrix variability. With these partitions known, social network analysts can better focus their time and resources towards the partitions which will yield the greatest amount of insight

into social networks. As more information on clandestine social networks is uncovered, the proposed methodology provides a gauge of both the quality and use of new information, ensuring analyst efforts achieve the greatest possible success.

5.1. Introduction

This thesis proposed a tractable, statistically sound method with which to test observed clandestine social networks for non-random structure by partitioning the network based on nodal attributes. Like many other statistical applications, this method employs hypothesis tests, seeking to discover if the observed network variability is explained by the underlying partitions. Partitions appearing to explain network variability were further investigated using social network analysis tools.

5.2. Methodology

The methodology introduced both a binomial probability mass function and likelihood function parameterized based on network variables. Proposing an alternative hypothesis, where network structure was explained by the partitions of the observed adjacency matrix, against a null hypothesis, where network structure was not explained by the partitions, a likelihood statistic was found for both hypotheses. Though comparing the two, a ratio statistic was found and used as the test statistic. By comparison to a set statistical significance threshold, the partition under investigation was evaluated to see if it explained non-random structure observed in the network.

5.3. Results

In order to test the network model, a "truth" network was constructed where the partitions influencing network formation were known, allowing the ability to gauge the quality of model results. Through the use of a confusion matrix, along with various other

metrics, the model was evaluated for proper operation. Though false positives, or type I error, did occur during some tests, the rate at which they occurred was acceptable based on the error threshold specifications of the evaluation. These results showed the network model to be well constructed and capable of identifying partitions able to identify and explain the presence of non-random structure.

5.4. Future Efforts

As mentioned in section 5.3, false positives occurred, resulting in finding an attribute responsible for explaining network structure when it actually is not. While this error is excusable, reducing or even eliminating the number of false positives is still desirable. Future studies using the network model should focus on the reduction of false positives.

The issue of missing data can be approached in several ways (Nysether, 2007: 2–19). While this effort implemented an assumption to fill gaps in the Sageman data, other options are at the analyst's disposal. For instance, a subject matter expert (SME) could be consulted to gain knowledge of the individual under investigation. Since SME knowledge comes from both outside knowledge and operational experience, it often provides the best way with which to fill in missing data. Assuming that the knowledge is sound, it can be incorporated into existing datasets. In the cases where SME advice is just an educated guess, a quantifiable method must be employed in which to gauge the quality of the estimate. To this end, it is suggested that a method such as linear regression be leveraged to fill in missing data. With the results from such a method on hand, SMEs can be consulted to verify that the results of the regression are both

reasonable and realistic. Using linear regression and SME knowledge in tandem will couple a statistically sound method with real-world knowledge.

Another area in which to expand research is in dealing with levels containing only one node. Due to the scope of this study, these cases were neglected. As proposed, the network model is more than capable of dealing with these cases, but not enough time was available to explore them.

Potential future research lies in using $\hat{P}$, the weighted dyad probability matrix. Since $\hat{P}$ is constructed based on the partitions explaining observed network structure, it details the probability of the existence of both observed and unobserved arcs. With these probabilities calculated, both the dyads with high probability of formation and the dyads with low probability of formation are known. This provides two more ways in which to investigate the observed network. One way is to explore the dyads which are known to exist but have a low probability of existence. This has a direct application to small-world theory which proposes that everyone is loosely connected with everyone else in the world. If an arc is known to exist between nodes, but the probability of that arc existing is minimal, those two nodes must have something in common which caused the tie to form but has not yet been revealed. What is the cause and how can it be leveraged to disrupt the network? It is these kinds of answers that the analyst seeks to discover.

The other application is to investigate the arcs which have a high probability of existing but are not observed. Is this due to operational security (OPSEC) and military deception (MILDEC) measures? Perhaps something occurred to drive a wedge between these two nodes. When the reasons for not observing these dyads are known, more insight is gained into the network, resulting in improved network analysis. The

computational and visual aspects of programs like UCINET and AGNA are perfect for exploring $\hat{P}$. By comparing the observed network to the one generated from whatever aspect of $\hat{P}$ is being investigated, it may perhaps be possible to see through the OPSEC and MILDEC measures employed in the operation of the clandestine network.

As specified, the adjacency matrix used to test partitions for network structure contained undirected arcs. As a result, the dyad probability matrices corresponding to each attribute are symmetric because $p_{ij} = p_{ji}$. The limitation of this representation is that dyads formed between nodes are assumed to have the same "weight" or authority. That is, an undirected arc does not take into account rank or influence one person (node) has over another. To incorporate this sort of relationship, Bayesian networks can be employed to show direction of dyads based on conditional probabilities between the two nodes. In addition, employing a weighted Bayesian network would incorporate both direction of dyads and the authority one node has over another. In either case, the resulting dyad probability matrices might no longer be symmetric because now it is possible that $p_{ij} \neq p_{ji}$. In the case of working with a weighted Bayesian network, both elements p_{ij} and p_{ji} must be investigated simultaneously to see which node is "higher ranked". The element with the higher probability of dyad occurrence might represent the higher ranked individual as the originating node. The cases where $p_{ij} = p_{ji}$ might represent peers because the two nodes have the same probability of dyad occurrence due to having the same amount of influence.

The alternative hypothesis used for this study was that at least one attribute contributed to network formation. Although this was the only hypothesis explored, the

alternative hypothesis is not limited to that form. Another such form of the alternative hypothesis is

$$H_A : p_1 < p_2 < p_3 < \ldots < p_h < p_{1,2} < p_{1,3} < \ldots < p_{ij} \tag{5.1}$$

Notice that the form of the alternative hypothesis in (5.1) specifies strength of dyad probability in respect to the other probabilities. Nothing mandates that the probabilities be in ascending or descending level or cross-level order, this was simply done for the sake of illustration to generalize this form for the alternative hypothesis. Another form of the null hypothesis could be to propose that p_h or p_{ij} are equal to some probability while the alternative hypothesis tests if this is truly the case. Either way, the network model is capable of handling such tests.

It may also be possible to use the network model to find out information about an observed network when nothing else is known other than the partition level each node is assigned to. To investigate this aspect, the testing of an observed network would have to be conducted where the correct results are known prior to testing.

For this research, dyads were assumed to be conditionally independent based on the partition. Certainly, factors exist in every network which influence dyad formation. The problem that arose, motivating this assumption, was that finding an overall generalized expression to sum this up was beyond the scope of this research. Work has been done in the field of correlated Bernoulli trials (Vireos, Balasubramanian, and Balakrishnan), but such studies were conducted with time as the main factor. That is not the case for social networks, where the ties between players are the main factors. If a

method were developed which incorporated variable correlation into the network model, many new doors would be opened for the methodology detailed in this study. One possible aspect would be the incorporation of other probability distributions, thus allowing the model to be expanded past the current bounds of the binomial distribution.

5.5. Conclusion

While this study focused on the clandestine social networks of terrorist organizations, the methodology proposed herein is applicable to many other social networks. For instance, with the increasing popularity of websites like www.myspace.com and programs used for "instant messaging", an abundant amount of data presents itself, as individuals list their "friends" or "contacts". From these lists, social networks can be constructed. Depending on how these networks are built, the strength of the ties between members can be incorporated. By examining a network constructed of weak ties, small-world theory is incorporated. It has been suggested that small-world theory is possibly the best representation of both group affiliation and connectivity. Since the model only tests the specified network inputted, the decision of what kind of network to both construct and investigate is left up to the analyst. Because of this, the analyst is responsible for the interpretation of the results yielded by the network model.

The major benefit of this model is reducing the amount of data required to investigate. As the level of data increases, the ability of analysts to separate the "wheat from the chaff" decreases. The unforeseen result of this is sometimes referred to as "data paralysis". For this reason, the network model presented here was developed and

demonstrated. As the quality of clandestine social network analysis increases, so does the analyst's ability to gain insight into the workings of terrorist organizations. With this improved insight, the analyst can present findings to decision makers, allowing them to make better and more informed decisions. As the quality of peacetime and wartime decisions increases, so will the ability of the United States to execute the Global War on Terror.

6. Appendix A: Supplemental Material

6.1. Maximum Likelihood Estimates for p_h and p_{ij} for the k = 3 Level Partition

$$l\left(\mathbf{p}\,|\,\mathbf{d}, z(k)\right) = \sum_{h=1}^{3}\left[d_h \log\left(p_h\right) + \left(N_h - d_h\right)\log(1-p_h)\right] + \sum_{i=1}^{2}\sum_{j=i+1}^{3}\left[d_{ij}\log\left(p_{ij}\right) + \left(N_{ij} - d_{ij}\right)\log(1-p_{ij})\right] \tag{6.1}$$

Expanding (6.1) out yields:

$$\begin{aligned} l\left(\mathbf{p}\,|\,\mathbf{d}, z(k)\right) = {} & d_1\log(p_1) + (N_1 - d_1)\log(1-p_1) + d_2\log(p_2) \\ & + (N_2 - d_2)\log(1-p_2) + d_3\log(p_3) + (N_3 - d_3)\log(1-p_3) \\ & + d_{12}\log(p_{12}) + (N_{12} - d_{12})\log(1-p_{12}) + d_{13}\log(p_{13}) \\ & + (N_{13} - d_{13})\log(1-p_{13}) + d_{23}\log(p_{23}) + (N_{23} - d_{23})\log(1-p_{23}) \end{aligned} \tag{6.2}$$

The next step is to take the derivative of (6.2) with respect to the probability parameter being estimated. If p_1 was the desired parameter, the derivative would be taken with respect to p_1. That is:

$$\frac{\partial l\left(\mathbf{p}\,|\,\mathbf{d}, z(k)\right)}{\partial p_1} = \frac{d_1}{p_1} + \frac{N_1 - d_1}{1-p_1} \tag{6.3}$$

Setting (6.3) equal to zero and solving for p_1 yields:

$$\hat{p}_1 = \frac{d_1}{N_1} \tag{6.4}$$

Note that in (6.4), the hat denotes parameter estimation. In general, the estimation of p_h is:

$$\hat{p}_h = \frac{d_h}{N_h} \tag{6.5}$$

To estimate cross-level probabilities, the derivative of (6.2) is taken with respect to the cross-level probability being estimated. If p_{12} was the desired parameter, the derivative would be taken with respect to p_{12}. That is:

$$\frac{\partial l\left(\mathbf{p} \mid \mathbf{d}, z\left(k\right)\right)}{\partial p_{12}} = \frac{d_{12}}{p_{12}} + \frac{N_{12} - d_{12}}{1 - p_{12}} \tag{6.6}$$

Setting (6.6) equal to zero and solving for p_{12} yields:

$$\hat{p}_{12} = \frac{d_{12}}{N_{12}} \tag{6.7}$$

Again, the hat in (6.7) denotes parameter estimation. In general, the estimation of p_{ij} is:

$$\hat{p}_{ij} = \frac{d_{ij}}{N_{ij}} \tag{6.8}$$

6.2. Reducing L_1 to L_0 for the $k = 3$ Level Partition

The necessary variables for the $k = 3$ case are:

- Level Variables: $d_1, d_2, d_3, p_1, p_2, p_3$
- Cross-Level Variables: $d_{12}, d_{13}, d_{23}, p_{12}, p_{13}, p_{23}$

L_1 is constructed using the above variables:

$$\begin{aligned} L_1 = {} & p_1^{d_1}(1-p_1)^{\binom{n_1}{2}-d_1} \bullet p_2^{d_2}(1-p_2)^{\binom{n_2}{2}-d_2} \bullet p_3^{d_3}(1-p_3)^{\binom{n_3}{2}-d_3} \\ & \bullet p_{12}^{d_{12}}(1-p_{12})^{\binom{n_{12}}{2}-\binom{n_1}{2}-\binom{n_2}{2}-d_{12}} \bullet p_{13}^{d_{13}}(1-p_{13})^{\binom{n_{13}}{2}-\binom{n_1}{2}-\binom{n_3}{2}-d_{13}} \\ & \bullet p_{23}^{d_{23}}(1-p_{23})^{\binom{n_{23}}{2}-\binom{n_2}{2}-\binom{n_3}{2}-d_{23}} \end{aligned} \tag{6.9}$$

Substituting $p_1 = p_2 = p_3 = p_{12} = p_{13} = p_{23} = p_0$ into (6.9):

$$\begin{aligned} L_1 = {} & p_0^{d_1}(1-p_0)^{\binom{n_1}{2}-d_1} \bullet p_0^{d_2}(1-p_0)^{\binom{n_2}{2}-d_2} \bullet p_0^{d_3}(1-p_0)^{\binom{n_3}{2}-d_3} \\ & \bullet p_0^{d_{12}}(1-p_0)^{\binom{n_{12}}{2}-\binom{n_1}{2}-\binom{n_2}{2}-d_{12}} \bullet p_0^{d_{13}}(1-p_0)^{\binom{n_{13}}{2}-\binom{n_1}{2}-\binom{n_3}{2}-d_{13}} \\ & \bullet p_0^{d_{23}}(1-p_0)^{\binom{n_{23}}{2}-\binom{n_2}{2}-\binom{n_3}{2}-d_{23}} \end{aligned} \tag{6.10}$$

Combining terms simplifies (6.10) down to:

$$L_1 = p_0^{(d_1+d_2+d_3+d_{12}+d_{13}+d_{23})} \bullet (1-p_0)^{\left[\binom{n_1}{2}+\binom{n_2}{2}+\binom{n_3}{2}+\binom{n_{12}}{2}-\binom{n_1}{2}-\binom{n_2}{2}+\binom{n_{13}}{2}-\binom{n_1}{2}-\binom{n_3}{2}+\binom{n_{23}}{2}-\binom{n_2}{2}-\binom{n_3}{2}-d_1-d_2-d_3-d_{12}-d_{13}-d_{23}\right]} \tag{6.11}$$

Using (3.2) simplifies (6.11) to:

$$L_1 = p_0^{\,d}(1-p_0)^{\left[\binom{n_1}{2}+\binom{n_2}{2}+\binom{n_3}{2}+\binom{n_{12}}{2}-\binom{n_1}{2}-\binom{n_2}{2}+\binom{n_{13}}{2}-\binom{n_1}{2}-\binom{n_3}{2}+\binom{n_{23}}{2}-\binom{n_2}{2}-\binom{n_3}{2}-d\right]} \tag{6.12}$$

Eliminating opposite choose operations reduces (6.12) to:

$$L_1 = p_0^{\,d}(1-p_0)^{\left[\binom{n_{12}}{2}+\binom{n_{13}}{2}+\binom{n_{23}}{2}-\binom{n_1}{2}-\binom{n_2}{2}-\binom{n_3}{2}-d\right]} \tag{6.13}$$

Though inserting summations, (6.13) becomes:

$$L_1 = p_0^{\,d}(1-p_0)^{\left[\sum_{i=1}^{2}\sum_{j=i+1}^{3}\binom{n_{ij}}{2}-\sum_{h=1}^{3}\binom{n_h}{2}-d\right]} \tag{6.14}$$

The summations of (6.14) could also be written as:

$$\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\binom{n_{ij}}{2}-(2-k)\sum_{h=1}^{k}\binom{n_h}{2}=\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\binom{n_{ij}}{2}+\sum_{h=1}^{k}\binom{n_h}{2}=\binom{n}{2}=N \tag{6.15}$$

Applying (6.15) to (6.14) yields:

$$L_1 = p_0^{\,d}(1-p_0)^{[N-d]} = L_0 \tag{6.16}$$

While this example was for the $k = 3$ case, it generalizes to all values of k.

Bibliography

Anderson, Carolyn J., Stanley Wasserman, and Bradley Crouch. "A p* Primer: Logit Models for Social Networks", *Social Networks*, 24: 37 – 66 (January 2002).

Ashar, Hanna and Jonathan Z. Shapiro. "Measuring Centrality: A Note on Hackman's Resource-Allocation Theory," *Administrative Science Quarterly*, 33: 275 – 283 (June 1988).

Bonacich, Phillip. "Factoring and Weighting Approaches to Status Scores and Clique Identification," *Journal of Mathematical Sociology*, 2: 113 – 120 (June 1972).

Borgatti, Stephen P. "Centrality and Network Flow," *Social Networks*, 27: 55 – 71 (January 2005).

Borgatti, Stephen P., Kathleen M. Carley, and David Krackhardt. "On the Robustness of Centrality Measures Under Conditions of Imperfect Data," *Social Networks*, 28: 124 – 136 (May 2006).

Buchanan, Mark. *Nexus: Small Worlds and the Groundbreaking Science of Networks*. New York: W.W. Norton & Company, 2002.

Cho, Catherine, Sooyoung Kim, Jaewook Lee, and Dae-Won Lee. "A Tandem Clustering Process for Multimodal Datasets," *European Journal of Operational Research*, 168: 998 – 1008 (February 2006).

Casella, George and Roger L. Berger. *Statistical Inference* (2nd Edition). Pacific Grove, CA: Duxbury, 2002.

Clark, Clinton R. *Modeling and Analysis of Clandestine Networks*. MS Thesis, AFIT/GOR/ENS/05-04. School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2005.

Costenbader, Elizabeth and Thomas W. Valente. "The Stability of Centrality Measures when Networks are Sampled," *Social Networks*, 25: 283 – 307 (October 2003).

Dillon, William R. and Matthew Goldstein. *Multivariate Analysis: Methods and Applications*. New York: John Wiley & Sons, Inc., 1984.

Dunham, Margaret H. *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall, 2003.

Faust, Katherine, Karin E. Willert, David D. Rowlee, and John Skvoretz. "Scaling and Statistical Models for Affiliation Networks: Patterns of Participation Among Soviet Politicians During the Brezhnev Era," *Social Networks*, 24: 231 – 259 (July 2002).

Fienberg, Stephen E. and Stanley S. Wasserman. "Categorical Data Analysis of Single Sociometric Relations," *Sociological Methodology*, 1: 156 – 192 (January 1981).

Fienberg, Stephen E., Michael M. Meyer, and Stanley S. Wasserman. "Statistical Analysis of Multiple Sociometric Relations," *Journal of the American Statistical Association*, 80: 51 – 67 (March 1985).

Frank, Ove and David Strauss. "Markov Graphs," *Journal of the American Statistical Association*, 81: 832 – 842 (September 1986).

Freeman, Linton C. "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, 1: 215 – 239 (January 1979).

Gamerman, Dani and Hedibert F. Lopes. *Markov Chain Monte Carlo* (2nd Edition). Boca Raton: Chapman & Hall/CRC, 2006.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman & Hall/CRC, 1996.

Giudici, Paolo and Robert Castelo. "Improving Markov Chain Monte Carlo Model Search for Data Mining," *Machine Learning*, 50: 127 – 158 (January 2003).

Gómez, Daniel, Enrique González-Arangüena, Manuel Conrado, Guillermo Owen, Mónica del Pozo, and Juan Tejada. "Centrality and Power in Social Networks: A Game Theoretic Approach," *Mathematical Social Sciences*, 46: 27 – 54 (August 2003).

Handcock, Mark S. "Statistical Models for Social Networks: Inference and Degeneracy," *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. 229 – 240. Washington, DC: The National Academies Press, 2003.

Holland, Paul W. and Samuel Leinhardt. "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association*, 76: 33 – 65 (March 1981).

Jain, A.K., M.N. Murty, and P.J. Flynn. "Data Clustering: A Review," *ACM Computing Surveys*, 31: 264 – 323 (September 1999).

Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An Efficient k-Means Clustering Algorithm: Analysis and Implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 881 – 892. New York: IEEE Press, 2002.

Kau, Lih-Jen. "Adaptive Predictor with Dynamic Fuzzy k-Means Clustering for Lossless Image Coding," *The 12th IEEE International Conference on Fuzzy Systems*. 944 – 949. New York: IEEE Press, 2003.

Krause, Eugene F. *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. New York: Dover Publications, Inc., 1986.

Kulkarni, Vidyadhar G. *Modeling and Analysis of Stochastic Systems*. New York: Chapman & Hall, 1995.

Li, Jianyu, Rui Lv, Zhanxin Yang, Shuzhong Yang, Hongwei Mo, and Xianglin Huang. "Construction of Small World Networks Based on k-Means Clustering Analysis," *Computational Science – ICCS 2006: 6th International Conference Proceedings*. 997 – 1000. New York: Springer, 2006.

Liben-Nowell, David and Jon Kleinberg. "The Link Prediction Problem for Social Networks," *Proceedings of the 12th International Conference on Information and Knowledge Management*. 556 – 559. New York: ACM Press, 2003.

Manly, Bryan F.J. *Multivariate Statistical Methods: A Primer*. Boca Raton: Chapman & Hall/CRC, 2005.

Marsden, Peter V. "Egocentric and Sociocentric Measures of Network Centrality," *Social Networks*, 24: 407 – 422 (October 2002).

Monge, Peter R. and Noshir S. Contractor. *Theories of Communication Networks*. New York: Oxford University Press, 2003.

Montgomery, Douglas C. *Design and Analysis of Experiments* (5th Edition). New York: John Wiley & Sons, Inc., 2001.

Montgomery, Douglas C, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis* (3rd Edition). New York: John Wiley & Sons, Inc., 2001.

Moy, Gary K. *A Specific Network Link and Path Likelihood Prediction Tool*. MS Thesis, AFIT/GCS/ENG/96D-21. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1996 (ADA325184).

Nysether, Nathan E. *Classifying Failing States*. MS Thesis, AFIT/GOR/ENS/07-19. School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2007.

Ove, Frank. "Using Centrality Modeling in Network Surveys," *Social Networks*, 24: 385 – 394 (October 2002).

Page, Lawrence and Sergey Brin. "Google searches more sites more quickly, delivering the most relevant results." n. pag. http://www.google.com/technology/. 25 November 06.

Pattison, Philippa and Stanley Wasserman. "Logit Models and Logistic Regressions for Social Networks: II. Multivariate Relations," *British Journal of Mathematical and Statistical Psychology*, 52: 169 – 193 (November 1999).

Pattison, Philippa, Stanley Wasserman, Garry Robins, and Alaina Michaelson Kanfer. "Statistical Evaluation of Algebraic Constraints for Social Networks," *Journal of Mathematical Psychology*, 44: 536 – 568 (December 2000).

Ressler, Steve. "Social Network Analysis as an Approach to Combat Terrorism: Past, Present, and Future Research," *Homeland Security Affairs*, 2: n. pag. (July 2006).

Robins, Garry, Philippa Pattison, and Stanley Wasserman. "Logit Models and Logistic Regressions for Social Networks: III. Valued Relations," *Psychometrika*, 64: 371 – 394 (September 1999).

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. "A Workshop on Exponential Random Graph (p*) Models for Social Networks." Workshop Given to the Social Networks and Complexity Seminar. Kennedy School of Government, Harvard University, Cambridge, MA, 17 August 2005.

Sageman, Marc. *Understanding Terrorist Networks*. Philadelphia: University of Philadelphia Press, 2004.

Sageman, Marc. "E-Notes: Understanding Terror Networks (1 November 2004)." Excerpt from unpublished article. n. pag. http://www.fpri.org/enotes/20041101. middleeast.sageman.understandingterrornetworks.html. 23 October 2006.

Schweinberger, Michael and Tom A. B. Snijders. "Settings in Social Networks: A Measurement Model," *Sociological Methodology*, 33: 307 – 341 (January 2003).

Serban, Nicoleta and Larry Wasserman. "CATS: Clustering After Transformation and Smoothing," *Journal of the American Statistical Association*, 100: 990 – 999 (September 2005).

Spall, James C. "Estimation via Markov Chain Monte Carlo," *IEEE Control Systems Magazine*, 23: 34 – 45 (April 2003).

Slater, Peter J. "A Survey of Sequences of Central Subgraphs," *Networks*, 34: 244 – 249 (November 1999).

Snijders, Tom A.B. "Explained Variation in Dynamic Network Models," *Mathematiques et Science Humanes / Mathematics and Social Science*, 42: 31 – 41 (December 2004).

Van de Bunt, Gerhard G., Marjitje A.J. Van Duijn, and Tom A.B. Snijders. "Friendship Networks Through Time: An Actor-Oriented Dynamic Statistical Network Model," *Computational and Mathematical Organization Theory*, 5: 167 – 192 (July 1999).

Viveros, Roman, K. Balasubramanian, and N. Balakrishnan. "Binomial and Negative Binomial Analogues Under Correlated Bernoulli Trials," *The American Statistician*, 48: 243 – 247 (August 1984).

Wackerly, Dennis D., William Mendenhall III, and Richard L Scheaffer. *Mathematical Statistics with Applications* (6th Edition). Pacific Grove, CA: Duxbury, 2002.

Walker, Michael E. and others. "Statistical Models for Social Support Networks," in *Advances in the Social and Behavioral Sciences: Research in the Social and Behavioral Sciences*. Ed. Stanley Wasserman and Joseph Galaskiewicz. Thousand Oaks, CA: Sage Publications, 1994.

Wang, Yuchung J. and George Y. Wong. "Stochastic Blockmodels for Directed Graphs," *Journal of the American Statistical Association*, 82: 8 – 19 (March 1987).

Wasserman, Stanley and Sheila O'Leary Weaver. "Statistical Analysis of Binary Relational Data: Parameter Estimation," *Journal of Mathematical Psychology*, 29: 406 – 427 (December 1985).

Wasserman, Stanley. "Conformity of Two Sociometric Relations," *Psychometrika*, 52: 3 – 18 (March 1987).

Wasserman, Stanley and Dawn Iacobucci. "Sequential Social Network Data," *Psychometrika*, 53: 261 – 282 (June 1988).

Wasserman, Stanley and Katherine Faust. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press, 1994.

Wasserman, Stanley and Philippa Pattison. "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p*," *Psychometrika*, 61: 401 – 425 (September 1996).

Wikipedia. "Inclusion-Exclusion Principle." n. pag. http://en.wikipedia.org/wiki/Inclusion-exclusion_principle. 14 December 2006.

Wikipedia. "Instant Messaging." n. pag. http://en.wikipedia.org/wiki/Instant_messaging. 3 February 2007.

Wikipedia. "Maximum likelihood." n. pag. http://en.wikipedia.org/wiki/Maximum_likelihood. 12 January 2007.

Wikipedia. "MySpace." n. pag. http://en.wikipedia.org/wiki/MySpace. 3 February 2007.

Wikipedia. "Taxicab Geometry." n. pag. http://en.wikipedia.org/wiki/Taxicab_geometry. 27 November 2006.

Wu, C. F. Jeff and Michael Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: John Wiley & Sons, Inc., 2000.

Vita

Captain Joshua Seder graduated from Asbury College in Wilmore, Kentucky in May 2002 with a Bachelor of Arts in Applied Mathematics. Commissioned by the United States Air Force on 27 September 2002 and stationed at Wright-Patterson AFB, Ohio, his first assignment was to the Air Force Research Laboratory Air Vehicle's Directorate as a Financial Analyst. In January 2004, he was reassigned to the Aeronautical Systems Center's Reconnaissance Systems Wing as a small unmanned aerial vehicle (SUAV) Program Manager where he oversaw a $9.3M program. His notable accomplishments were the 2004 Reconnaissance Systems Wing's Junior Company Grade Officer of the Year award and assigning a DoD level Mission Designation Series (MDS) of RQ-11A to the "Pathfinder Raven". From August 2005 to March 2007, he attended the Air Force Institute of Technology where he was awarded a Master of Science in Operations Research. His follow-on assignment is to Air Force Special Operations Command (AFSOC) Headquarters at Hurlburt Field, Florida as a quiet professional.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
22-03-2007	**Master's Thesis**	Jun 2006 - Mar 2007

4. TITLE AND SUBTITLE
Examining Clandestine Social Networks for the Presence of Non-Random Structure

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Seder, Joshua, S., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 642
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/07-24

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/HECS
Attn: Gregory D. Sullivan, 2Lt, USAF
2689 G Street
WPAFB OH 45433-7022 DSN: 785-8015

NASIC/FCEB
Attn: Billy D. Darnell, Major, USAF
4180 Watson Way
WPAFB OH 45433-5648 DSN: 986-1023

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This thesis develops a tractable, statistically sound hypothesis testing framework for the detection, characterization, and estimation of non-random structure in clandestine social networks. Network structure is studied via an observed adjacency matrix, which is assumed to be subject to sampling variability. The vertex set of the network is partitioned into *k* mutually exclusive and collectively exhaustive subsets, based on available exogenous nodal attribute information. The proposed hypothesis testing framework is employed to statistically quantify a given partition's relativity in explaining the variability in the observed adjacency matrix relative to what can be explained by chance. As a result, valuable insight into the true structure of the network can be obtained. Those partitions that are found to be statistically significant are then used as a basis for estimating the probability that a relationship tie exists between any two vertices in the complete vertex set of the network. The proposed methodology aids in the reduction of the amount of data required for a given network, focusing analyses on those attributes that are most promising. Ample effort is given to both model demonstration and application, including an example using open-source data, illustrating the potential use for the defense community and others.

15. SUBJECT TERMS
social network analysis, hypothesis test, probability mass function, maximum likelihood estimation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Marcus B. Perry (ENS)
U	**U**	**U**	**UU**	101	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636, ext 4588; e-mail: Marcus.Perry@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18