**ABSTRACT**

KOBAN, DONALD DANIEL. Accounting for Uncertainty in Social Network Analysis through Replication. (Under the direction of Dr. Thom Hodgson).

Over the past decade, the US military has primarily been focused on counterinsurgency operations in Iraq and Afghanistan and as a result has markedly improved its understanding of irregular warfare. One of the most significant changes in the way the military understands irregular warfare is captured in the 2013 Army Strategic Planning Guidance which states, “success depends as much on understanding the social and political fabric of the situation as it does on the ability to physically dominate it”. As a result, the Army has placed great emphasis on the use of social network analysis (SNA) as a tool to aid leaders in understanding the organizational dynamics of an insurgency. The methods presented in the following paper provide a technique to quantify the certainty of SNA outputs in an environment where the structure of a social network is uncertain.

Social network analysis is based on the axiom that an individual’s position in a network determines in part the opportunities and constraints he or she encounters. Using network structure as the foundation for network analysis, it follows that an individual’s importance in a network is dependent upon how an analyst models the structure of a network. When networks are built from a collection of reliable data, calculating centrality measures is a fairly simple process that produces useful, quantitative measures for determining critical members in a network. However, when the accuracy of network data is uncertain, there is a high probability that an analyst’s representation of a network will be flawed. When centrality measures are based on an imperfect representation of a network, it becomes difficult to express the uncertainty of SNA measurements.

Current SNA techniques are limited because they do not adequately account for the uncertainty of information and they do not quantify the uncertainty associated with SNA outputs. Basing targeting recommendations on a single representation of a network with unknown structure causes SNA outputs to be heavily dependent on how accurately an analyst identifies the true network. Replication improves upon existing SNA techniques because of its ability to account for numerous possible representations of a network while placing proportionally greater emphasis on the most reliable information. Furthermore, replication enables an analyst to express the uncertainty associated with centrality-based targeting recommendations by generating a distribution of rankings for each individual.

Accounting for Uncertainty in Social Network Analysis through Replication

by
Donald Daniel Koban

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Operations Research

Raleigh, North Carolina

2014

APPROVED BY:

| | |
|---|---|
| Dr. Thom Hodgson<br>Committee Chair | Dr. Michael Kay |
| Dr. Shu-Cherng Fang | |

## BIOGRAPHY

Major Donald Daniel Koban graduated from the United States Military Academy in 2004 with a Bachelor of Science in Engineering Management and was commissioned as a Second Lieutenant in the infantry corps.

His first assignment was to the 1st Battalion of the 502nd Infantry Regiment at Fort Campbell, Kentucky, where he served as an infantry platoon leader prior to deploying to Iraq in support of Operation Iraqi Freedom. During his deployment Donald served as a military transition team maneuver advisor to the 1/4/6 Iraqi Army Battalion in Lutifiyah, Iraq.

Following his redeployment in September of 2006, Donald served as a company executive officer and assistant battalion operations officer before deploying to northwest Baghdad in October of 2007. During his second deployment, Donald served six months as a military transition team maneuver advisor to the 5/22/6 Iraqi Army and seven months as a brigade all-source intelligence analyst. After redeploying in November of 2008, Donald served as the brigade deputy chief of operations prior to joining the 49th Quartermaster Group at Fort Lee, Virginia. Donald served as the group intelligence officer and later took command of the group headquarters company in June of 2011.

Upon graduation, Donald will be assigned to the United States Military Academy as a mathematics instructor.

## ACKNOWLEDGMENTS

To my advisor, Dr. Thom Hodgson, thank you for your patience and support. Your advice throughout the research process was exactly what was needed to guide me on the path to successfully completing my thesis while allowing me the freedom to pursue a topic I am passionate about.

To Dr. John Lavery, thank you for discussing and reviewing drafts of my research. Your continuous and timely constructive criticism, editing, and coaching helped me identify areas to improve my research.

To Dr. Michael Kay, thank you for providing me with the knowledge and ability to code in MATLAB. Without these skills, I would have struggled completing my research.

To Colonel John Graham, thank you for guidance and advice on how to integrate my military perspective into the academic world. Your insight on how to approach the research process provided me with the context to establish effective research goals and objectives, and your help in introducing me to other members in the network analysis field was absolutely critical to my effort.

To Dr. Kathleen Carley, your training on social network analysis introduced me to field I did not know existed and served as the starting point for my research. Thank you and your staff for providing professional instruction that facilitated a quick and thorough understanding of network analysis.

To Randy Munch and the Training Brains Operations Center Staff, your advanced network analysis training was the final piece I needed to tie everything together. Your team showed me how the concepts I learned from Dr. Carley's training are being applied to military operations.

## TABLE OF CONTENTS

## LIST OF TABLES

**LIST OF FIGURES**

# CHAPTER 1 INTRODUCTION

Over the past decade, the United States military has primarily been focused on counterinsurgency operations in Iraq and Afghanistan and as a result has markedly improved its understanding of irregular warfare. One of the most significant changes in the way the military understands irregular warfare is the realization that "success depends as much on understanding the social and political fabric of the situation as it does on the ability to physically dominate it" (Army Strategic Planning Guidance [ASPG], 2013, p. 4). As the objectives of warfare extend beyond physically dominating an adversary, it is no longer adequate to simply identify where an enemy is located or to focus solely on neutralizing, disrupting, or eliminating a threat. Leaders must understand "the human environment and dynamics in the entire engagement space" and must pursue lethal and nonlethal interventions to achieve complex objectives that involve social influence (Defense Science Board, 2009, p. 1). Counterinsurgency doctrine found in Field Manual (FM) 3-24 advocates that social network analysis (SNA) is a tool that can aid leaders in understanding "the organizational dynamics of an insurgency" by "formalizing the informality of insurgent networks" (Field Manual [FM] 3-24, 2006, p. B-10). The process of understanding an operational environment is further complicated as analysts study networks that are based on assumptions and incomplete, imperfect information. According to the 2013 Army Strategic Planning Guidance, today's leaders are required to be "comfortable making decisions with imperfect information in any situation" (ASPG, 2013, p. 13). The methods presented in the following paper provide a technique to quantify the certainty of SNA outputs in an environment where the structure of a social network is uncertain.

The military has greatly improved its doctrine on irregular warfare and has incorporated analytical concepts found throughout SNA academic literature dating back to the early 20$^{th}$ century. The concept of mapping out individuals and their interactions is a proven and accepted technique in the field of sociology, and the military has successfully incorporated SNA techniques in Iraq and Afghanistan. The Army's counterinsurgency manual defines key SNA metrics such as density, degree centrality, and betweenness centrality and provides context to explain how different SNA metrics can be used to identify key individuals and assess the effects of targeting interventions. Unfortunately, current doctrine is limited by the fact that it offers no explanation of how to mathematically calculate SNA metrics.

In a similar fashion, current counterinsurgency doctrine conveys the importance of distinguishing confirmed associations from suspected associations, but offers a limited description on the methods for classifying associations and offers no explanation of how to incorporate the assessed reliability of information into SNA. FM 3-24 provides no formal definition of what constitutes a known association and instead gives two examples of known associations which include "face-to-face meetings" and "confirmed telephonic conversations" (FM 3-24, 2006, p. B-17). FM 2-22.3, Human Intelligence Collector Operations, provides a similar description of known and suspected associations and defines a "suspected" association as those "in which there are indicators that individuals may have had associations but there is no way to confirm that association" (Field Manual [FM] 2-22.3, 2006, p. 12-7). One example of a suspected association found in FM 2-22.3 is "a face-to-face meeting where one party can be identified, but the other party can only be tentatively identified" (FM 2-22.3, 2006, p. 12-7). When considering the varying degrees of uncertainty associated with different intelligence sources, it is difficult to definitively classify information as certain while adhering to the "analytic principles" outlined in Joint Publication (JP) 2-01, Joint and National Intelligence Support to Military Operations (Joint Publication [JP] 2-01, 2012, p. D-7).

The analytic principle of "being precise in what is known" requires analysts to clearly inform leaders of the source reliability of known information which includes specifying whether information is direct or indirect. Additionally, JP 2-01 warns against the possibility of deception and recommends that analysts guard against it by determining "whether all the sources and collection platforms that should be reporting on a matter have indeed done so" (JP 2-01, 2012, p. D-8). An easier way to describe the process of screening information for deception is the concept of corroboration. When independent sources report the same information about a relationship between two individuals, an analyst can be more certain that the reported information is not fabricated. However, corroborating reporting may still not increase the certainty of information if circular reporting[1] is involved. In some cases where information is not readily

[1] Circular reporting is a situation where a piece of information appears to come from multiple independent sources, but in fact is coming from only one source (Circular reporting, 2013). For instance if source A reports an association and source B reports the same association that he or she learned from source A, then source B can no longer be considered an independent source since their knowledge is dependent on source A.

available or accessible, confirming information as fact may only be justified in retrospect.

A conservative approach for dealing with information uncertainty might be to only include information that is considered factual. However, only including confirmed information could prevent an analyst from including important pieces of information that would provide a more accurate representation of a true network. According to FM 2-22.3, the “rationale for depicting suspected associations is to get as close as possible to an objective analytic solution while staying as close as possible to known or confirmed facts” (FM 2-22.3, 2006, p. 12-7). Including uncertain information in network analysis may produce a slightly flawed representation of a network, but using only known associations provides a limited representation of a network where key valid associations may be omitted simply because an analyst is uncertain they exist. Stated more simply, if information cannot be proportionally emphasized, then leaders are faced with a choice of making a misinformed decision or an uninformed decision.

The use of mathematically calculated SNA metrics known as centrality measures is a well-established practice for analyzing networks and SNA metrics have produced validated results in the field of sociology. Wasserman & Faust (1994) highlight the topics listed below as examples of past network analysis research:

- Occupational mobility (Breiger, 1981, 1990)
- The impact of urbanization on individual well-being (Fischer, 1982)
- The world political and economic system (Snyder and Kick, 1979; Nemeth and Smith, 1985)
- Community elite decision making (Laumann, Marsden, and Galaskiewicz, 1977; Laumann and Pappi, 1973)
- Consensus and social influence (Friedkin, 1986; Friedkin and Cook, 1990; Doreian, 1981)

After the September 11th terrorist attack, academics in the field of sociology felt that SNA could be extended to the analysis of covert networks. One of the most notable initial applications of SNA to covert networks is Valdis Krebs’ analysis of the 9/11 hijacker cell (Krebs, 2002). Krebs built a network using publicly available media to construct a social network of the terrorist cell and then used mathematical metrics related to an individual's location in the overall network structure to determine the key members of the group.

While many advances have been made on how to optimally destabilize a network, there has been little research focused on addressing the significant issue of accounting for the uncertainty of reported information. Collecting information on covert networks that intentionally conceal relationships, intentions, and activities jeopardizes the validity of SNA techniques since network analysis can be skewed by the improper inclusion or exclusion of individuals and associations in a network. The following paper proposes replication as a means to analyze network metrics across a broad range of possible network structures.

# CHAPTER 2 SOCIAL NETWORK ANALYSIS CONTEXT

Social networks are a way of thinking about social systems that focus attention on the relationships among entities that make up a system, which are called actors or nodes (Borgatti, Martin, & Johnson, Analyzing Social Networks, 2013, p. 1). Social scientists use social network analysis (SNA) to analyze the "relationships among social entities, and on the patterns and implications of these relationships" (Wasserman & Faust, 1994, p. 3). In many cases, social scientists focus on answering research questions such as "What is the basis of friendship ties?", and they use quantitative network measures, known as centrality measures, to explain things such as the distribution of power and influence within an organization. (Borgatti, Brass, Mehra, & Giuseppe, 2009, p. 894). Likewise, military forces embrace SNA as an effective approach for studying and destabilizing terrorist groups and credit SNA for contributing to successes such as the capture of Saddam Hussein (Borgatti et al., 2009, p 893). In both cases, a fundamental axiom of SNA is that a node's position in a network determines in part the opportunities and constraints that it encounters (Borgatti et al., 2009, p. 893). SNA combines social theory with mathematical methods to provide quantitative measures of social properties that might otherwise be defined only in qualitative terms (Wasserman & Faust, 1994, p. 10). This chapter will introduce basic SNA concepts and methods of how SNA is used to calculate quantitative measures of an entire network and measures to assess the importance of individual actors within a network.

### 2.1 SNA TERMS AND CONCEPTS

SNA methods are based on graph theory which is used to represent a network with a set of vertices (called nodes or actors) and a set of edges (called links) (Borgatti et al., 2013, p.12). When conducting social network analysis there are some basic terms that should be understood. These include:

**One-mode and two-mode networks.** Networks are classified as one-mode if they involve measurements on a single set of actors. Relations in a one-mode network are measured directly at the level of pairs of actors where relations represent specific substantive connections between pairs of actors. An example of a one-mode network relation is kinship where the relation is identified by means such as interviewing actors or examining archival data (Borgatti et al., 2013, p.231). In situations where relations cannot be collected directly, two-mode networks are used. Networks are classified as two-mode if they involve measurements on two sets of nodes such as a set of actors and a set of events. Relations in a two-mode network are a function of dyads in which relations between actors are implied indirectly through shared events. Two-mode network relations are sometimes referred to as affiliation networks where two actors are assumed to interact because they share events such as attending a meeting together or belonging to a common organization. The most common technique for analyzing two-mode networks is to convert a two-mode matrix to a one-mode matrix by using linear algebra (Borgatti et al., 2013, p.233). Section 2.3 provides a detailed description of how to transform a two-mode network. Military doctrine advocates the use of “association matrices” for one-mode data and “activities matrices” for two-mode data. (FM 3-24, 2006).

One-mode data is more common in sociology studies where researchers solicit relations between individuals based on surveys or interviews. One-mode data is also used when studying terrorist networks, but the nature of intelligence reporting on covert network often leads to a large amount of two-mode data.

**Nodes.** In traditional social network analysis, nodes are typically individuals or organizations (Borgatti et al., 2013, p. 30). However, nodes can represent many different things such as locations, events, resources, actions, knowledge, tasks, roles, or beliefs. The selection of types of nodes is dependent upon the objectives and nature of research.

**Edges.** An edge represents a social relation between nodes and it is usually viewed as representing specific substantive connections (Wasserman & Faust, 1994, p. 37). Some examples of types of relations include: (Wasserman & Faust, 1994, p. 37)

- Individual evaluations: friendship, linking, respect, and so forth
- Transactions or transfer of material resources: lending or borrowing, buying or selling
- Transfer of non-material resources: communications, sending/receiving information
- Interactions

- Movement: physical (migration from place-to-place), social (movement between occupations or statuses)
- Formal roles
- Kinship: marriage, descent

When the type of relation can be determined, edges are often assigned values that represent the strength of a relation because different types of relations often represent varying strengths or frequencies of interaction. In other cases where the types of relation are less clear or overly complex, it is common to use binary weighting for nodes where a value of 1 represents the existence of a tie and a value of 0 represents the absence of a tie. Another important characteristic of edges is that they can be represented as a directed link or undirected link. Directed edges (also known as arcs) are used to represent relations that "logically have a sense of direction – for example, 'is the parent of' and 'gives advice to'"(Borgatti et al., 2013, p. 12). Undirected links are used when "direction does not make sense or logically must always be reciprocated, as in 'was seen with' or is 'kin to'" (Borgatti et al., 2013, p. 12).

**Path.** A path is a sequence of adjacent nodes in which no single node is revisited. In many cases a pair of nodes may be connected by multiple paths with varying length. The length of a geodesic path is the number of edges in the path and is called the geodesic distance.

**Graphs.** A graph is a visual representation of a network in which nodes are represented with points and edges are represented with lines. An edge is either present or absent between each pair of nodes.

**Association matrices.** In Army doctrine, an association matrix refers to a square, symmetric adjacency matrix that is used to represent one-mode data. In an association matrix, rows and columns represent actors, and an entry in row i and column j represents a link from actor i to actor j (Borgatti et al., 2013, p. 18). When two vertices are connected by an edge, the vertices are adjacent (Borgatti et al., 2013, p. 12). Army doctrine does not provide an example of a directed network, but if a network contains directed links then the association matrix can be asymmetric or symmetric. If a network is undirected, then the association matrix will be symmetric.

**Activities matrices.** An activities matrix is a military term that refers to a two-mode adjacency matrix, which is also referred to as an affiliation matrix in academic literature. In an activities matrix, rows represent actors, columns represent events, and an entry in row i and column j represents a shared event between actor i and event j. A one-mode matrix of co-occurrences of shared events between actors can be constructed by post-multiplying a two-mode matrix by its transpose (Borgatti et. al, 2013, p. 235). It is important to note that two-mode data is based on the assumption that shared events are an indicator of social interactions and relationships. Consequently, there is usually a greater degree of uncertainty associated with two-mode data when compared to one-mode data which is based on direct associations.

### 2.2 SNA METHODS & NETWORK MEASURES

Social network measures are used to analyze and describe networks at the organizational-level or individual-level depending on the nature of a research question. When a researcher is comparing two or more independent networks, network measures such as density are used. Density is the proportion of ties in a network relative to the total number of possible ties (FM 3-24, 2006, p. B-11). When a researcher is interested in identifying key influential individuals or knowledgeable individuals in a network, individual-level measures called centrality measures are used. There are numerous centrality measures used to describe an individual's importance, but Army network training emphasizes the following four measures: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality.

Degree centrality describes how active an individual is in the network and is measured using the concept of degrees – the number of direct connections a node has (FM 3-24, 2006, p. B-14). Mathematically, the normalized degree centrality of a node is calculated as the row (or column) sums of an adjacency matrix divided by maximum number of possible links (for a network of size $n$, this is $n$-1) (Wasserman & Faust, 1994, p. 178):

$$C_D(n_i) = \frac{\sum_j x_{ij}}{n-1}$$

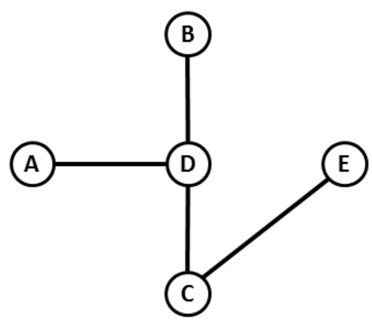
where:

$C_D(n_i) = \textit{the degree centrality of node } n_i$

$x_{ij} = \textit{the } (i,j) \textit{ entry of the adjacency matrix}$

$n = \textit{the total number of nodes in the network}$

A node that has a high degree centrality is known as a "hub". In military targeting applications, individuals with high degree centrality are believed to have larger exposure to a network and opportunities to directly influence other actors. They are believed to be valuable targets for collecting information on a network. Actor D has the highest degree centrality in the graph depicted in Figure 1.




| Degree Centrality | |
|---|---|
| Node A | 0.25 |
| Node B | 0.25 |
| Node C | 0.50 |
| Node D | 0.75 |
| Node E | 0.25 |

**Figure 1: Degree Centrality for Example Undirected Binary Network**

Closeness centrality is a measure of how close an actor is to all other actors in a set of actors (Wasserman & Faust, 1994, p. 183). Closeness provides an indication of which actors are most central in a network because it is assumed that individuals with high closeness centrality can interact more quickly with all other actors. Only geodesic paths are considered when calculating closeness centrality. Mathematically, the normalized closeness centrality of a node is calculated as the sum of geodesic distances from a node to all other nodes divided into the minimum possible sum of distances (for a network of size *n*, *n*-1 is the minimum sum of distances and is associated with a node that is adjacent to all other actors) (Wasserman & Faust, 1994, p. 185):

$$C_C(n_i) = \frac{n-1}{\left[\sum_{j=1}^{n} d(n_i, n_j)\right]}$$

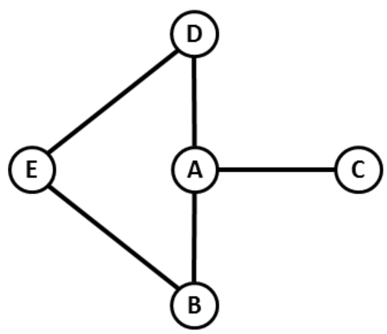
where:

$C_C(n_i) = \textit{the closeness centrality of actor } n_i$

$d(n_i, n_j) = \textit{minimum number of links between actors } n_i \textit{ and } n_j$

$n = \textit{the total number of nodes in the network}$

In military targeting applications, individuals with high closeness centrality are believed to be valuable targets for collecting information on a network. Actor A has the highest closeness centrality in the graph depicted in Figure 2.




| Closeness Centrality | |
|---|---|
| Node A | 0.80 |
| Node B | 0.66 |
| Node C | 0.50 |
| Node D | 0.66 |
| Node E | 0.57 |

**Figure 2: Closeness Centrality for Example Undirected Binary Network**

Betweenness centrality indicates the extent to which an individual lies between other individuals in a network, serving as an intermediary, liaison, or bridge (FM 3-24, 2006, p. B-14). When calculating betweenness centrality, only geodesic paths are considered because it is assumed that information or resources will always opt to use the shortest path between two nodes. An actor with high betweenness centrality has great influence over the flow of information and/or resources in a network and is considered to have potential for informal power. Betweenness centrality is calculated for a focal node by computing, for each pair of nodes other than the focal node, what proportion of all the geodesic paths from one to the other pass through the focal node. These proportions are summed across all pairs and the result is a single value for each node in the network (Borgatti et al., 2013, p. 174). The normalized formula for betweenness centrality of node i is obtained by dividing betweenness values by the maximum number of pairs of actors not including the $i^{th}$ node (Wasserman & Faust, 1994, p. 190):

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}} \cdot \frac{1}{[(n-1)(n-2)/2]}$$

where:

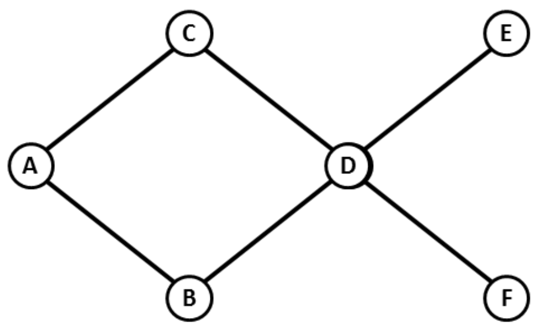
$C_B(n_i) = \textit{the betweenness centrality of actor } n_i$

$g_{jk} = \textit{the number of geodesic paths linking nodes } j \textit{ and } k$

$g_{jk}(n_i) = \textit{the number of geodesic paths linking two nodes that contain node } i$

$n = \textit{the total number of nodes in the network}$

In military targeting applications, individuals with high betweenness centrality are believed to be valuable targets for disrupting and fragmenting a network. Actor D has the highest betweeness centrality in the graph depicted in Figure 3.




| Betweenness Centrality | |
|---|---|
| Node A | 0.05 |
| Node B | 0.15 |
| Node C | 0.15 |
| Node D | 0.75 |
| Node E | 0 |
| Node F | 0 |

**Figure 3: Betweenness Centrality for Example Undirected Binary Network**

Eigenvector centrality can be described in many ways, but Borgatti et al. (2013) provides one of the clearer descriptions. Borgatti et al. (2013) describes eigenvector centrality as a measure of popularity in the sense that a node with high eigenvector centrality is connected to nodes that are themselves well connected (p. 168). Mathematically, eigenvector centrality is calculated by solving a system of homogenous linear equations where each node's popularity contributes to the popularity of other nodes (Bonacich, 1972, 115).

$$e_i = x_{i1}e_1 + x_{i2}e_2 + x_{i3}e_3 \ldots + x_{in}e_n$$

where:

$e_i = the\ popularity\ score\ of\ actor\ i$

$x_{ij} = the\ (i,j)\ entry\ of\ the\ adjacency\ matrix$

The above equation in matrix form is $e = Xe$, or $(X - I)e = 0$, where $e$ is the vector or unknown popularity scores. If the system of equations is structured in this manner, the solution vector will be nonzero only if the $det|X - I| = 0$. For this reason, Bonacich (1972) incorporated eigenvalues into the system of equations in a manner that guarantees a nonzero solution vector of popularity scores. The first step of Banacich's method is to add a constant to the left hand side of the popularity equation.

$$\lambda e = Xe, \quad or \quad (X - \lambda I)e = 0$$

where:

$e = the\ vector\ of\ unknown\ popularity\ scores$

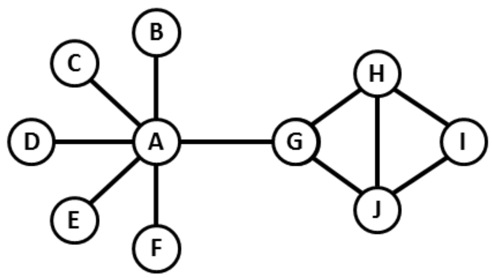
$X = the\ network\ adjacency\ matrix$

$\lambda = a\ constant$

By adding a constant to the left hand side, it is possible to solve for values that make the $det|X - \lambda I| = 0$ (i.e. the eigenvalues of the adjacency matrix). Once the eigenvalues have been obtained, solution vectors (i.e. eigenvectors) of popularity scores can be determined by solving the system of equations with any of the eigenvalues serving as the constant term. However, since the intention is to have positive values for popularity scores, Bonacich (1972) specifies that the eigenvector associated with the largest eigenvalue should be used as the solution. This concept of using the largest eigenvalue is based on the Perron-Frobenius theorem which states that a real square matrix with positive entries will have a unique largest eigenvalue with a corresponding eigenvector consisting of strictly positive elements (Perron, 1907).

Once the eigenvector centrality values have been calculated, they can be normalized by dividing each value by the maximum possible eigenvector centrality value of $\sqrt{1/2}$. This maximum occurs when a network consists of a single pair of adjacent actors and it is the maximum value regardless of the network size (Wei, Pfeffer, Reminga, & Carley, 2011).

In military targeting applications, individuals with high eigenvector centrality are believed to be valuable targets for collecting information on a network. In the graph depicted in figure 4, actor A is well connected in the network, but actor G has a higher eigenvector centrality because actor G is associated with actors that that are also well connected. Since actors A is associated with "less important" individuals, it is possible for a node that is only adjacent to three nodes to outrank a node that is adjacent to twice as many nodes.



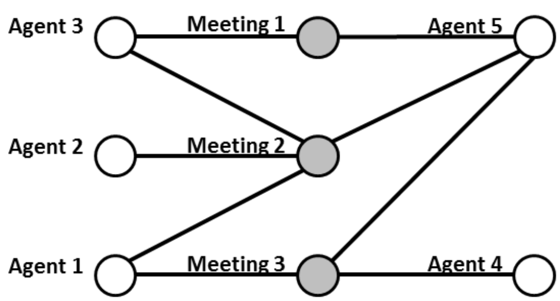

| Eigenvector Centrality | | | |
|---|---|---|---|
| Node A | 0.639 | Node G | 0.655 |
| Node B | 0.228 | Node H | 0.599 |
| Node C | 0.228 | Node I | 0.427 |
| Node D | 0.228 | Node J | 0.599 |
| Node E | 0.228 | | |
| Node F | 0.228 | | |

**Figure 4: Eigenvector Centrality for Example Undirected Binary Network**

## 2.3 TRANSFORMING AFFILIATION NETWORKS TO SOCIAL NETWORKS

When the goal of SNA is to identify key actors, two-mode affiliation networks must be transformed to a one-mode, social network. As such, it is important to understand that there is an inherent level of uncertainty associated with links between individuals when they are generated from affiliation networks. The following section provides a detailed overview of how to transform two-mode data to one-mode data and a brief discussion of the uncertainty associated with affiliation data.

Affiliation networks refer to two-mode networks where links represent relations between actors and a set of events. Examples of two-mode networks include: "agent by location", "agent by event", and "agent by organization", where associations between agents can be derived from shared locations, shared events, or shared organization. Two-mode data is represented using bipartite graphs which include two node classes where links represent one node's affiliation with a separate node. Bipartite graphs are useful because indirect connections are more apparent when analysts can see the shared nodes that are being used to infer a relation. However, if an analyst is attempting to use SNA to identify key individuals, the network must be a one-mode network where nodes represent individuals and links representing relations between individuals.  Figure 5 depicts a bipartite graph and affiliation matrix for an "actor by meeting" two-mode network.

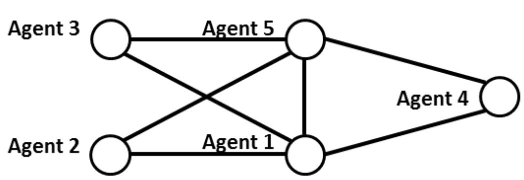



**Affiliation Matrix**

| | Meeting 1 | Meeting 2 | Meeting 3 |
|---|---|---|---|
| Agent 1 | 0 | 1 | 1 |
| Agent 2 | 0 | 1 | 0 |
| Agent 3 | 1 | 1 | 0 |
| Agent 4 | 0 | 0 | 1 |
| Agent 5 | 1 | 1 | 1 |

**Figure 5: Example Affiliation Matrix**

Transforming two-mode data to one-mode data is a fairly simple process that involves matrix multiplication.  When the affiliation matrix is multiplied by its transpose, the result is a one-mode, symmetric, valued association matrix where links represent the number of shared meetings. Figure 6 depicts a graph and association matrix derived from the two-mode network shown in Figure 5.




**Association Matrix**

| | Agent 1 | Agent 2 | Agent 3 | Agent 4 | Agent 5 |
|---|---|---|---|---|---|
| Agent 1 | 2 | 1 | 1 | 1 | 2 |
| Agent 2 | 1 | 1 | 1 | 0 | 1 |
| Agent 3 | 1 | 1 | 2 | 0 | 2 |
| Agent 4 | 1 | 0 | 0 | 1 | 1 |
| Agent 5 | 2 | 1 | 2 | 1 | 3 |

**Figure 6: Example Association Matrix**

The diagonal entries of the derived association matrix represent the number of meetings attended by each agent, and the individual entries in the matrix correspond to the number of shared events between two actors.  It is important to note the diagonal values of the association matrix do not affect centrality measures, but the individual entries in the matrix do. When considering the meaning of edge values discussed in section 2.1, it makes sense to assess a stronger relation between individuals that share multiple events compared to individuals that share only a single event. However, analysts often use binary edge values because trying to accurately represent varying strengths of interactions can become tedious and overly complicated, especially when social networks are created by transforming several affiliation networks.

Using shared events to infer associations between individuals may be a reasonable assumption, but this method also carries a degree of uncertainty that must be understood when performing network analysis. When associations are derived from large events, the potential that two individuals actually interact is less than when associations are derived from small events. For example two individuals could have attended the same university and never met each other. On the other hand, if two individuals were in the same department and attended several classes together, then there is a greater potential that the two individuals actually interacted. The uncertainty inherent in two-mode data generally makes one-mode data solicited through surveys or interviews preferable to one-mode data derived from inferred associations.

# CHAPTER 3 DARK NETWORK DATA COLLECTION AND RELIABILITY ASSESSMENTS

Dark network is a metaphorical term used to describe covert and illegal networks that seek to conceal themselves from authorities (Everton, 2012, p. xxv). The methods used for analyzing dark networks are the same as other social networks, but the main difference lies with the availability and reliability of data. The main challenge when studying dark networks is gathering timely, accurate, and complete data (Everton, 2012, p. xxvi). The following chapter provides a discussion of how the military collects data on covert networks and discusses how reliability and validity are assessed for various types of reporting.

Before continuing, it is important to understand the difference between information and "intelligence". Joint Publication (JP) 2-0, Joint Intelligence, defines intelligence as "the product resulting from the collection, processing, integration, evaluation, analysis, and interpretation of available information concerning foreign nations, hostile or potentially hostile forces or elements, or areas of actual or potential operations." (Joint Publication [JP] 2-0, 2007, p. I-15) The key distinction is that intelligence involves a "synthesis of quantitative analysis and qualitative judgment that is subject to competing interpretation" (JP 2-0, 2007, p. I-2). The challenge of accurately analyzing information is further complicated by the fact that an analyst almost always bases assessments on incomplete information. As such, intelligence inherently has uncertainty associated with the information that is on hand as well as uncertainty associated with unknown information.

Intelligence ultimately drives the conduct of operations and influences how a commander organizes forces, employs forces to accomplish objectives, and protects forces from threats. Ironically, the best information is usually generated as a result of effective operations. Conversely, inaccurate intelligence produces ineffective operations which can negatively affect a host nation population and expose military forces to unnecessary risk. For example, when innocent civilians are improperly identified as threats and detained, the host nation may lose confidence in security forces. Therefore, commanders must understand that intelligence is an estimate and they must accept an amount of risk when formulating plans based on an analyst's assessment (JP 2-0, 2007, p.

II-10). Intelligence analysts are tasked with the responsibility of minimizing and precisely articulating the uncertainty associated with their assessments.

### 3.1 DATA COLLECTION

Unlike traditional social network analysis that usually involves data obtained through surveys, interviews, and direct observation, network data in a military environment is collected from multiple sources known as intelligence sources. Intelligence sources can be people, documents, equipment, or technical sensors, and are grouped according to one of the seven major intelligence disciplines: geospatial intelligence (GEOINT); human intelligence (HUMINT); signals intelligence (SIGINT); measurement and signature intelligence (MASINT); open-source intelligence (OSINT); technical intelligence (TECHINT); and counterintelligence (CI). (JP 2-0, 2007, p. I-6). Figure 7 shows a summary of intelligence disciplines, subcategories, and sources.

GEOINT -- Geospatial Intelligence
- -- Imagery
- -- IMINT - Imagery Intelligence
- -- Geospatial Information

HUMINT -- Human Intelligence
- -- Debriefings
- -- Interrogation Operations
- -- Source Operations
- -- Document and Media Exploitation

SIGINT -- Signals Intelligence
- -- COMINT - Communications Intelligence
- -- ELINT - Electronic Intelligence
  - ** Technical ELINT
  - ** Operational ELINT
- -- FISINT - Foreign Instrumentation Signals Intelligence

MASINT -- Measurement and Signature Intelligence
- -- Electromagnetic Data
- – Geophysical Data
- – Materials Data
- – Radio Frequency Data
- – Radar Data
- – Nuclear Radiation Data

OSINT -- Open-Source Intelligence
- – Academia
- – Interagency
- – Newspapers/Periodicals
- – Media Broadcasts
- – Internet

TECHINT -- Technical Intelligence

CI -- Counterintelligence

**Figure 7: JP 2-0 Intelligence Disciplines, Subcategories, and Sources**

Intelligence collection produces various types of information that is organized by five main categories. JP 2-01 provides detailed descriptions and examples of each type of information which is extremely useful in governing when it is appropriate for an analyst to consider information as factual. Table 1 provides the doctrinal types of intelligence data as described in JP 2-01.

**Table 1: JP 2-01 Types of Intelligence Data**

**TYPES OF INTELLIGENCE DATA**

| Term | Definition | Example |
|---|---|---|
| Fact | Verified information; something known to exist or to have happened. | A confirmed inventory of a resource of one's own service |
| Direct Information | Information relating to an intelligence issue under scrutiny the details of which can, as a rule, be considered factual, because the nature of the source, the source's direct access to the information, and the concrete and readily verifiable character of the contents. | Foreign official report providing a specific piece of information within their purview or human intelligence from a US diplomatic officer who directly observed an event. |
| Indirect Information | Information relating to an intelligence issue the details of which may or may not be factual, the doubt reflecting some combination of the source's questionable reliability, the source's lack of direct access, and the complete character of the contents. | Human intelligence from a reliable agent, citing secondhand information from a source of undetermined reliability. |
| Direct Data | Organized information that provides context for evaluating the likelihood that a matter under scrutiny is factual. | Charts, graphs, or tables depicting organized data collected by US personnel or trusted agents. |
| Indirect Data | Organized information that provides context for evaluating the likelihood that a matter under scrutiny is factual. | Charts, graphs, or tables depicting organized data collected by a liaison intelligence service. |

Even though doctrine provides definitive descriptions and an example of each type of data, applying those definitions to actual intelligence reports still involves tedious assessments of source reliability and continuous efforts to gather corroborating intelligence reports. Assessing source reliability is difficult because verifying reported information often requires retrospective analysis or corroboration through independent sources or collection platforms. Considering the large amount of intelligence reporting that is collected from various sources on a daily basis, it is unrealistic to expect that every detail of every report can be corroborated through independent sources. This is particularly true for information concerning activities that dark networks are attempting to conceal. Finding a single source with access to information on the operations of terrorist organization is difficult in itself let alone finding additional sources that can corroborate the information. In some cases when a single source is classified as "reliable" and by definition has "a history of complete reliability", it may be reasonable to place a greater degree of emphasis on newly reported information provided by that particular source. However, if a single source is classified as "usually reliable" and by definition has "a history of valid information most of the time", then it may only be reasonable to trust the information if the content of the information is logical and consistent with previously reported information. In both cases, an analyst should always question the reliability of the source and the content of reported information. They also should be precise in describing source reliability and distinguishing between verified and unverified information.

When indirect, second-hand information is reported, the burden of proof for verification is greater. Even if a reliable source reports second-hand information, the analyst should never consider the information to be factual because the source lacks direct access to the information. If there are multiple corroborating intelligence reports of second-hand information, the information still represents indirect information and should not be considered factual. The key principle is that the nature of reported information has a significant effect on the proportion of emphasis that an analyst should place on each piece of information.

The process of assessing known and suspected associations for social network analysis (SNA) adds an additional layer of uncertainty especially when associations are based on affiliation data. In a best case scenario a confirmed association would be based on corroborating, first-hand information about a direct association between two individuals. However, a common scenario might be that a single-source has second-hand information about an indirect association. As it can be seen, accurately assessing the likelihood of an association is a very complex process that requires subjective analysis. Military doctrine does not provide a method for quantifying the likelihood of suspected associations, but it does recommend that analysts identify associations as either "confirmed", "probable", or "suspected". In order to incorporate these principles into quantitative analysis, this paper defines quantitative limits for the assessed likelihood of each type of association. Figure 8 shows an illustration of what the relationship between the nature of available reporting and reported associations might look like. The quantitative likelihood scales are derived from confidence-level scales found in JP 2-0.

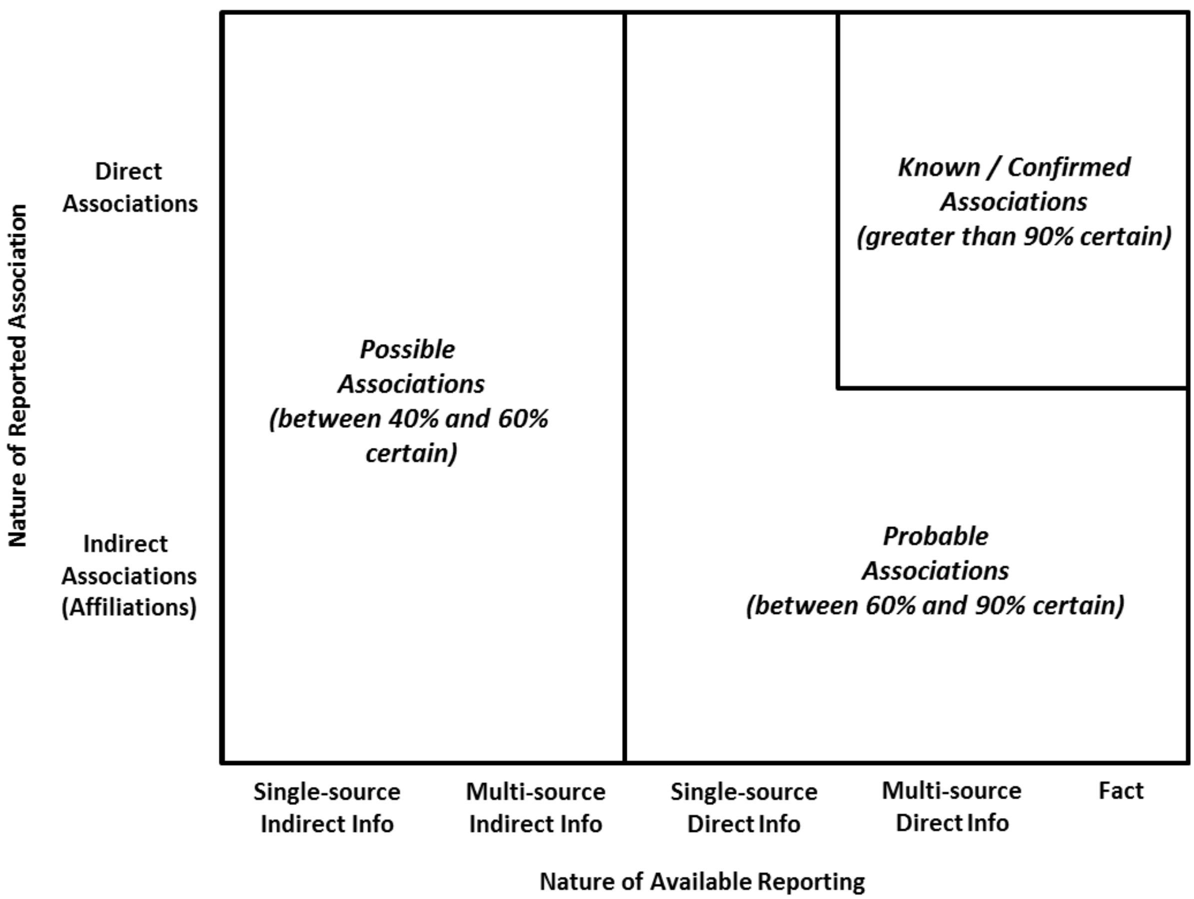



**Figure 8: Relationship Between Reported Associations and Available Reports**

Last, an additional characteristic of military data collection on dark networks is that information is more susceptible to deception and reporting errors.  Accounting for the uncertainty of reported information is one of the most difficult challenges when analyzing covert network.  The following section provides a description of doctrinal methods used to account for uncertainty.

### 3.2 INTELLIGENCE FUSION & CONFIDENCE ESTIMATES

Intelligence analysts use a process known as intelligence fusion to mitigate the uncertainty of their analytical assessments.  JP 2-0 defines fusion as “the process of examining all sources of intelligence and information to derive a complete assessment of activity” (JP 2-0, 2007, p. xiv).  Intelligence products that are based on all sources of information, both hard (physics-based, such as video, infrared imaging, synthetic aperture radar or acoustic) and soft (human-based such as oral reports, written reports or annotation) are known as “all-source” products.  All-source intelligence can reduce the influence of deception by assuring that “all the sources and collection platforms that should be reporting on a matter have indeed done so” (JP 2-01, 2012, p. D-8).  JP 2-0 also warns against analysts “placing unquestioned trust in a single-source intelligence report”.  When multiple, independent sources provide consistent information about an adversary, an intelligence analyst can be more confident that reported information is valid, which in turn improves the confidence of assumptions and conclusions that are based on the reported information.

An additional way to improve the certainty of intelligence estimates is to place “proportionally greater emphasis on information reported through the most reliable sources” (JP 2-0, 2007, II-7).  Source reliability assessments are the responsibility of the collector who produces a report at the tactical level, and these assessments are included in the report body.  Source reliability refers to a collector’s assessment of whether or not a source is providing accurate information as determined by a source’s previous reporting accuracy.   Table 2 provides a doctrinal source reliability matrix as described in FM 2-22.3 (p. B-1).

**Table 2: FM 2-22.3 Source Reliability Matrix**

| | | |
|---|---|---|
| **A** | **Reliable** | **No doubt** of authenticity, trustworthiness, or competency; has a history of complete reliability |
| **B** | **Usually Reliable** | **Minor doubt** about authenticity, trustworthiness, or competency; has a history of valid information most of the time |
| **C** | **Fairly Reliable** | **Doubt** of authenticity, trustworthiness, or competency but has provided valid information in the past |
| **D** | **Not Usually Reliable** | **Significant doubt** about authenticity, trustworthiness, or competency but has provided valid information in the past |
| **E** | **Unreliable** | **Lacking** in authenticity, trustworthiness, and competency; history of invalid information |
| **F** | **Cannot Be Judged** | **No basis** exists for evaluating the reliability of the source |

Intelligence collectors are also required to express a confidence level in the content of reported information. This confidence level assessment relates closely to the concept of intelligence fusion but only involves a collector's assessment of the consistency of information contained in a single independent report. In contrast, an all-source product represents the synthesis of multiple independent reports. Table 3 provides a doctrinal information reliability matrix as described in FM 2-22.3 (p. B-2).

**Table 3: FM 2-22.3 Information Reliability Matrix**

| | | |
|---|---|---|
| **1** | **Confirmed** | **Confirmed** by other independent sources; **logical** in itself; **Consistent** with other information on the subject |
| **2** | **Probably True** | Not confirmed; **logical** in itself; **consistent** with other information on the subject |
| **3** | **Possibly True** | Not confirmed; **reasonably logical** in itself; **agrees with some** other information on the subject |
| **4** | **Doubtfully True** | Not confirmed; possible but **not logical**; **no other information** on the subject |
| **5** | **Improbable** | Not confirmed; **not logical** in itself; **contradicted** by other information on the subject |
| **6** | **Cannot Be Judged** | **No basis** exists for evaluating the reliability of the information |

Just as collectors are required to report confidence-level assessments with precise terminology and codes, all-source analysts should also provide the commander with precise confidence-level estimates related to their analysis. All-source assessments represent the synthesis of multiple independent reports which all have varying degrees of uncertainty. An all-source analyst must also use his or her judgment to determine what reports to include in an assessment, and should always consider alternative hypotheses to account for the uncertainty of the information they include in their analysis. The doctrinal terms used to express confidence levels in all-source assessments are still based on subjective assessments, but they allow analysts to communicate uncertainty levels in a uniform and consistent manner. The qualitative and quantitative confidence-level scales found in JP 2-0 are show in Table 4.

Table 4: JP 2-0 Confidence-level Scales

| INTELLIGENCE CONFIDENCE LEVELS | | |
|---|---|---|
| **Description of Probability or Confidence** | **Synonyms** | **Percent** |
| HIGHLY LIKELY | ◆ Highly Probable<br>◆ We Are Convinced<br>◆ Virtually Certain<br>◆ Almost Certain<br>◆ High Confidence<br>◆ High Likelihood | >90% |
| LIKELY | ◆ Probable<br>◆ We Estimate<br>◆ Chances Are Good<br>◆ High-Moderate Confidence<br>◆ Greater Than 60% Likelihood | 60-90% |
| EVEN CHANCE | ◆ Chances Are Slightly Greater (or Less) Than Even<br>◆ Chances Are About Even<br>◆ Moderate Confidence<br>◆ Possible | 40-60% |
| UNLIKELY | ◆ Probably Not<br>◆ Not Likely<br>◆ Improbable<br>◆ We Believe…Not<br>◆ Low Confidence<br>◆ Possible but Not Likely | 10-40% |
| HIGHLY UNLIKELY | ◆ Highly Improbable<br>◆ Nearly Impossible<br>◆ Only a Slight Chance<br>◆ Highly Doubtful | <10% |

# CHAPTER 4 EDGE LIKELIHOOD REPLICATION ANALYSIS

As stated previously, social network analysis (SNA) is based on the axiom that a node's position in a network determines in part the opportunities and constraints that it encounters (Borgatti et al., 2009, p. 894). Using network structure as the foundation for network analysis, it follows that an individual's importance in a network is dependent upon how an analyst models the structure of a network. When networks are built from a collection of reliable data, calculating centrality measures is a fairly simple process that produces useful, quantitative measures for determining critical members in a network. However, when the accuracy of network data is uncertain, there is a high probability that the representation of the network will be flawed. When centrality measures are based on an imperfect representation of a network, it becomes difficult to express the uncertainty of social network analysis measurements. The following section provides a method to quantify the uncertainty of SNA through the use of replication that takes into account a broad range of possible network structures.

Section 3.2 provides an overview of the methods used to express uncertainty in individual intelligence reports and all-source intelligence products. Military doctrine places a great deal of emphasis on precisely describing the uncertainty associated with intelligence estimates, so there is also a need to quantify uncertainty in SNA. To some degree, the uncertainty of SNA measures can be controlled by the deliberate inclusion of only network data that are very certain to be accurate. However, the exclusion of relevant, uncertain information equally affects a person's understanding of a network. Objectively determining what data provides the most accurate representation of a network can usually only be achieved through retrospective analysis. The best approach is most likely the approach recommended in FM 2-22.3, which recommends including suspected associations while staying as close as possible to known or confirmed facts.

If an analyst chooses to include suspected associations, it is difficult to accurately represent uncertain associations when calculating centrality measures. Compensating for uncertainty by using valued edges that represent the probability that interaction occurs between two members is a reasonable technique for degree and eigenvector centrality measures which are based on first order associations. However, measures that are more sensitive to network structure such as closeness and betweenness can vary greatly based on the inclusion or omission of edges that connect subgroups in the network.

Borgatti et al. (2006) specifically studied the effects of missing nodes and edges in his paper "On the Robustness of Centrality Measures under Conditions of Imperfect Data" by using replication to randomly add and remove a specified percentage of nodes and edges from random Erdos-Renyi networks of varying sizes. His study demonstrated that centrality measures are very robust under small amounts of error (10% and under), and found that the accuracy of centrality measures decrease predictably and monotonically with increases in error. Another result was that network density tends to reduce accuracy for all kinds of error except edge addition. (Borgatti, Carley, & Krackhardt, 2006).

The methods for edge-likelihood replication analysis are very closely related to the methods used by Borgatti et al. (2006) but differ in the type of networks they are applied to and in the how edges are removed from a network. Borgatti et al. (2006) focused his study on randomly generated Erdos-Renyi networks which involve edges that are stochastically formed between pairs of nodes with equal probability. As a result, on average all nodes will have the same the degree and the networks will be well connected. The analysis presented in this paper is based specifically on the analysis of dark networks characterized by a cellular structure with limited connections between subgroups. Borgatti et al. (2006) also randomly removed nodes and edges in his study. Edge-likelihood techniques are based on a "weighted" edge removal where the weight associated with an edge corresponds with an analyst's subjective assessment of the likelihood that it exists. These weights do not represent edge values and are recorded as an edge attribute which is referenced in replication algorithms.

This chapter is focused on the process of identifying key individuals in dark networks that include known and suspected associations. The first process, referred to as Universal Likelihood Replication Analysis (ULRA), involves a less intensive approach where all suspected associations are assigned a universal likelihood of existing. A second process, referred to as Variable Likelihood Replication Analysis (VLRA), involves a more detailed approach where an analyst assigns separate likelihood estimates for each suspected association based their assessment of the reliability of specific intelligence reports and the nature of reported information. Both processes use replication to analyze various possible structures of a network as a means to generate a distribution of rankings for each node.

## 4.1 REPLICATION PROCESS

Replication involves the repetition of an experimental condition so the variability associated with a phenomenon can be estimated. Replication is used to randomly generate thousands of possible representations of a network, and to calculate individual centrality measures on each replicate. Centrality measures are then ordinally ranked and stored in a results matrix. In cases where nodes have identical measures, nodes are assigned the same rank. The end result is a mean ranking for each node, and a histogram of the number of times a node ranks in each position. Figure 9 shows a flow chart of the replication procedure.

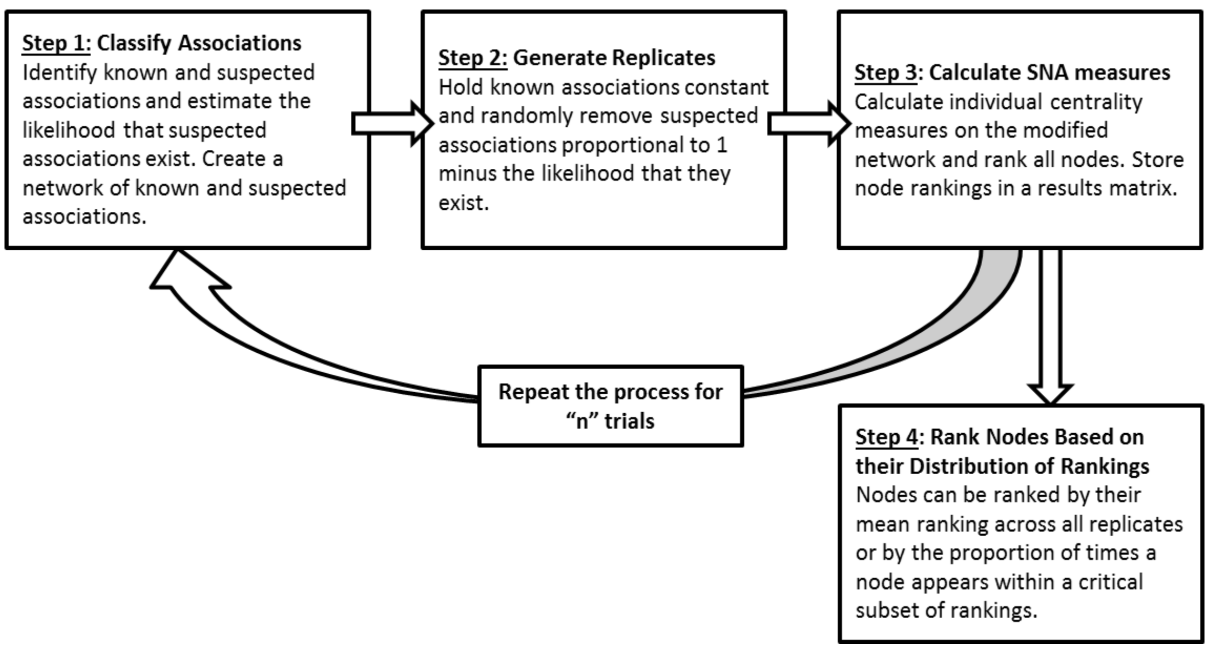



**Figure 9: Replication Procedure Flow Chart**

## 4.2 NODE PRIORITIZATION

Nodes are prioritized in each replicate by ordinal rankings as opposed to normalized centrality values. Ordinal rankings are used because they represent a node's relative importance in a given network, and they are discrete measures of importance that can easily be organized into bins. Also, when the node rankings of each replicate are stored in a results matrix, the distribution of rankings for each node can be displayed in a histogram by counting the number of times each node ranks in every position.

The results matrices are then used to prioritize nodes based on the distribution of rankings for each node. Since the goal of SNA is usually to identify a subset of the most critical actors in the network, nodes are ranked based on the proportion of replicates that a node is included in the subset of critical actors. An alternative would be to use mean node rankings across all replicates. However, this approach has some limitations since results can easily be skewed by outliers in the distribution of rankings.

### 4.4 CERTAINTY MODEL & VALUED NETWORK MODELS

Throughout this chapter replication results are compared with a binary network model (referred to as the "certainty model") and a valued network model in order to highlight how replication differs from existing analytical techniques. The certainty model is the most naïve of all models where all associations are treated equally and assigned binary edge values. Including the certainty model serves as a point of reference to highlight how results would vary if uncertainty is not incorporated into network analysis. The valued network attempts to incorporate uncertainty by assigning values to edges based on strength of a relations or the probability that interaction occurs between two members. Although this approach seems reasonable, the results in the following sections highlight how replication improves upon the valued network model.

### 4.5 UNIVERSAL LIKELIHOOD REPLICATION ANALYSIS (ULRA)

In networks that involve a large number of nodes and suspected associations, time or resources may not always available to individually assess the likelihood of each separate association. ULRA provides a means to conduct hasty sensitivity analysis on a network composed of confirmed and suspected associations to determine how uncertainty affects centrality rankings. When using the ULRA method, associations are organized into one of two categories, either confirmed or suspected. Edges that are classified as suspected are then assigned a universal likelihood of existence. This universal likelihood estimate determines the proportion of suspected edges that are randomly removed from the network when generating replicates. Thousands of replicates are generated and nodes are ranked by their mean rank across all replicates. The following section provides an overview of this process applied to a simple example network.

Using a universal likelihood assessment for suspected associations is useful because, it shows the relative impact of removing a specified percentage edges. In targeting applications, the ultimate goal is to accurately identify the most important individuals in a network. By analyzing varying levels of edge likelihood, ULRA identifies what levels of uncertainty would result in a change to the top ranked individuals. When models are based primarily on confirmed associations, it is expected that suspected edges will have little to no effect on the top ranked individuals. However, if models include a large proportion of suspected edges, then different levels of edge likelihood may produce different rankings of the most important individuals.

Below is an example network which is used to demonstrate how ULRA can be used to determine what edge likelihood levels would result in a change to the top ranked node (see Figure 10).

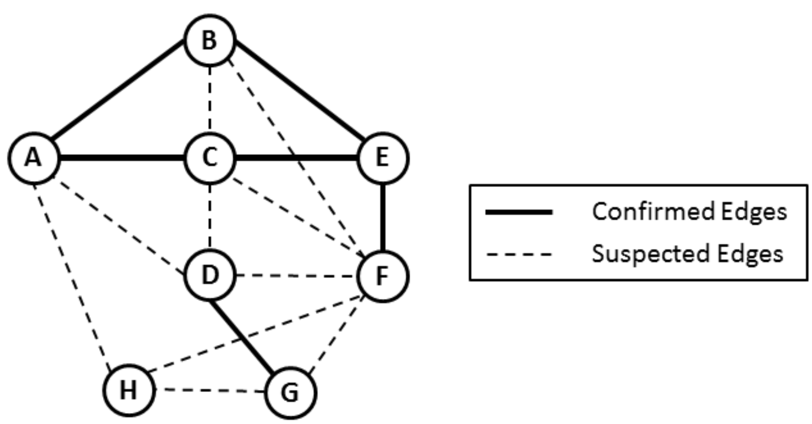



**Figure 10: Example Eight Node Network with Confirmed and Suspected Edges**

In this example, a proportion of edges are randomly removed from the network based on the overall likelihood of all suspected associations as a group. For instance, when edge likelihood is assessed to be 33%, 7 of the 10 suspected edges are randomly removed from each replicate. Nodes are then ordinaly ranked over 1000 replicates and prioritized based on the mean of centrality rankings across all replicates. Figure 11 shows sensitivity analysis for the universal likelihood replication method (ULRA) applied to degree centrality measures.

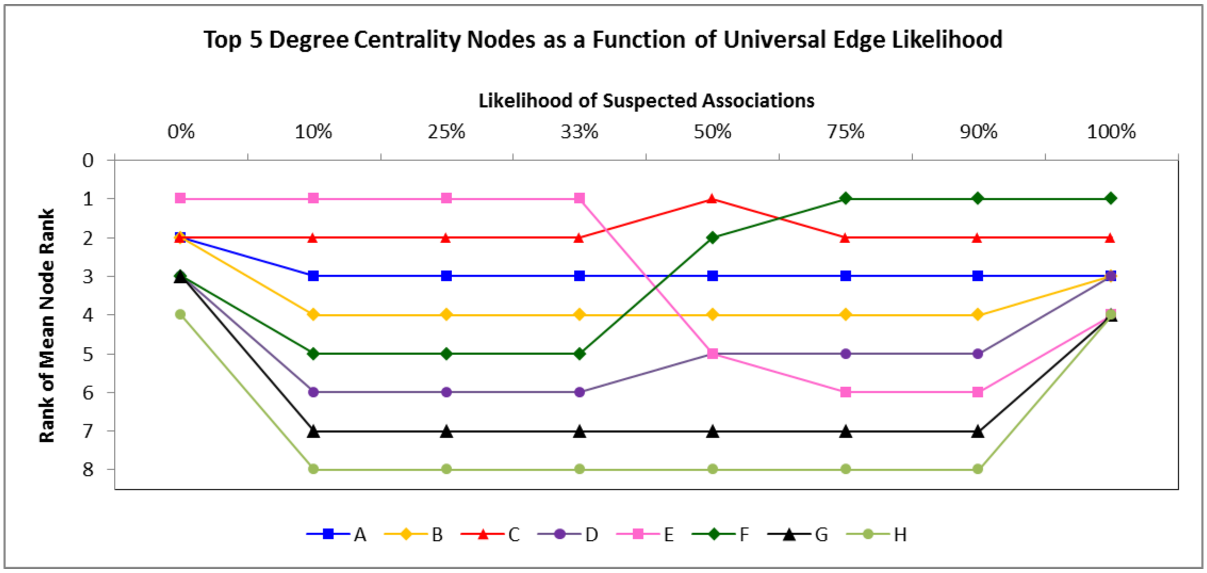



**Figure 11: ULRA Sensitivity Analysis for an Eight Node Example Network**

The results show that ULRA provides useful information about how uncertainty affects centrality measures. For instance, at the 0% likelihood level only confirmed edges are considered, and node E is considered the most important node. Whereas, at the 100% likelihood level all edges are considered confirmed, and node F is considered most important. Clearly, an analyst's recommendation is tied to their assessment of the likelihood of suspected associations. In this example involving only eight nodes, the decision is somewhat basic. If an analyst assesses that suspected associations are accurate more often than they are not (i.e. greater than 50% likelihood), then node F would be considered the most important.

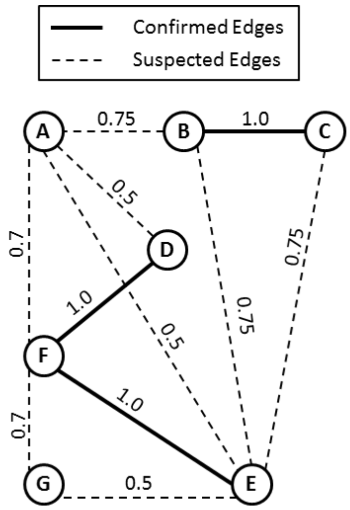
When considering other levels of edge likelihood at 33% and lower, node E is considered the top ranked node. This is expected because node E is initially confirmed to be adjacent to the greatest number of nodes in the network. When suspected edges are included in the network, it is possible for other nodes to outrank node E, but in the majority of replicates node E either outranks or is tied with other nodes. When replication is performed at the 50% level, node C and F are considered the most important nodes. This is expected because nodes C and F have the potential to be adjacent to five and six nodes respectively, and the largest degree attainable by any other node is four. Therefore, node C and F have a higher mean ranking because there are many possible scenarios where node C and F would have degrees ranging from four or six. Conversely, node E can only decrease or maintain its original rank because at most it will only be adjacent to three nodes. Finally at the 75% level and higher, node F is considered the most important edge.

There is no question that the inclusion or exclusion of edges in a network affects centrality measures. ULRA sensitivity allows an analyst to assess how the inclusion or exclusion of edges affect who is ranked as a priority target. The main limitation of this model is that uncertainty is not universal. Some intelligence sources are more reliable than others and some information is more substantive than others. ULRA does not make distinctions between likelihoods of individual suspected association. A more precise way to model the uncertainty of associations is to use variable likelihood replications analysis (VLRA) where each association is assigned an individual likelihood estimate.

**4.6 VARIABLE LIKELIHOOD REPLICATION ANALYSIS (VLRA)**

The following section discusses the logic behind the VLRA technique and demonstrates how VLRA expands upon existing techniques by more precisely incorporating information uncertainty and expressing confidence in SNA outputs.

Estimating what node is most active in Figure 12 is simple if every edge is assumed to have the same value and all edges are assumed to be equally likely (i.e. the certainty model). Degree centrality for binary edge values suggests that node E is the most active because it is adjacent to the largest number of nodes and therefore would have more opportunities for interaction. However, if some edges are certain and others are uncertain, then determining what node is most important is more complex. In the following demonstration, the numbers shown above each edge represent the likelihood of a binary edge existing in the VLRA model and they represent edge strength in the valued network model.




**Degree Centrality Rankings**

| Certainty Model | | | Valued Network Model | | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Binary Edge Values | | | Numeric Edge Values | | | Binary Values Randomly Assigned Based on Likelihoods | | | | |
| Node ID | Value | Rank | Node ID | Value | Rank | Node ID | Mean Rank | P(1) | P(2-3) | P(4-5) |
| E | 0.83 | 1 | E | 0.58 | 1 | E | 1.38 | 0.68 | 0.32 | 0.00 |
| A | 0.67 | 2 | F | 0.57 | 2 | F | 1.42 | 0.63 | 0.37 | 0.00 |
| F | 0.67 | 2 | B | 0.42 | 3 | A | 2.30 | 0.14 | 0.80 | 0.06 |
| B | 0.50 | 3 | A | 0.41 | 4 | B | 2.22 | 0.07 | 0.93 | 0.00 |
| C | 0.33 | 4 | C | 0.29 | 5 | C | 2.95 | 0.00 | 0.82 | 0.18 |
| D | 0.33 | 4 | D | 0.25 | 6 | D | 3.20 | 0.00 | 0.67 | 0.33 |
| G | 0.33 | 4 | G | 0.20 | 7 | G | 3.49 | 0.00 | 0.45 | 0.55 |

* P(1) = probability that a node is ranked 1st or tied for 1st
* P(2-3) = probability that a node is ranked 2nd or 3rd or tied for 2nd or 3rd
* P(4-5) = probability that a node is ranked 4th or 5th or tied for 4th or 5th

**Figure 12: VLRA Degree Centrality Results for a Seven Node Example Network**

When VLRA is used, nodes E, F, and A are the top three ranking nodes. The network graph shows that that node E is definitely connected to node F, probably connected to nodes B and C, and possibly connected to nodes A and G. In other words, node E is definitely connected to one node, probably connected to three nodes, and possibly connected to five nodes. Similarly, node F is definitely connected to two nodes and probably connected to four nodes. The question is what node is more important? Is the more important node, the node that is probably connected to three and possibly connected to five, or the node that is probably connected to four? The answer is dependent on the likelihoods of each association. In this scenario, node E is considered more important than node F. On the other hand, if it was later determined that the likelihood that node F was connected to nodes A and G was 0.8 instead of 0.7, then node F would be considered more important than node E even though node F is still only probably connected to four nodes.

When the valued network model is used, the results are almost identical to the VLRA model with the exception of the rankings for nodes A and B. The valued network model uses the expected degree of each node as the basis for ranking nodes. Therefore, node B outranks node A, because node B has an expected degree of 2.5 and node A has an expected degree of 2.45. In contrast, VLRA ranks nodes based on the number of times a node ranks as the top node over a series of experiments. VLRA suggests that node A ranks as the most important node or tied for the most important node in 14% of replicates compared to node B which ranks as the most important node or tied for the most important node in 7% of replicates. This difference in rankings between the two methods is based on the question that is being asked. If VLRA rankings were based on the mean rankings of each node, then the valued network model and VLRA model would produce identical results because node B has a mean ranking of 2.22 and node A has a mean ranking of 2.30. However, since the goal of network analysis is usually to determine the subset of actors that are most important, it is more appropriate to analyze the frequency that a node is included in the set of most important actors.

It is also important to consider the frequency that a node is included in the set of unimportant actors. For example, node A ranks as most important in 14% of replicates, but it is also ranks fourth in 6% of replicates. Node B on the other hand ranks third or higher in all replicates. Therefore, there is some assessed risk that node A is not important at all and there is no assessed risk that node B is not important. VLRA is useful because an analyst can express the expected payoff and risk associated with targeting actions in a manner that is easy to communicate to a decision maker. The alternative might be to use the numerical centrality scores for each node in the valued network (i.e. node A scores 0.42 and node B scores 0.41), but these measures cannot be translated to a 7% or 14% chance of being the most active actor in the network.

An additional benefit of using VLRA is that an analyst can express the uncertainty of their analysis with a histogram of rankings for each node. Figure 13 shows histograms of rank frequencies for nodes A, E, and F over 100 replicates.

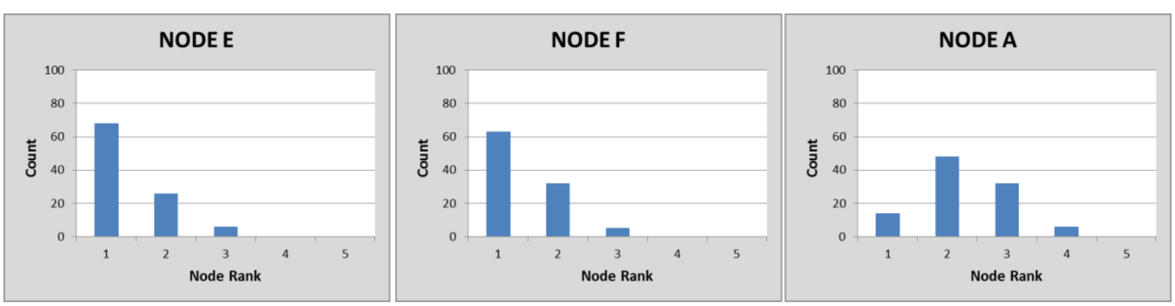
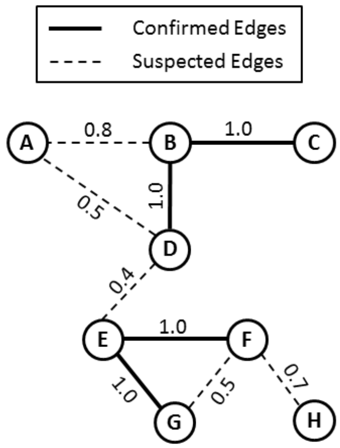



**Figure 13: VLRA Degree Centrality Histograms for a Seven Node Example Network**

These histograms visually show the range of rankings for each node and the rankings that occur most often. For instance, an analyst can see that there is a greater amount of variation in rank of node A. Node A is the most important or tied for the most important node in 14 of the replicates, but node A also ranks fourth in 6 replicates. The rank of node A is more uncertain because node A is only connected to other nodes through suspected associations. Therefore, node A could possibly have a degree of as low as zero or a degree as high as four. Conversely, there is less variance in the rank of node F, because it has the largest number of certain associations. Node F ranks 3rd or higher for all replicates.

Without knowing the true network, it is not possible to definitively say which model produces the most accurate rankings for degree centrality. There is very little difference in rankings between the valued network model and VLRA model for degree and eigenvector centrality rankings. Using the expected degree for each node is a reasonable way to proportionally emphasize edges based on the strength of associations, and it requires less computational effort than generating replicates. So at first, it does not appear that VLRA would significantly improve the bottom line for targeting recommendations for degree and eigenvector centrality. However, when the goal is to identify which nodes are closest to all other nodes or the goal is to identify which nodes have the most importance as a liaison or bridge, VLRA can produce extremely different results compared to existing models.

Estimating what node has the largest significance as a liaison or bridge in Figure 14 is easy if the certainty model is used. In certainty model, node D is most important because it lies on the most geodesic paths between nodes and the network is fully connected. However, if some edges are certain and others are uncertain, then it is possible for the network to be disconnected which makes it more difficult to determining what node is most important.




**Betweenness Centrality Rankings**

| Certainty Model | | | Valued Network Model | | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Binary Edge Values | | | Numeric Edge Values | | | Binary Values Randomly Assigned Based on Likelihoods | | | | |
| **Node ID** | **Value** | **Rank** | **Node ID** | **Value** | **Rank** | **Node ID** | **Mean Rank** | **P(1)** | **P(2)** | **P(3-5)** |
| **D** | 0.57 | 1 | **E** | 0.62 | 1 | **B** | 1.71 | 0.56 | 0.17 | 0.27 |
| **E** | 0.57 | 1 | **D** | 0.57 | 2 | **E** | 1.70 | 0.42 | 0.46 | 0.12 |
| **B** | 0.29 | 2 | **B** | 0.29 | 3 | **F** | 2.27 | 0.30 | 0.27 | 0.43 |
| **F** | 0.29 | 2 | **F** | 0.29 | 4 | **D** | 2.13 | 0.18 | 0.51 | 0.31 |
| **A** | 0.0 | 3 | **A** | 0.0 | 5 | **A** | 3.05 | 0.00 | 0.33 | 0.67 |
| **G** | 0.0 | 3 | **G** | 0.0 | 5 | **C** | 3.05 | 0.00 | 0.33 | 0.67 |
| **H** | 0.0 | 3 | **H** | 0.0 | 5 | **G** | 3.05 | 0.00 | 0.33 | 0.67 |

* P(1) = probability that a node is ranked $1^{st}$ or tied for $1^{st}$
* P(2) = probability that a node is ranked $2^{nd}$ or tied for $2^{nd}$
* P(3-5) = probability that a node is ranked $3^{rd}$ to $5^{th}$ or tied for $3^{rd}$ to $5^{th}$

**Figure 14: VLRA Betweenness Centrality Results for an Eight Node Example Network**

When VLRA is used, node B is ranked as most important or tied for most important. It can be seen that node B is definitely the bridge between nodes C and D and nodes A and C, and is possibly a bridge between node C and the nodes in the bottom subgroup. Similarly, node E is possibly a bridge that connects all nodes of the top and bottom subgroups. The question is what node is more important for bridging the network together? When the two subgroups in the network are connected (i.e. the E-D edge exists), node E is the most important node in the network in all replicates. However, in 60% of the possible scenarios the network is disconnected (i.e. the E-D edge does not exist). In a scenario of a disconnected network, node B is always more important than node E. Therefore, VLRA highlights that the importance of node E is directly related to the likelihood of the E-D edge. The VLRA histograms for 100 replicates reflect this unique situation related to how node B ranks in the network and is illustrated in Figure 15.

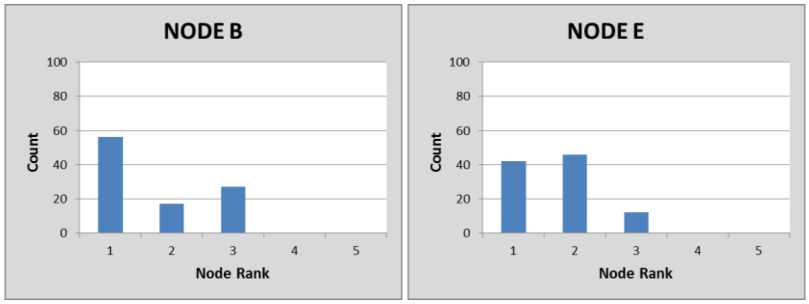



**Figure 15: VLRA Betweenness Centrality Histograms for an Eight Node Example Network**

The histogram of node B shows that node B is the top ranked node or tied for top ranked node in 56% of replicates.  However, targeting node B also involves some risk as it ranks third or tied for third in 27% of replicates.  The variance in ranking for node B is a result of the potential importance of node E.  Conversely, node E ranks as the top ranked node or tied for the top ranked node in 42% of replicates and is the third ranked node or tied for the third ranked node in only 12% of replicates.  In this scenario determining what node to target requires some interpretation of the network graph and interpretation of the distribution of possible rankings.  One could argue that node E should be the top priority because removing it would guarantee that the network is fragmented into two groups.  However, if the E-D edge doesn't exist, then targeting node B would fragment the network into three groups by isolating C from the network. In the end, the decision hinges on the likelihood of the E-D edge.

The results of this simple experiment demonstrate why the valued network model is limited in its ability to incorporate uncertainty into betweenness centrality rankings and suggest that VLRA is a more accurate model.  The valued network model only considers a scenario of a fully connected network.  Likewise, representing the network as fully connected based on an association that is only 40% likely does not adequately model the complexity of the situation.  VLRA is a stronger model because it is able to account for rankings across a broad range of possible network structures.

# CHAPTER 5 ANALYSIS OF FM 3-24 NETWORK

The following section provides Universal Likelihood Replication Analysis (ULRA) and Variable Likelihood Replication Analysis (VLRA) applied to degree, closeness, and eigenvector centrality measures for an example of a network found in chapter FM3-24.

### 5.1 OVERVIEW OF THE FM 3-24 NETWORK

Figure 16 provides a doctrinal example of a 16-node association matrix consisting of 24 confirmed, 21 probable, and 28 possible associations that can be found in FM 3-24 (p. B-17). It is important to note that the doctrinal example only presents the lower triangle of a matrix and does not numerically represent entries in the association matrix. In order to use this matrix for centrality calculations, binary edge values are used to represent that an association either is present or absent. The triangular representation of the association matrix is converted to a symmetric, square matrix by assuming that associations are undirected.

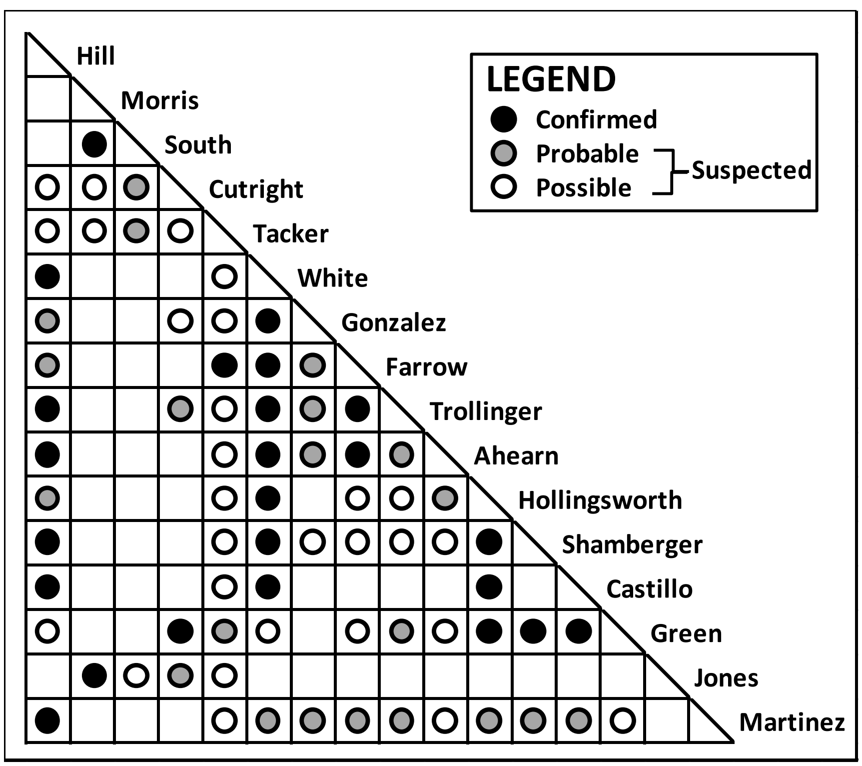



**Figure 16: FM 3-24 Doctrinal Example of Association Matrix**

Since FM 3-24 does not provide the supporting intelligence reports used to create the doctrinal example, likelihoods were randomly assigned based on the confidence levels associated with each type of association as described in Figure 8. These randomly assigned likelihoods are intended to represent the likelihood estimates of an analyst performing social network analysis (SNA). Known associations are assigned likelihoods of 1.0, probable associations are randomly assigned likelihoods between 0.6 and 0.9, and possible associations are assigned likelihoods between 0.4 and 0.6. Figure 17 displays a network graph of the FM 3-24 network, and a table of edge-likelihoods for each association can be found in Appendix A.

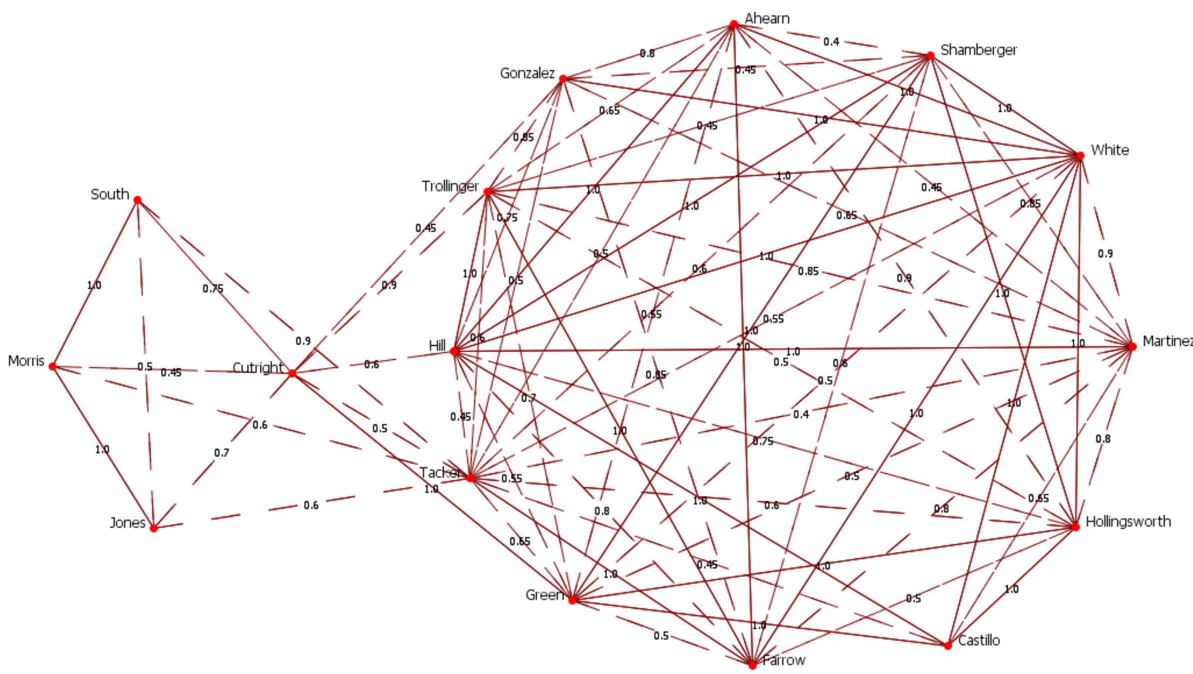



**Figure 17: Network Graph of FM 3-24 Association Matrix**

### 5.2 FM 3-24 ULRA FINDINGS

Graphs of ULRA sensitivity analysis for varying levels of edge likelihood are provided in Appendix B. The following section provides a summary of the key findings related to the graphs found in Appendix B.

### 5.2.1 DEGREE CENTRALITY ULRA SENSITIVITY ANALYSIS

A change to the top three ranking nodes occurs at the 50% confidence level. When considering the levels of edge likelihood in Appendix B, there are two possible groups of actors that rank in the top three: {White, Hill, Green} and {White, Hill, Tacker}. White and Hill appear in both sets, so White and Hill can be considered important for all possible scenarios. Green does not appear in all sets, but consistently ranks third or fourth across all scenarios. The reason why the rankings for White, Hill, and Green are resilient to the inclusion of suspected edges is because they have the largest number of confirmed associations. The degree of a node is an individual measure that does not depend on the structure of adjacent nodes, so the expected value of a node's degree centrality is directly related to the number of edges tied to a node and the likelihood that they exist. Therefore, nodes that have a large number of suspected edges have the greatest potential to increase in importance as the likelihood of suspected edges increases. For instance, Tacker is ranked as the least important individual at the 0% confidence level but enters the top three rankings at the 50% confidence level and becomes the most important individual at the 75% confidence levels. This behavior is expected since Tacker has the largest number of suspected associations (13 out of 15). As increasingly larger percentages of suspected edges are included in the network, Tacker increases in importance. ULRA sensitivity analysis allows an analyst to identify what edge likelihood levels would result in a change to the top three rankings and it allows an analyst to see how the top three rankings vary across a broad range of scenarios.

### 5.2.2 CLOSENESS CENTRALITY ULRA SENSITIVITY ANALYSIS

A change to the top three ranking nodes occurs at the 25%, 50%, and 75% confidence levels. When considering the levels of edge likelihood in Appendix B, there are four possible groups of actors that rank in the top three: {White, Hill, Shamberger}, {White, Hill, Green}, {Tacker, Hill, White}, and {Tacker, Hill, Green}. Hill is the only individual in all four sets, so Hill can be considered important for all possible scenarios. Green appears in the top three sets at the 25%, 33%, 75%, 90%, and 100% levels. It may seem unexpected for Green to appear in the top three at lower levels of edge likelihood since he has 11 total associations with only 36% (4 out of 11) of them being "confirmed". However, closeness centrality measurements depend on the overall structure of the entire network, so inclusion of nodes not tied to Green can affect his importance. ULRA sensitivity analysis suggests that if even 25% (13 of the 49) of the suspected associations in the overall network exist, then Green has a high probability of being a very important individual.

### 5.2.3 BETWEENNESS CENTRALITY ULRA SENSITIVITY ANALYSIS

A change to the top three ranking nodes occurs at the 10%, 25%, 50%, and 75% confidence levels. When considering the levels of edge likelihood in Appendix B, there are five possible groups of actors that rank in the top three: {White, Hill, Farrow}, {White, Green, Hill}, {White, Green, Tacker}, {Tacker, Cutright, Green}, and {Tacker, Cutright, Hill}. There is a great deal of variance in the set of the top three ranked individuals across the varying levels of likelihood, so it is more difficult to determine what individuals are most important. The large amount of variance for betweenness measurements is most likely due to the fact that betweenness centrality measures the extent to which an individual serves as an intermediary, liaison, or bridge between groups. Therefore in a disconnected network, any edge that ties a node from one group to a node in a disconnected group will immediately increase in importance. Similarly, when groups are connected by a single intermediary, the addition of an edge that creates an alternate intermediary can cause the original intermediary to decrease in importance. In the FM 3-24 network, when all suspected edges are omitted (i.e. 0% confidence level), the network consists of two subgroups. Jones, Morris, and South have no confirmed associations with the rest of the network. Tacker has a suspected association with all three individuals, so he has a large potential to be extremely important as a bridge in the network. ULRA sensitivity analysis suggests that if even 25% (13 of the 49) of the suspected associations in the overall network exist, then Tacker has a high probability of being a very important individual. At the 33% confidence level and higher, Tacker is ranked as the most important individual.

### 5.2.4 EIGENVECTOR CENTRALITY ULRA SENSITIVITY ANALYSIS:

A change to the top three ranking nodes occurs at the 75% confidence levels. When considering the levels of edge likelihood in Appendix B, there are two possible groups of actors that rank in the top three: {White, Hill, Shamberger} and {White, Hill, Tacker}. White and Hill appear in all sets, so they can be considered important in all scenarios. At the 75% level, Tacker replaces Shamberger in the top three rankings. The limited variability in eigenvector rankings is most likely because eigenvector centrality is calculated by proportionally weighting adjacent nodes based on their degree centrality. Since degree centrality is directly related to the number of edges tied to a node and the likelihood that they exist, it follows that for replications that use low edge likelihood confidence levels, eigenvector centrality will be higher for nodes that are connected to nodes with a large number of confirmed association. White and Hill have the most confirmed associations in the network (8 confirmed for White; 6 confirmed for Hill), and they are also confirmed to be associated with each other. Therefore, their eigenvector rank is very resilient to the inclusion of suspected edges, and they rank in the top three for degree centrality for all levels of edge likelihood. Shamberger does not stand out as very important in other centrality measures, but he is definitely associated with White, Hill. Therefore, Shamberger's rank is resilient to the inclusion of suspected edges because he is connected to two individuals who rank in the top three for degree centrality across all confidence levels. Tacker enters the top three ranking at the 75% level because he has a possible association with every individual in the network. Therefore, with each replicate, Tacker associates with more individuals, many of whom are also increasing in degree centrality.

### 5.2.5 SUMMARY OF ULRA SENSITIVITY ANALYSIS:

All four centrality measures are sensitive to the uncertainty of information because no measure produces a consistent group of the top three ranking individuals for all levels of edge likelihood. Of the four measures, betweenness centrality has the greatest variance in the top three ranking individuals. In general, White and Hill stand out as important individual based on degree, closeness, and eigenvector centrality measures across all levels of edge likelihood. White and Hill's consistent ranking and resilience to the inclusion of suspected associations is due to their number of confirmed associations. Tacker's importance in the network is the most uncertain, because he ranks outside the top 5 for low edge likelihoods and rapidly moves up in rank as edge likelihood increases. This is expected because Tacker is associated with every individual in the network which would make him the central figure as an information hub and liaison if all of his reported associations were valid. However, since 12 of his 15 associations are categorized as "possible" and only 1 of his 15 associations is categorized as confirmed, Tacker's importance in the network is tied very closely to edge likelihood.

ULRA sensitivity analysis is useful because it allows an analyst to see how varying levels of edge likelihood affect who is ranked as a priority target. By observing the amount of variance in rankings for numerous possible structures, an analyst can better understand and express the uncertainty associated with the set of top ranking individuals. The main limitation of ULRA is that determining the most accurate level of edge likelihood is difficult because all suspected associations are treated equally. There is no distinction between associations that are considered probable and associations that are considered possible. VLRA which assesses likelihoods for each individual edge is a preferred way to be more precise in how edges are proportionally emphasized in analysis. Nevertheless, in networks where an analyst does not have the time, resources, or ability to individually assess edge likelihoods, ULRA sensitivity analysis is a reasonable way to take into account various possible representations of a network when conducting SNA.

**5.3 FM 3-24 VLRA FINDINGS**

The following section provides a summary of the key findings pertaining to VLRA applied to a network that considers confirmed, probable, and possible associations (See Appendix C) and VLRA applied to a network that considers only confirmed and probable associations (See Appendix D). For simplicity, the network of confirmed, probable, and possible associations will be referred to as scenario 1 and the network of confirmed and probable associations will be referred to as scenario 2. The ratio of suspected to total edges in scenario 1 is 67% (24 – confirmed, 21 – probable; 28 – possible), and the ratio of suspected to total edges in scenario 2 is 47% (24 – confirmed; 28 – probable). These two scenarios are presented to illustrate how VLRA performs under various conditions of uncertainty and to demonstrate how VLRA can be modified for different targeting applications. For instance, if the goal is to identify which individuals should be targeted for further intelligence collection, then a network that includes all types of associations might be appropriate to focus attention on individuals with the greatest potential to be important. Whereas, if the goal is to identify what individuals should be targeted for detention, then a network that includes only probable and confirmed associations might be appropriate to identify the individuals with the greatest probable importance.

### 5.3.1 DEGREE CENTRALITY VLRA

White, Hill, and Tacker are estimated to be in the top three rankings for scenario 1 with 99.2%, 96.0%, 74.6% confidence, and White, Hill, and Trollinger are estimated to be in the top three rankings for scenario 2 with 100%, 99.7%, and 84.9% confidence. These results suggest that White and Hill are the best individuals to target based on both probable and potential importance. Conversely, Trollinger probably ranks in the top three, but Tacker has potential to be included in the top three ranked individuals if low likelihood associations are considered.

All three models identify identical sets of the top three ranking individuals in each scenario, so it does not appear that using the VLRA model affects the bottom line targeting recommendation. However, using VLRA is arguably a more precise model because it allows an analyst to quantify a distribution of rankings for each targeting recommendation.

### 5.3.2 CLOSENESS CENTRALITY VLRA

Tacker, Hill, and White are estimated to be in the top three rankings for scenario 1 with 88.6%, 77.8%, 68.7% confidence, and White, Trollinger, and Farrow are estimated to be in the top three rankings for scenario 2 with 99.5%, 94%, and 90.3% confidence. White is the only individual who is assessed to be in the top three on the basis of both probable and potential importance. Trollinger and Farrow are ranked in the top three based on probable importance, but Tacker and Hill are potentially more important than Trollinger and Farrow.

The VLRA results applied to closeness centrality differ greatly in comparison to the valued network model. The valued network model identifies Tacker, Martinez, and Cutright as the top three individuals in scenario 1, but the VLRA model assesses that Martinez and Cutright, would rank in the top three in only 25.1% and 10.9% of possible scenarios. The valued network model rankings appear to be pretty far away from confirmed information about the network. Martinez and Cutright are pendant nodes in the network graph of only confirmed associations, so they would need to possess a great deal of potential importance to outrank nodes that have more probable importance. Martinez has limited potential to be an important hub, because he is not adjacent to any of the individuals in the top subgroup. Similarly, Cutright is suspected to be close to every individual in the smaller top subgroup, but he is only adjacent to four of the individuals in the bottom network. Interestingly, Tacker, who is the highest ranking node in the VLRA model, is also a pendant node in the graph of only confirmed associations. The reason why Tacker ranks so high is because unlike Martinez and Cutright, Tacker's potential importance outweighs the probable importance of all other nodes. Tacker is reported to be possibly connected to every individual in the entire network. Therefore, if even half of Tacker's reported associations are valid, no other individual would be more important than him.

In scenario 2, the valued network model also produces different results than the VLRA model. The valued network model identifies Trollinger, Green, and Martinez as the top three individuals, but the VLRA model assesses that Green and Martinez would rank in the top three in only 63.6% and 30.9% of possible scenarios. Based on Green and Martinez' location in the network graph of confirmed associations, it appears that the VLRA recommendations are closer to the confirmed structure of the network. White is clearly the most central individual in the network because he is adjacent to eight of the 16 individuals and therefore can reach 50% of the network in one step. Since Trollinger and Farrow are confirmed to be adjacent to White, in theory they will always be able to take advantage of the efficiencies of communicating through White. In contrast, Green and Martinez are not confirmed to be adjacent to White, so they would only be able to take advantage of efficiencies gained by communicating through the White in a small proportion of the possible network structures.

It is also interesting to note that top three ranking individuals of the VLRA model are contained in the set of top three ranking individuals in the certainty model, so it is possible to conclude that the same VLRA results could be obtained by simply using the certainty model. However, the main difference in the performance of the two models is that the VLRA model is more discriminating for the relative importance of each node. For instance, the certainty model lists four individuals as being tied for the second most important node in scenario 2, so the critical set for the certainty model actually consists of five individuals instead of three. It is not possible to definitively say which three individuals are most important because four nodes share the same rank. In comparison, the VLRA model allows an analyst to express confidence estimates for the likelihood that each individual ranks in the top three. Using the VLRA model enables an analyst to compare

Farrow, who ranks in the top three with 90.3% confidence, with Green who ranks in the top three with 63.6% confidence.

The VLRA model appears to be the most precise model because it is discriminating and provides an objective solution that is reasonable close to confirmed information. There are clear differences between the valued network and VLRA closeness centrality rankings. The valued network appears to overemphasize suspected associations and identifies nodes that would only rank in the top three in a small fraction of possible network representations. The differences between the certainty model and VLRA model are somewhat subtle, but the main limitation of the certainty model is its inability to discriminate between nodes that share the same rank. For these reasons the VLRA model appears to be a better performing model.

### 5.3.3 BETWEENESS CENTRALITY VLRA

Tacker, Cutright, and Hill are estimated to be in the top three rankings for scenario 1 with 97.1%, 69.7%, 42.5% confidence, and Cutright, Trollinger, and Green are estimated to be in the top three rankings for scenario 2 with 78.2%, 67.1%, and 60.5% confidence. Cutright is the only individual who is assessed to be in the top three based on his probable and potential importance. Trollinger and Green are probably in the top three for closeness centrality rankings, but Tacker and Hill are potentially more important than Trollinger and Green.

Interpreting the VLRA results in both targeting scenarios is difficult for the FM 3-24 network because it is unknown if the network is fully connected or disconnected. For this reason some nodes can have a great deal of importance in the network if the network is fully connected or the same nodes might have little to no importance if the network is disconnected. This relationship can easily be observed in the scenario 2 by examining the bi-modal structure of the histograms for South, Jones, and Morris. For example, South's most frequent rank is fifth (1,194 out of 5000 replicates), but he also ranks first in 18.8% of replicates (904 out of 5000). The bi-modal nature of rankings is somewhat discouraging from the standpoint of being able to definitively determine which nodes are most important. However, from a modeling standpoint, VLRA histograms appear to tell a more complete story about the importance of each node. If the certainty or valued network models are used, an analyst is left with a simple ordinal ranking of nodes and these simple rankings fail to capture the range or frequency of possible ranks for each node.

As was the case with closeness centrality rankings, the VLRA results for betweenness centrality rankings differ in comparison to the valued network model for the scenario 1. The valued network model identifies Tacker, Cutright, and White as the top three individuals, but the VLRA model assesses that White would rank in the top three in only 12.6% of possible scenarios. The most likely reason why White is not ranked higher in the VLRA model is because White is only possibly connected to Tacker, but Hill is possibly connected to both Cutright and Tacker. Hill's possible association with Cutright and Tacker in turn makes Hill potentially important as an intermediary to gain access to Cutright and Tacker. For these reasons, it appears that the VLRA model rankings are more consistent with the confirmed structure of the network.

When VLRA is applied to the network in scenario 2, the valued network model also produces different results than the VLRA model. The valued network model identifies Cutright, Farrow, and South as the top three individuals, but the VLRA model assesses that Farrow and South would rank in the top three in only 36.0% and 29.0% of possible replicates. In the end determining the rank of Farrow and South is a matter of examining the confirmed and probable importance of each individual. Trollinger and Green have more confirmed importance than Farrow and South and they both have potential for increased importance. Farrow is similar in that he has confirmed and potential importance, but Trollinger and Green would outrank Farrow in a greater proportion of the possible networks. South has no confirmed importance, and his importance is directly related to his associations with Cutright and Tacker. In short, it appears as though the valued Network overemphasizes the probable importance of individuals.

The most interesting finding is that the VLRA and certainty models produce almost identical results for the top six ranking individuals in the network. Unlike scenario 1, the certainty model rankings are discriminating in scenario 2 and they are very close to the VLRA model. In other words, the random inclusion of probable associations produces almost identical ranking as including all probable associations. However, closer inspection of the VLRA distribution of rankings demonstrates how generating a distribution of rankings is more useful than generating an ordinal ranking. The certainty model includes Farrow in the set of top three ranking individuals, and Green is assessed as less important than Farrow. However, the VLRA model assesses that Green is in the top three ranking with 60.5% confidence and Farrow is in the top three ranking with 36.0% confidence. The VLRA model assesses Green is almost twice as likely as Farrow to be ranked in the top three individuals.

In summary, the VLRA model appears to be the strongest model because it is the only model that clearly accounts for a node's importance in disconnected and fully connected networks. Betweenness centrality is a measure that can be very sensitive to inclusion of key bridging associations. The certainty and valued network model fail to account for scenarios of a disconnected network. Ranking nodes based on the assumption that a network is fully connected is not accurate even if key bridging associations are assessed to be weak associations. The VLRA model is the only model that accurately models the existence or absence of key bridging associations.

### 5.3.4 EIGENVECTOR CENTRALITY VLRA

White, Hill, and Trollinger are estimated to be in the top three rankings for scenario 1 with 87.4%, 76.4%, 29.6% confidence, and White, Hill, and Martinez are estimated to be in the top three rankings for scenario 2 with 100%, 98.1%, and 46.8% confidence. These results suggest that White and Hill are the best individuals to target based on both probable and potential importance. Conversely, Martinez is probably ranks in the top three most important individuals, but Trollinger has potential to be important if low likelihood associations are considered.

All three SNA models identify the same set of top ranking individuals, so it does not appear that using VLRA affects the bottom line targeting recommendation. However, using VLRA is arguably a more precise model because it allows an analyst to convey the certainty of targeting recommendations.

### 5.3.5 SUMMARY OF VLRA FINDINGS

In the context of the FM 3-24 network, the valued network model appears to be a reasonable technique for degree and eigenvector centrality measures, the certainty model appears to be a reasonable technique for betweenness and closeness centrality measures, and the VLRA model appears to be a reasonable technique for all four centrality measures.  The VLRA model produces identical results for degree and eigenvector centrality rankings when compared to the valued network model, and it produces almost identical results for closeness and betweenness centrality rankings when compared to the certainty model.  On the surface, it does not appear that VLRA greatly improves the bottom line targeting recommendations in comparison to existing techniques.  However, VLRA is the only model that quantifies uncertainty associated with targeting recommendations.  The VLRA model is arguable a superior model because it can be applied to all centrality measures and it produces results that proportionally emphasize associations based on the likelihood that they exist.  VLRA rankings consider a node's potential for importance while simultaneously accounting for a node's confirmed importance.

# CHAPTER 6 CONCLUSION

Conceptually, social network analysis (SNA) centrality measures are fairly easy to comprehend and easy to calculate using basic linear algebra. In many cases where a network is composed of a small number of individuals, identifying the most central or the most important bridges in a network can be identified through visual inspection of a network and does not require mathematical calculations. An analyst probably does not need to calculate the column or row sums of an association matrix to determine what individuals are most active if a network graph clearly highlights that some individuals are adjacent to a greater proportion of individuals than others. The value of current SNA techniques is that they provide a way to cue an analyst to important individuals in complex networks where important individuals cannot be easily identified through visual inspection. However, when a network graph is uncertain, determining who is most important is difficult in even small networks.

Current SNA techniques are limited because they do not adequately account for the uncertainty of information and they do not quantify the uncertainty associated with SNA outputs. It is proven that network analysis can be skewed by the improper inclusion or exclusion of individuals and associations. Basing targeting recommendations on a single representation of a network with unknown structure causes SNA outputs to be heavily dependent on how accurately an analyst identifies the true network. Therefore, the most common way to mitigate targeting individuals based on uncertain information is to be deliberate in what nodes and edges are included in the network graph. Using only known and probable associations may seem reasonable, but this approach may lead an analyst to omit key valid associations simply because they are uncertain. In the context of dark networks where individuals attempt to conceal their activities, there is a decent chance that many associations will not be able to be confirmed. Additionally, current SNA techniques only provide an ordinal ranking of individuals and do not consider the range or frequency of rankings for each node.

Replication improves upon existing SNA techniques because of its ability to account for numerous possible representations of a network. Furthermore, replication enables an analyst to express the uncertainty associated with centrality-based targeting recommendations by generating a distribution of rankings for each individual. The main difference between the analysis presented in this research and previous replication research is that edge removal is not entirely random in this research. The Universal Likelihood Replication Analysis (ULRA) and Variable Likelihood Replication Analysis (VLRA) models are both based on the concept that known associations are held constant in the replication process, and only suspected edges are randomly removed from the network. By holding known associations constant, ULRA and VLRA achieve the desired end state of emphasizing known associations more than suspected associations.

ULRA sensitivity analysis improves upon existing techniques by allowing an analyst to assess how the inclusion of suspected edges might affect SNA rankings for varying levels of edge likelihood. The main limitation of ULRA is that all suspected associations are treated equally, so it does not account for the fact that some suspected associations may be more reliable than others. However, if an analyst does not have the time, resources, or ability to individually assess edge likelihoods, ULRA sensitivity analysis is a reasonable way to take into account various possible representations of a network when conducting SNA.

VLRA which assesses likelihoods for each individual edge is a preferred way to be more precise in how edges are proportionally emphasized in analysis. By individually assessing likelihoods, VLRA includes the most likely associations in the greatest proportion of replicates. As a result VLRA rankings represent which individuals are most important based on their confirmed and potential importance. The most interesting finding of this research is that VLRA model produces identical results for degree and eigenvector centrality rankings when compared to the valued network model, and it produces almost identical results for closeness and betweenness centrality rankings when compared to the certainty model. This finding may be a random occurrence specific to the FM 3-24 network, but this study suggests that the valued network model may a better model for degree and eigenvector centrality and the certainty model may a better model for betweenness and closeness centrality.

Even though the VLRA model did not provide considerably different results compared to some existing models, the VLRA model is arguably the most conceptually sound of all of the methods and it is flexible enough to be applied to the analysis of all four centrality measures.

Ultimately, the VLRA model achieves the goal of effectively accounting for the uncertainty of information and the goal of quantifying the uncertainty of SNA outputs.

# CHAPTER 7 FUTURE WORK

The replication techniques presented in this study were only applied to a single fictional network. The FM 3-24 network was chosen primarily because the research objective was to create a model that incorporates analytical principles and concepts found in Army doctrine. It was difficult to find other networks composed of confirmed, probable, and possible associations, so the FM 3-24 network was used simply because it appeared to have the characteristics of a dark network, and it was readily available.

This research could be improved by testing these techniques on additional networks, on larger networks, and on networks with varying proportions of suspected edges. The VLRA model will always be useful because it improves upon existing models by quantifying the certainty of social network analysis (SNA) outputs. However, the benefit of using VLRA might be negligible if it does not affect the bottom line targeting recommendations. If the VLRA model produces similar results in other network, then the simpler course of action might be to use a valued network for degree and eigenvector centrality rankings and convert valued edges to binary edges when analyzing betweenness and closeness.

## REFERENCES


*Army Strategic Planning Guidance* [ASPG], (Headquarters, Department of Defense, Washington, DC, 7 February 2013). Retrieved from http://usarmy.vo.llnwd.net/e2/rv5_downloads/info/references/army_strategic_planning_guidance.pdf

Bonacich, P. (1972). Factoring and Weighting Approaches to Status Scores and Clique Identification. *The Journal of Mathematical Sociology*, *2*(1), 113–120.

Borgatti, S. P., Brass, D. J., Mehra, A., & Giuseppe, L. (2009). Network Analysis in the Social Sciences. *Science, 323*, 892-895.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006, May). On the Robustness of Centrality Measures under Conditions of Imperfect Data. *Social Networks, 28*(2), 124-136.

Borgatti, S. P., Martin, E. G., & Johnson, J. C. (2013). *Analyzing Social Networks.* Thousand Oaks, CA: SAGE Publications Inc.

Breiger, R. L. (1990). *Social Mobility and Social Structure*. Cambridge, England: Cambridge University Press.

*Circular reporting*. (2013, September 6). Retrieved September 6, 2013, from wikipedia: http://en.wikipedia.org/wiki/Circular_reporting

Doreian, P. (1981). Estimating Linear Models with Spatially Distributed Data. *Sociological Methodology*, *12*(1981), 359–388.

Everton, S. F. (2012). *Disrupting Dark Networks.* New York: Cambridge University Press.

Field Manual [FM] 2-22.3, *Human Intelligence Collector Operations,* (Headquarters, Department of the Army, Washington, DC, September 2006).

Field Manual [FM] 3-24, *Counterinsurgency,* (Headquarters, Department of the Army, Washington, DC, December 2006).

Fischer, C. S. (1982). *To Dwell Among Friends: Personal Networks in Town and City*. Chicago, IL: University of Chicago Press.

Friedkin, N. E., & Cook, K. S. (1990). Peer Group Influence. *Sociological Methods & Research*, *19*(1), 122–143.

Friedkin, Noah E. (1986). A Formal Theory of Social Power. *Journal of Mathematical Sociology1*, *12*(2), 103–126.

Joint Publication [JP] 2-0, *Joint Intelligence,* (Headquarters, Department of Defense, Washington, DC, 22 June 2007).

Joint Publication [JP] 2-01, *Joint and National Intelligence Support to Military Operations,* (Headquarters, Department of Defense, Washington, DC, 05 January 2012).

Krebs, V. (2002). Uncloaking Terrorist Networks. *First Monday, 7*(4). Retrieved from http://journals.uic.edu/ojs/index.php/fm/article/view/941/863

Laumann, E. O., Marsden, P. V, & Galaskiewicz, J. (1977). Community-Elite Influence Structures : Extension of a Network Approach. *American Journal of Sociology*, *83*(3), 594–631.

Laumann, E. O., & Pappi, F. U. (1973). New Directions in the Study of Community Elites. *American Sociological Review*, *38*(2), 212–230.

MIT Strategic Engineering. (2011). Matlab Tools for Network Analysis (2006-2001). Retrieved February 12, 2014, from http://strategic.mit.edu/downloads.php?page=matlab_networks

Nemeth, R. J., & Smith, D. A. (1985). International Trade and World-System Structure: A Multiple Network Analysis. *Reveiw (Fernand Braudel Center)*, *8*(4), 517–560.

Perron, O. (1907). Zur Theorie der Matrices. *Mathematische Annalen1*, *64*(2), 248–263.

*Report of the Defense Science Board on Understanding Human Dynamics,* (Defense Science Board, Washington, DC, March 2009). Retrieved from http://www.acq.osd.mil/dsb/reports/ADA495025.pdf

Rubinov, M. (2010). Betweenness_wei.m. Retrieved February 12, 2014, from http://visualconnectome.googlecode.com/svn-history/r2/trunk/Plugins/BCT/

Snyder, D., & Kick, E. L. (1979). Structural Position in the World System and Economic Growth , 1955-1970 : A Multiple-Network Analysis of Transnational Interactions. *American Journal of Sociology*, *84*(5), 1096–1126.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Explanations.* New York, NY: Cambridge University Press.

Wei, W., Pfeffer, J., Reminga, J., & Carley, K. M. (2011). Handling Weighted, Asymmetric, Self-Looped and Disconnected Networks in ORA. Carnegie Mellon University, SCS/ISR, Technical Report, CMU-ISR-11-113

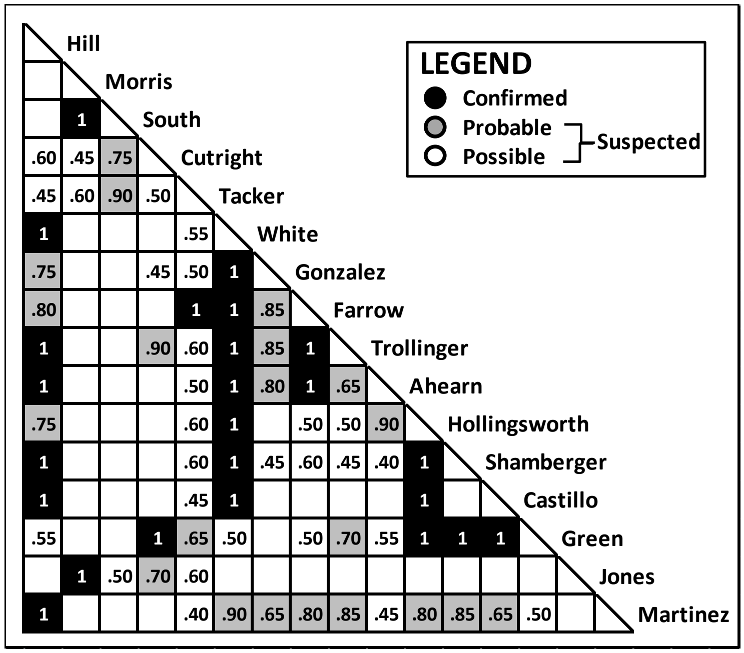
# APPENDICES

## APPENDIX A: FM 3-24 EDGE LIKELIHOODS

| Edge Likelihood (Probable Associations) | |
|---|---|
| Hill-Gonzalez | 0.75 |
| Hill-Farrow | 0.80 |
| Hill-Hollingsworth | 0.75 |
| South-Cutright | 0.75 |
| South-Tacker | 0.90 |
| Cutright-Trollinger | 0.90 |
| Gonzalez-Farrow | 0.85 |
| Gonzalez-Trollinger | 0.85 |
| Ahearn-Gonzalez | 0.80 |
| Ahearn-Trollinger | 0.65 |
| Ahearn-Hollingsworth | 0.90 |
| Hollingsworth-Martinez | 0.80 |
| Shamberger-Martinez | 0.85 |
| Castillo-Martinez | 0.65 |
| Green-Tacker | 0.65 |
| Green-Trollinger | 0.70 |
| Jones-Cutright | 0.70 |
| Martinez-White | 0.90 |
| Martinez-Gonzalez | 0.65 |
| Martinez-Farrow | 0.80 |
| Martinez-Trollinger | 0.85 |

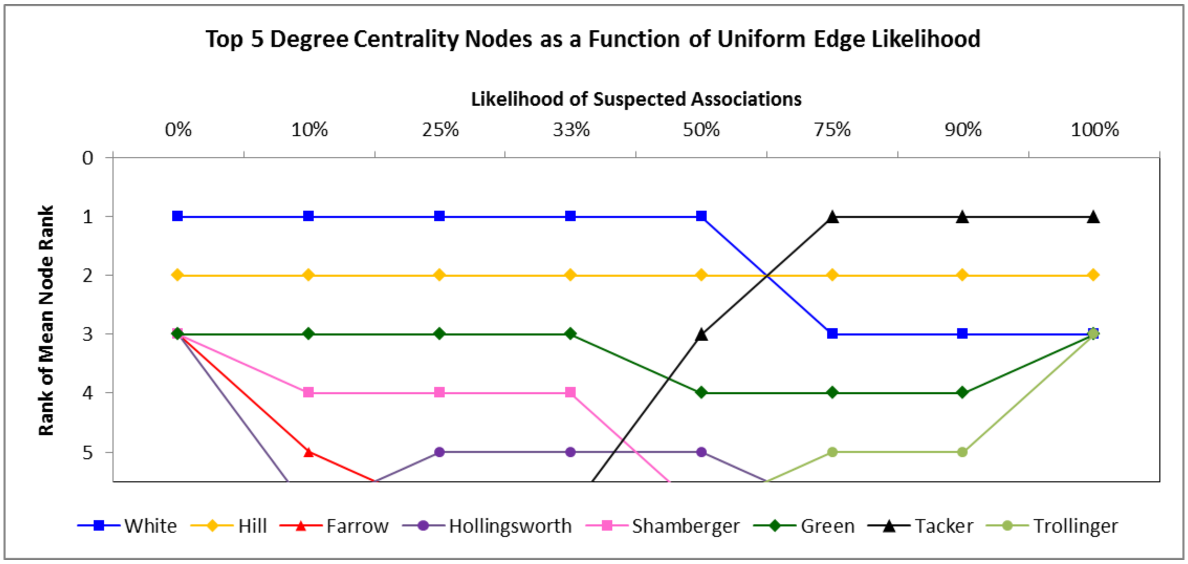
| Edge Likelihood (Possible Associations) | |
|---|---|
| Hill-Cutright | 0.60 |
| Hill-Tacker | 0.45 |
| Hill-Green | 0.55 |
| Morris-Cutright | 0.45 |
| Morris-Tacker | 0.60 |
| Cutright-Tacker | 0.50 |
| Cutright-Gonzalez | 0.45 |
| Tacker-White | 0.55 |
| Tacker-Gonzalez | 0.50 |
| Tacker-Trollinger | 0.60 |
| Ahearn-Tacker | 0.50 |
| Ahearn-Shamberger | 0.40 |
| Ahearn-Green | 0.55 |
| Ahearn-Martinez | 0.45 |
| Hollingsworth-Tacker | 0.60 |
| Hollingsworth-Farrow | 0.50 |
| Hollingsworth-Trollinger | 0.50 |
| Shamberger-Tacker | 0.60 |
| Shamberger-Gonzalez | 0.45 |
| Shamberger-Farrow | 0.60 |
| Shamberger-Trollinger | 0.45 |
| Castillo-Tacker | 0.45 |
| Green-White | 0.50 |
| Green-Farrow | 0.50 |
| Green-Martinez | 0.50 |
| Jones-South | 0.50 |
| Jones-Tacker | 0.60 |
| Martinez-Tacker | 0.40 |

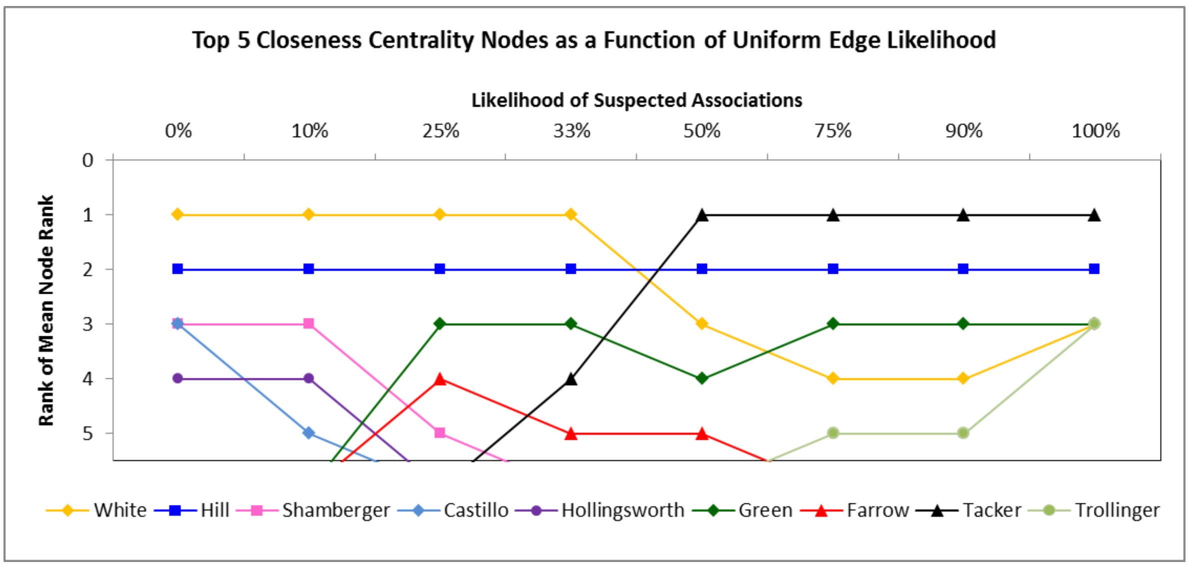
## APPENDIX A: FM 3-24 EDGE LIKELIHOODS (Continued)



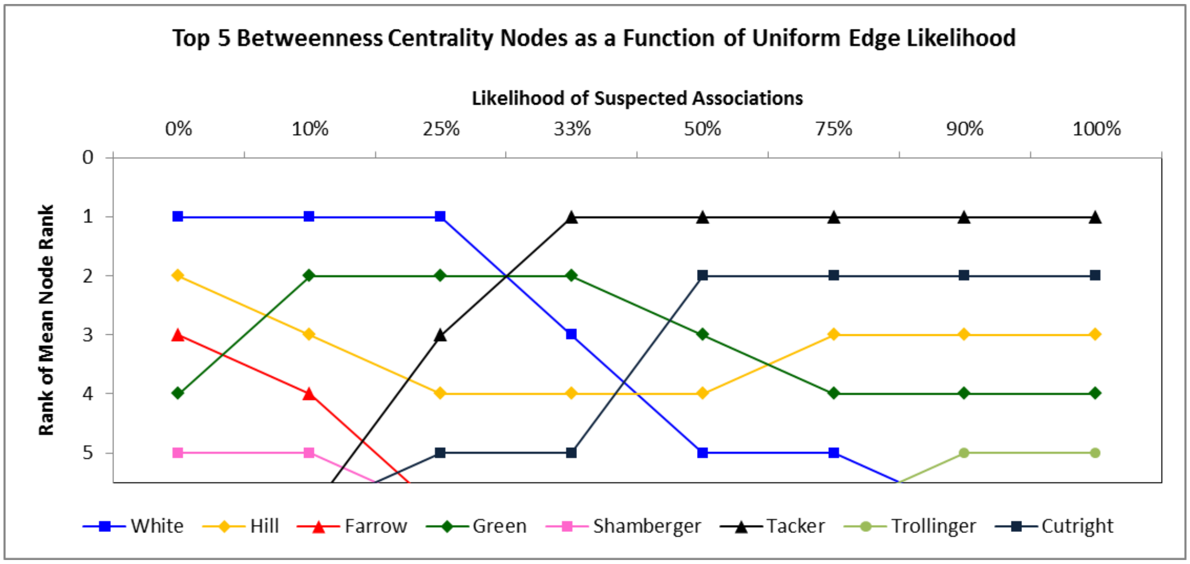

| Hill | Morris | South | Cutright | Tacker | White | Gonzalez | Farrow | Trollinger | Ahearn | Hollingsworth | Shamberger | Castillo | Green | Jones | Martinez | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | .60 | .45 | 1 | .75 | .80 | 1 | 1 | .75 | 1 | 1 | .55 | 0 | 1 | **Hill** |
| 0 | 0 | 1 | .45 | .60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | **Morris** |
| 0 | 1 | 0 | .75 | .90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .50 | 0 | **South** |
| .60 | .45 | .75 | 0 | .50 | 0 | .45 | 0 | .90 | 0 | 0 | 0 | 0 | 1 | .70 | 0 | **Cutright** |
| .45 | .60 | .90 | .50 | 0 | .55 | .50 | 1 | .60 | .50 | .60 | .60 | .45 | .65 | .60 | .40 | **Tacker** |
| 1 | 0 | 0 | 0 | .55 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | .50 | 0 | .90 | **White** |
| .75 | 0 | 0 | .45 | .50 | 1 | 0 | .85 | .85 | .80 | 0 | .45 | 0 | 0 | 0 | .65 | **Gonzalez** |
| .80 | 0 | 0 | 0 | 1 | 1 | .85 | 0 | 1 | 1 | .50 | .60 | 0 | .50 | 0 | .80 | **Farrow** |
| 1 | 0 | 0 | .90 | .60 | 1 | .85 | 1 | 0 | .65 | .50 | .45 | 0 | .70 | 0 | .85 | **Trollinger** |
| 1 | 0 | 0 | 0 | .50 | 1 | .80 | 1 | .65 | 0 | .90 | .40 | 0 | .55 | 0 | .45 | **Ahearn** |
| .75 | 0 | 0 | 0 | .60 | 1 | 0 | .50 | .50 | .90 | 0 | 1 | 1 | 1 | 0 | .80 | **Hollingsworth** |
| 1 | 0 | 0 | 0 | .60 | 1 | .45 | .60 | .45 | .40 | 1 | 0 | 0 | 1 | 0 | .85 | **Shamberger** |
| 1 | 0 | 0 | 0 | .45 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | .65 | **Castillo** |
| .55 | 0 | 0 | 1 | .65 | .50 | 0 | .50 | .70 | .55 | 1 | 1 | 1 | 0 | 0 | .50 | **Green** |
| 0 | 1 | .50 | .70 | .60 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **Jones** |
| 1 | 0 | 0 | 0 | .40 | .90 | .65 | .80 | .85 | .45 | .80 | .85 | .65 | .50 | 0 | 0 | **Martinez** |

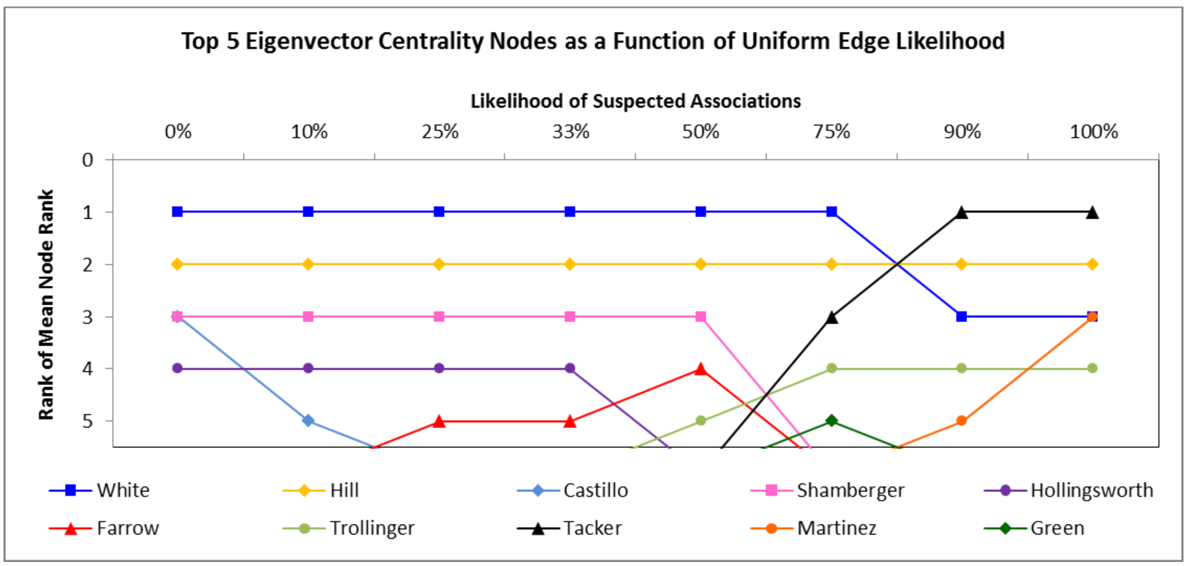
## APPENDIX B: ULRA RESULTS FOR THE FM 3-24 NETWORK



## APPENDIX B: ULRA RESULTS FOR THE FM 3-24 NETWORK (Continued)



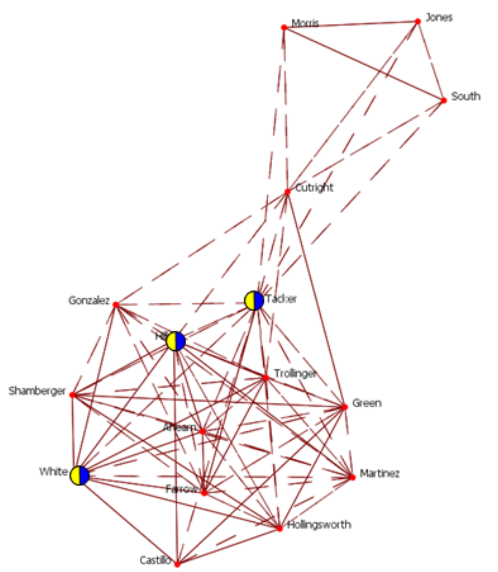



## APPENDIX C: FM 3-24 VLRA RESULTS FOR SCENARIO 1 (ALL CONFIRMED AND SUSPECTED ASSOCIATIONS CONSIDERED)

### C.1 DEGREE CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| Tacker | 1 | White | 1 | White | 1.61 | 99.2% | 0.8% | 0.0% |
| Hill | 2 | Hill | 2 | Hill | 1.70 | 96.0% | 4.0% | 0.0% |
| White | 3 | Tacker | 3 | Tacker | 2.60 | 74.6% | 25.4% | 0.0% |
| Trollinger | 3 | Trollinger | 4 | Trollinger | 2.93 | 69.5% | 30.5% | 0.0% |
| Green | 3 | Farrow | 5 | Farrow | 3.37 | 56.9% | 43.1% | 0.0% |
| Martinez | 3 | Hollingsworth | 5 | Hollingsworth | 3.37 | 55.6% | 44.4% | 0.0% |
| Farrow | 4 | Green | 6 | Green | 3.46 | 52.7% | 47.3% | 0.0% |
| Hollingsworth | 4 | Martinez | 7 | Martinez | 3.53 | 50.8% | 49.2% | 0.0% |
| Shamberger | 4 | Shamberger | 8 | Shamberger | 4.00 | 34.8% | 65.2% | 0.0% |
| Ahearn | 4 | Ahearn | 9 | Ahearn | 4.11 | 32.3% | 67.7% | 0.0% |
| Gonzalez | 5 | Gonzalez | 10 | Gonzalez | 4.94 | 12.2% | 87.8% | 0.0% |
| Cutright | 6 | Cutright | 11 | Cutright | 5.74 | 2.2% | 97.8% | 0.0% |
| Castillo | 7 | Castillo | 12 | Castillo | 5.95 | 0.2% | 99.8% | 0.0% |
| South | 8 | South | 13 | Morris | 7.56 | 0.0% | 100.0% | 0.0% |
| Morris | 8 | Morris | 14 | South | 7.46 | 0.0% | 99.9% | 0.1% |
| Jones | 8 | Jones | 15 | Jones | 7.77 | 0.0% | 99.6% | 0.4% |

**Full Network (Confirmed & Suspected)**




Known associations

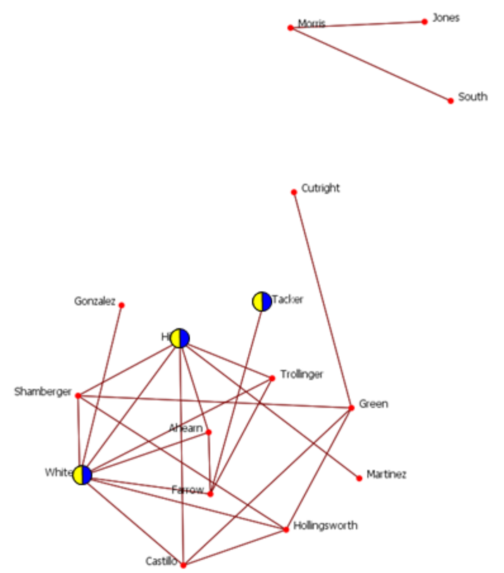
Suspected Associations

VLRA / Valued Top 3

VLRA Model Top 3

Valued Model Top 3

**Confirmed Associations**




Known associations

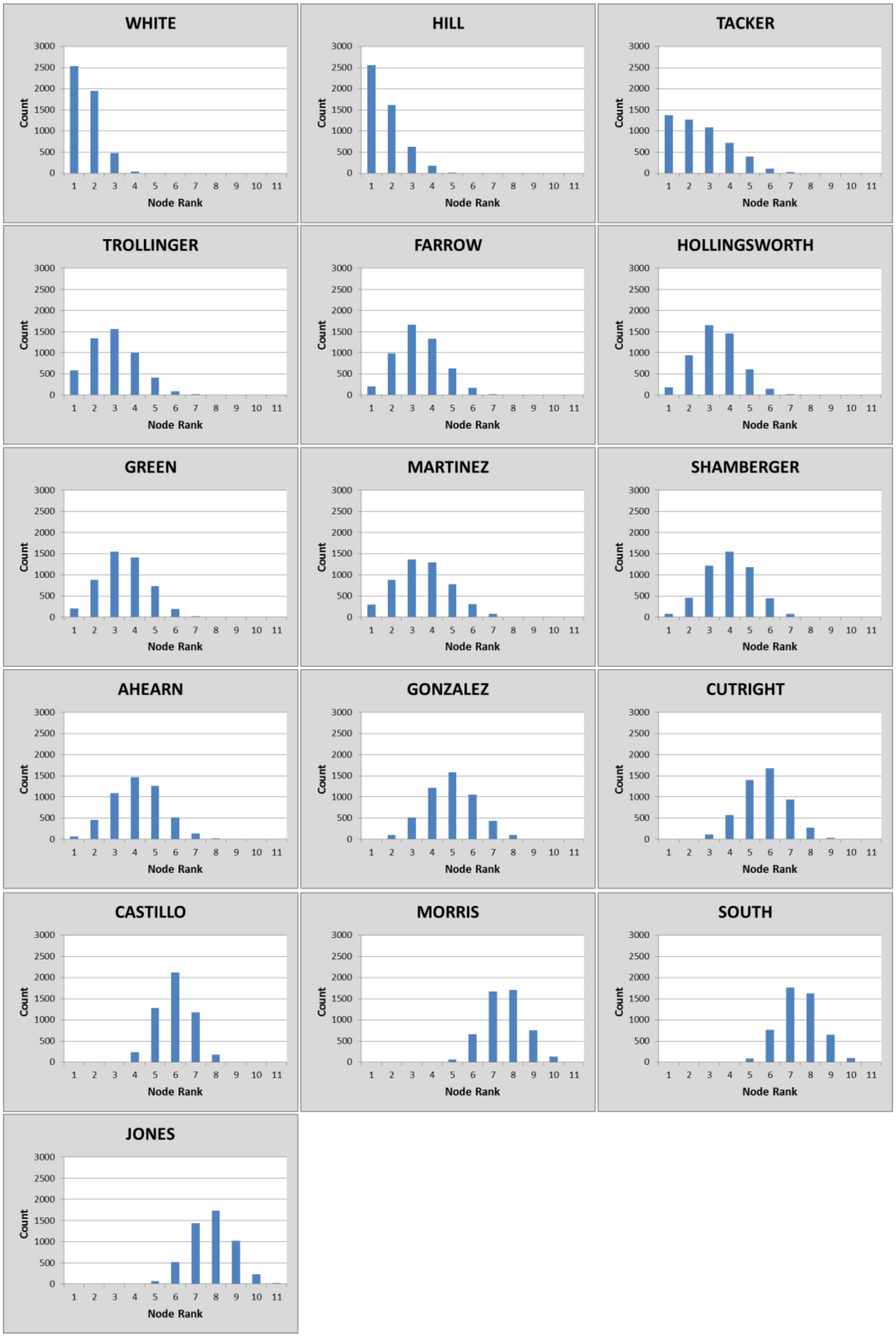
Suspected Associations

VLRA / Valued Top 3

VLRA Model Top 3

Valued Model Top 3

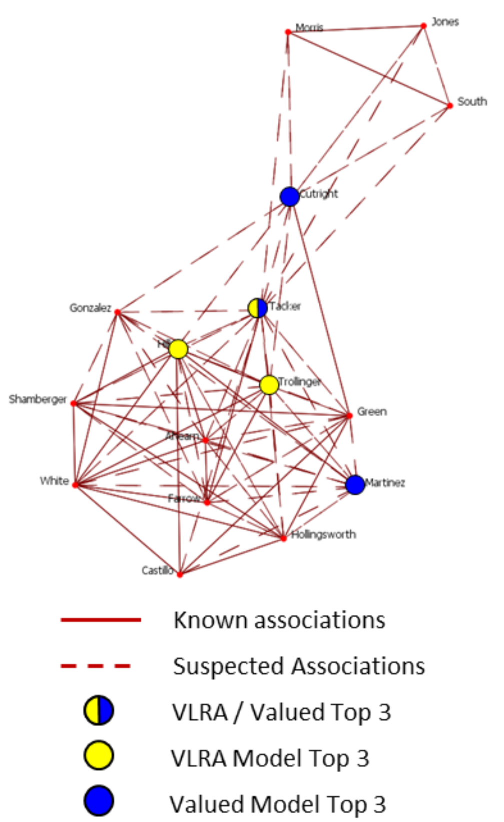
## Histograms of Degree Centrality Rank Frequency



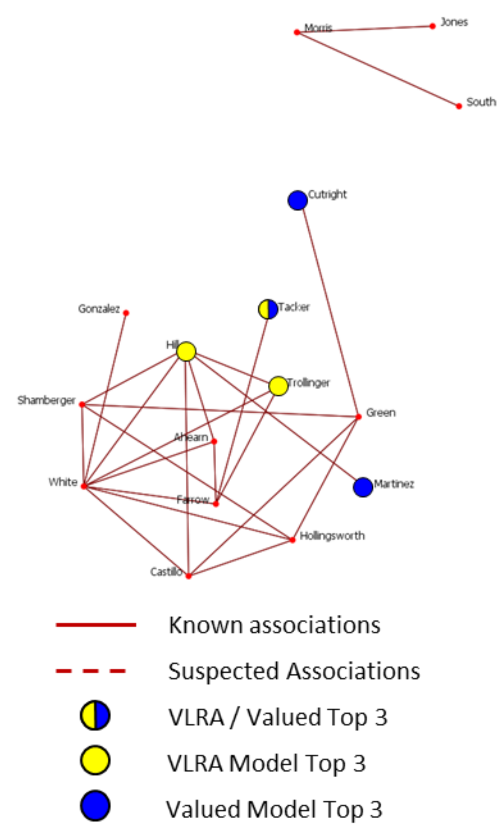
## C.2 CLOSENESS CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| Tacker | 1 | Tacker | 1 | Tacker | 2.09 | 88.6% | 11.4% | 0.0% |
| Hill | 2 | Martinez | 2 | Hill | 2.36 | 77.8% | 22.2% | 0.0% |
| White | 3 | Cutright | 3 | Trollinger | 2.96 | 68.7% | 31.3% | 0.0% |
| Trollinger | 3 | Gonzalez | 4 | White | 2.78 | 66.6% | 33.4% | 0.0% |
| Green | 3 | Ahearn | 5 | Green | 3.31 | 58.3% | 41.7% | 0.0% |
| Martinez | 3 | Hill | 6 | Farrow | 3.47 | 52.2% | 47.8% | 0.0% |
| Hollingsworth | 4 | Shamberger | 7 | Hollingsworth | 4.27 | 38.6% | 61.4% | 0.0% |
| Farrow | 4 | Green | 8 | Martinez | 4.84 | 25.1% | 74.8% | 0.1% |
| Shamberger | 4 | Trollinger | 9 | Shamberger | 4.83 | 24.7% | 75.2% | 0.1% |
| Ahearn | 4 | Hollingsworth | 10 | Ahearn | 5.15 | 20.3% | 79.4% | 0.3% |
| Gonzalez | 5 | White | 11 | Gonzalez | 5.39 | 16.1% | 83.0% | 1.0% |
| Cutright | 6 | Castillo | 12 | Cutright | 5.19 | 10.9% | 88.5% | 0.6% |
| Castillo | 7 | Jones | 13 | Castillo | 6.71 | 0.7% | 97.9% | 1.4% |
| South | 8 | Morris | 14 | South | 8.62 | 0.0% | 76.6% | 23.4% |
| Morris | 8 | Farrow | 15 | Morris | 9.04 | 0.0% | 64.2% | 35.8% |
| Jones | 8 | South | 16 | Jones | 9.08 | 0.0% | 62.0% | 38.0% |

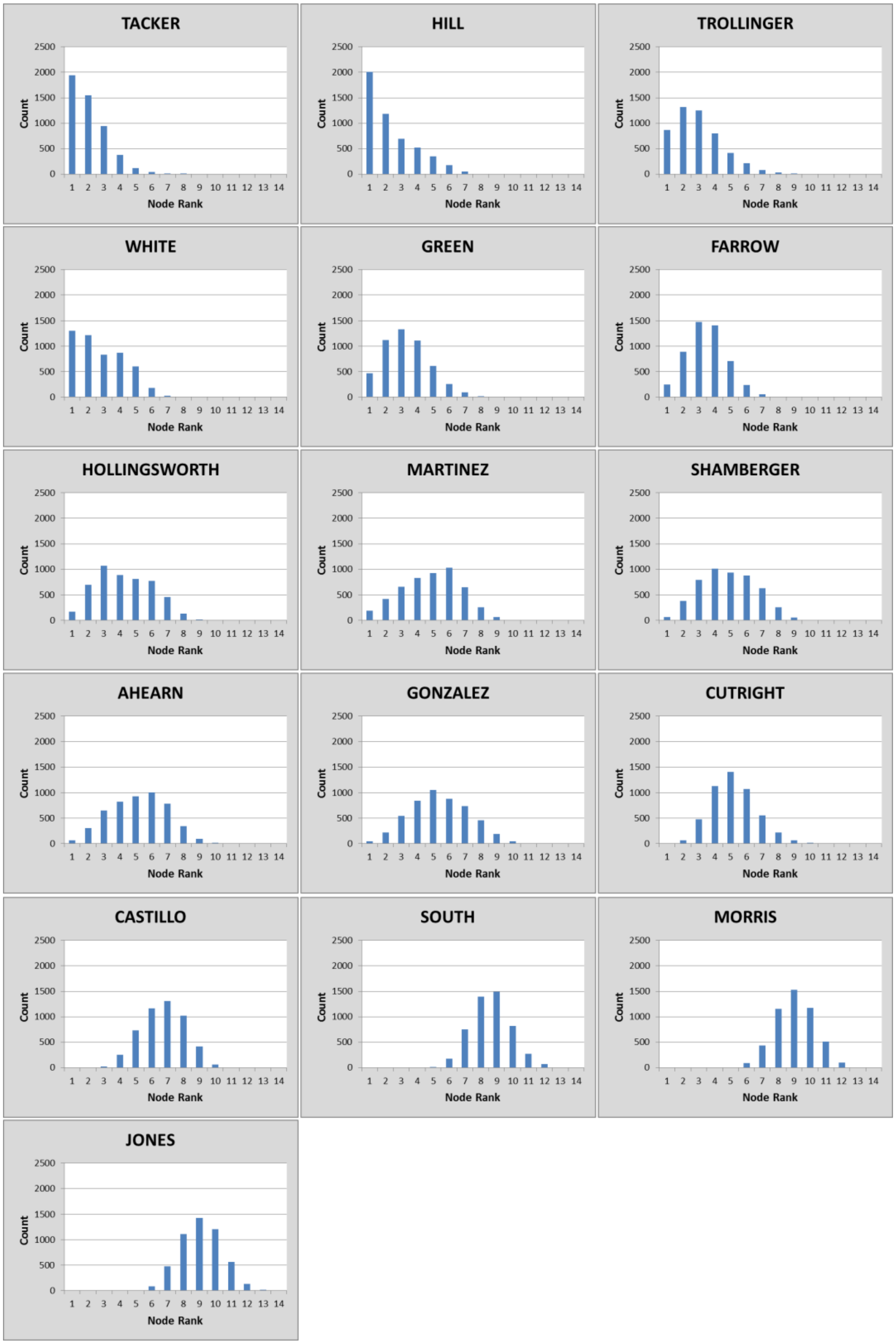
**Full Network (Confirmed & Suspected)**




**Confirmed Associations**



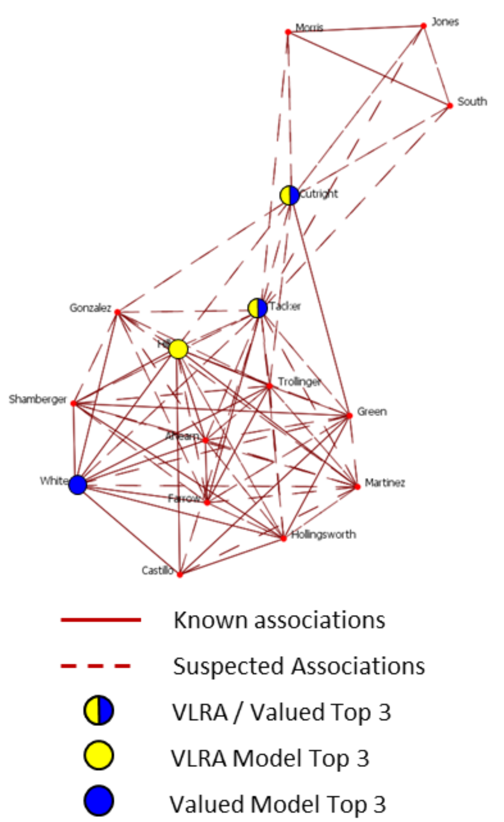
**Histograms of Closeness Centrality Rank Frequency**



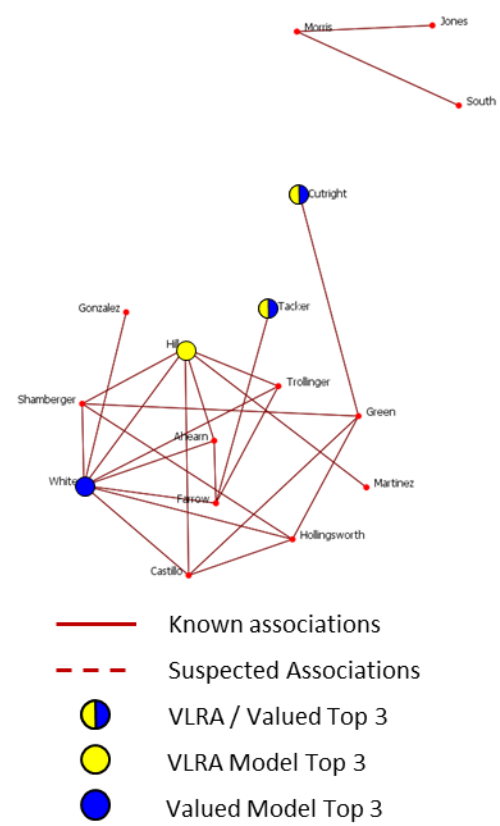
## C.3 BETWEENNESS CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| Tacker | 1 | Tacker | 1 | Tacker | 1.35 | 97.1% | 2.3% | 0.6% |
| Cutright | 2 | Cutright | 2 | Cutright | 3.37 | 69.7% | 26.5% | 3.8% |
| Hill | 3 | White | 3 | Hill | 4.85 | 42.5% | 53.2% | 4.3% |
| Green | 4 | Trollinger | 4 | Green | 4.68 | 28.9% | 70.3% | 0.8% |
| Trollinger | 5 | Green | 5 | Trollinger | 5.76 | 20.0% | 73.7% | 6.3% |
| White | 6 | Farrow | 6 | South | 10.03 | 13.1% | 31.3% | 55.6% |
| Martinez | 6 | Hill | 7 | White | 6.00 | 12.6% | 84.5% | 2.9% |
| Gonzalez | 7 | Morris | 8 | Morris | 11.54 | 8.0% | 22.0% | 70.0% |
| Hollingsworth | 8 | Hollingsworth | 8 | Jones | 11.82 | 5.0% | 23.7% | 71.3% |
| Farrow | 9 | Castillo | 9 | Hollingsworth | 8.84 | 1.8% | 68.3% | 29.9% |
| Shamberger | 9 | Shamberger | 9 | Farrow | 8.46 | 1.5% | 77.5% | 21.0% |
| Ahearn | 9 | Gonzalez | 10 | Gonzalez | 10.52 | 1.2% | 45.1% | 53.7% |
| Castillo | 10 | Martinez | 10 | Martinez | 9.98 | 1.0% | 52.3% | 46.7% |
| Morris | 10 | Jones | 10 | Shamberger | 10.42 | 0.6% | 46.0% | 53.3% |
| Jones | 10 | Ahearn | 10 | Ahearn | 11.00 | 0.2% | 38.7% | 61.1% |
| South | 10 | South | 10 | Castillo | 12.92 | 0.0% | 11.6% | 88.4% |

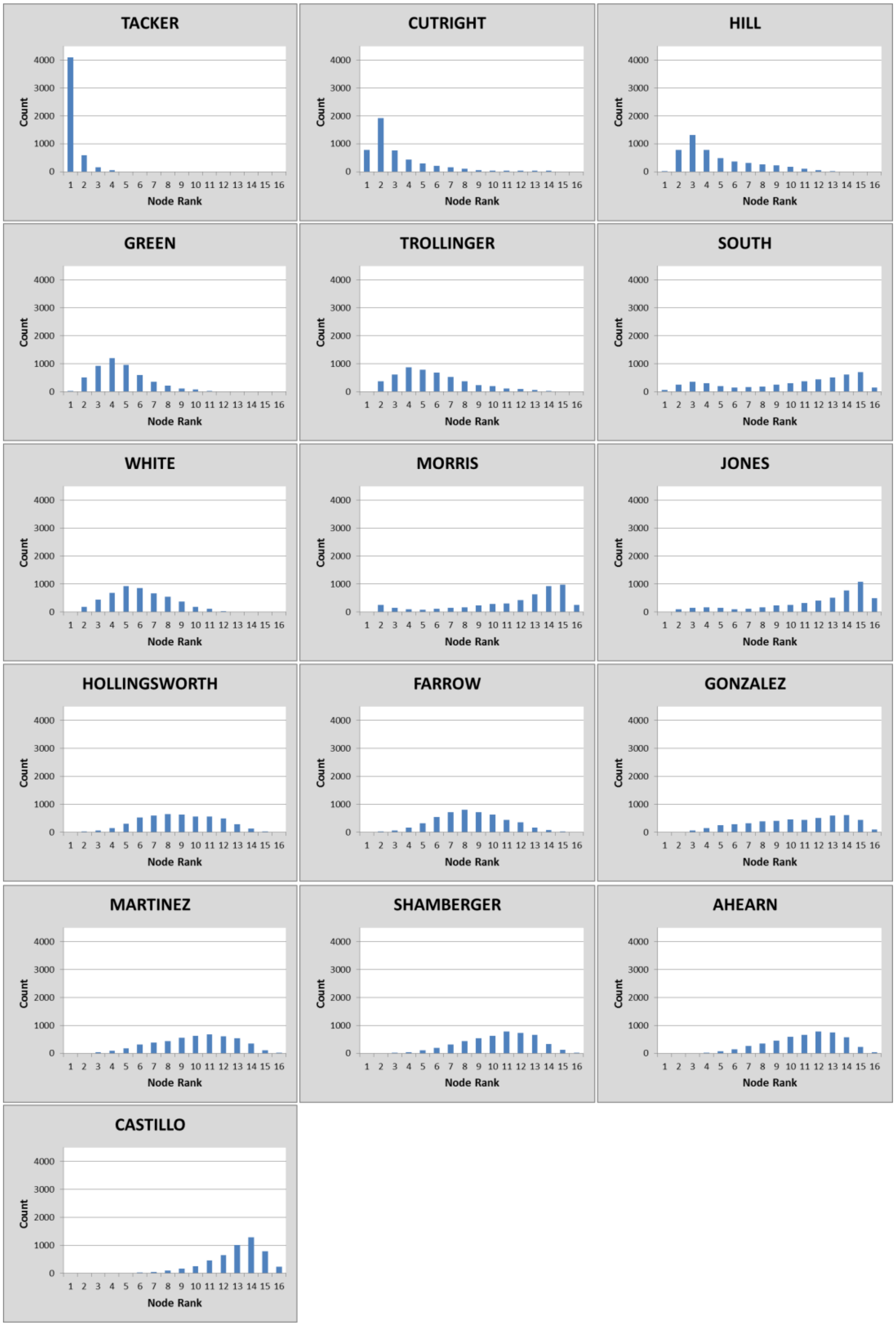
**Full Network (Confirmed & Suspected)**




**Confirmed Associations**



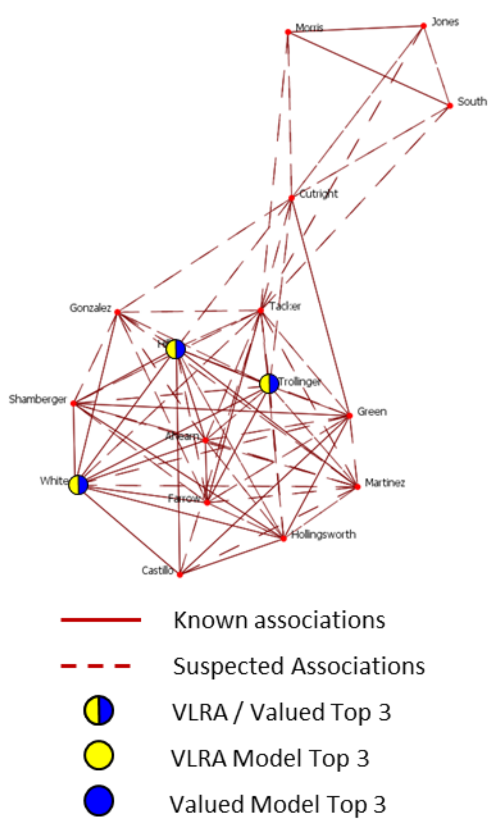
## Histograms of Betweenness Centrality Rank Frequency



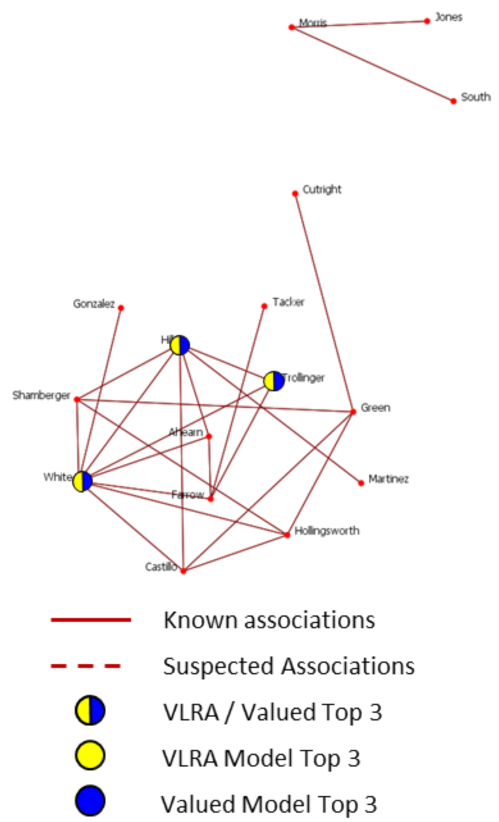
## C.4 EIGENVECTOR CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| Tacker | 1 | White | 1 | White | 2.03 | 87.4% | 12.6% | 0.0% |
| Hill | 2 | Hill | 2 | Hill | 2.59 | 76.4% | 23.5% | 0.1% |
| White | 3 | Trollinger | 3 | Trollinger | 5.26 | 29.6% | 67.9% | 2.5% |
| Martinez | 3 | Farrow | 4 | Farrow | 5.49 | 24.2% | 73.2% | 2.6% |
| Trollinger | 4 | Hollingsworth | 5 | Martinez | 6.02 | 22.5% | 71.6% | 5.9% |
| Green | 5 | Martinez | 6 | Hollingsworth | 6.11 | 17.8% | 77.3% | 4.9% |
| Farrow | 6 | Shamberger | 7 | Tacker | 7.99 | 11.1% | 63.2% | 25.7% |
| Shamberger | 6 | Ahearn | 8 | Shamberger | 6.89 | 10.4% | 82.9% | 6.7% |
| Ahearn | 6 | Green | 9 | Ahearn | 7.14 | 10.1% | 80.0% | 9.9% |
| Hollingsworth | 7 | Tacker | 10 | Green | 7.63 | 9.6% | 72.5% | 17.9% |
| Gonzalez | 8 | Gonzalez | 11 | Gonzalez | 9.57 | 0.9% | 60.6% | 38.4% |
| Castillo | 9 | Castillo | 12 | Castillo | 11.46 | 0.0% | 14.1% | 85.9% |
| Cutright | 10 | Cutright | 13 | Cutright | 12.82 | 0.0% | 0.5% | 99.5% |
| South | 11 | South | 14 | South | 14.60 | 0.0% | 0.0% | 100.0% |
| Jones | 11 | Jones | 15 | Jones | 15.25 | 0.0% | 0.0% | 100.0% |
| Morris | 11 | Morris | 16 | Morris | 15.14 | 0.0% | 0.0% | 100.0% |

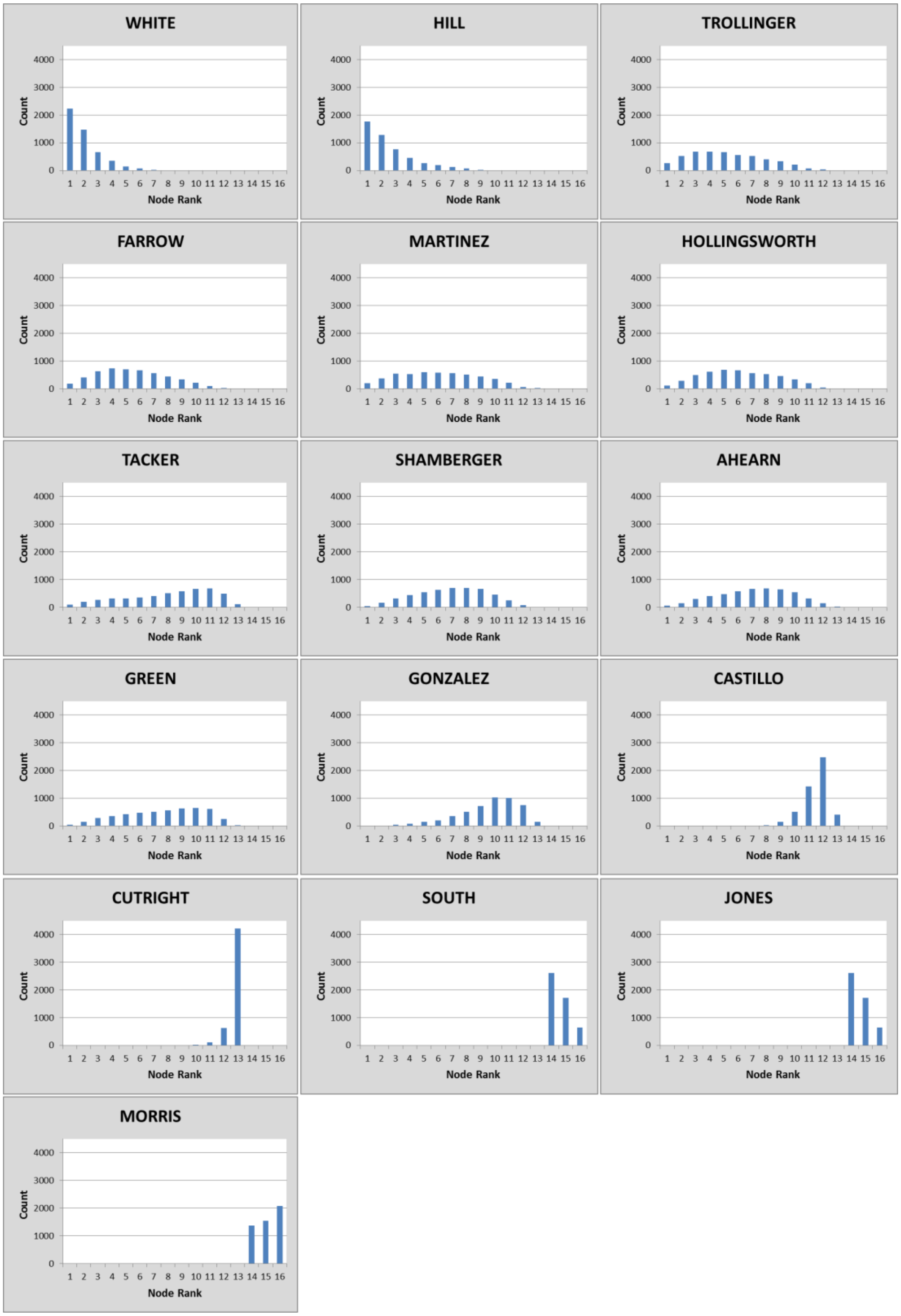
**Full Network (Confirmed & Suspected)**




**Confirmed Associations**



**Histograms of Eigenvector Centrality Rank Frequency**

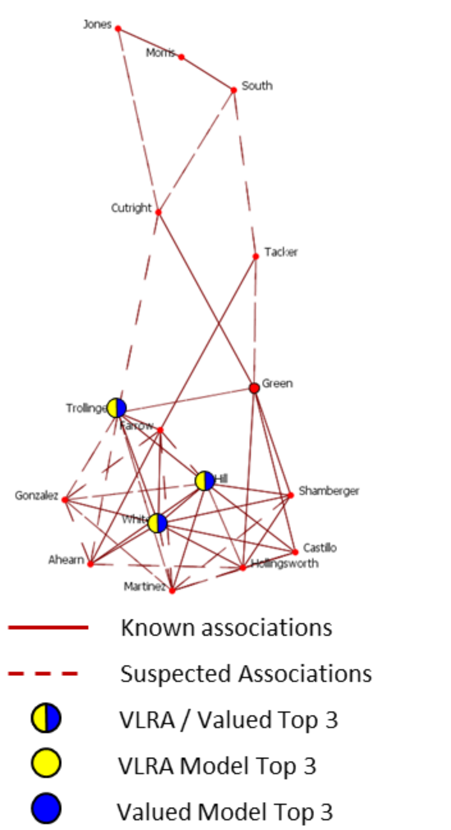


## APPENDIX D: FM 3-24 VLRA RESULTS FOR SCENARIO 2 (ONLY PROBABLE AND CONFIRMED ASSOCIATIONS CONSIDERED)

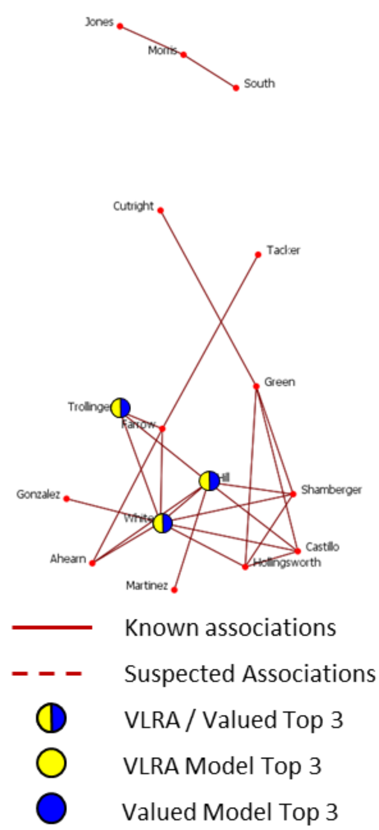
### D.1 DEGREE CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| White | 1 | White | 1 | White | 1.05 | 100.0% | 0.0% | 0.0% |
| Hill | 1 | Hill | 2 | Hill | 1.58 | 99.7% | 0.3% | 0.0% |
| Trollinger | 2 | Trollinger | 3 | Trollinger | 2.65 | 84.9% | 15.1% | 0.0% |
| Martinez | 2 | Martinez | 4 | Farrow | 3.13 | 72.1% | 27.9% | 0.0% |
| Farrow | 3 | Farrow | 5 | Hollingsworth | 3.12 | 72.0% | 28.0% | 0.0% |
| Hollingsworth | 3 | Hollingsworth | 5 | Martinez | 3.07 | 69.1% | 30.9% | 0.0% |
| Ahearn | 4 | Ahearn | 6 | Green | 4.19 | 18.2% | 81.8% | 0.0% |
| Green | 4 | Green | 6 | Ahearn | 4.19 | 17.4% | 82.3% | 0.3% |
| Gonzalez | 4 | Gonzalez | 7 | Gonzalez | 4.60 | 11.7% | 85.3% | 3.1% |
| Shamberger | 5 | Shamberger | 8 | Shamberger | 4.69 | 2.4% | 97.6% | 0.0% |
| Castillo | 5 | Castillo | 9 | Castillo | 4.89 | 1.8% | 98.2% | 0.0% |
| Cutright | 6 | Cutright | 10 | Cutright | 5.97 | 0.0% | 78.2% | 21.8% |
| South | 7 | South | 11 | South | 6.62 | 0.0% | 41.8% | 58.2% |
| Tacker | 7 | Tacker | 12 | Tacker | 6.73 | 0.0% | 38.3% | 61.7% |
| Morris | 8 | Morris | 13 | Morris | 7.28 | 0.0% | 12.5% | 87.5% |
| Jones | 8 | Jones | 14 | Jones | 7.59 | 0.0% | 7.4% | 92.6% |

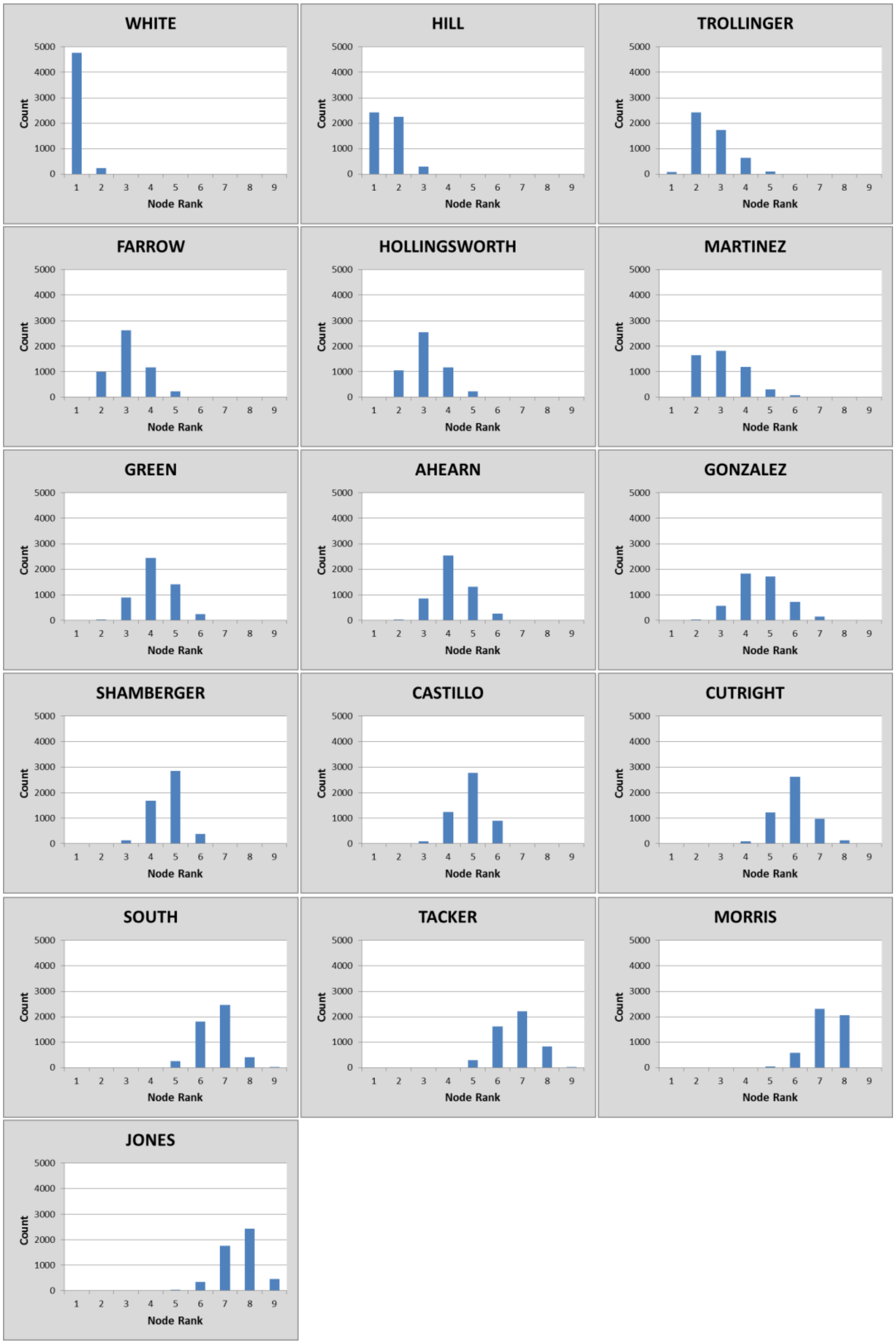
Full Network (Confirmed & Suspected)




Confirmed Associations

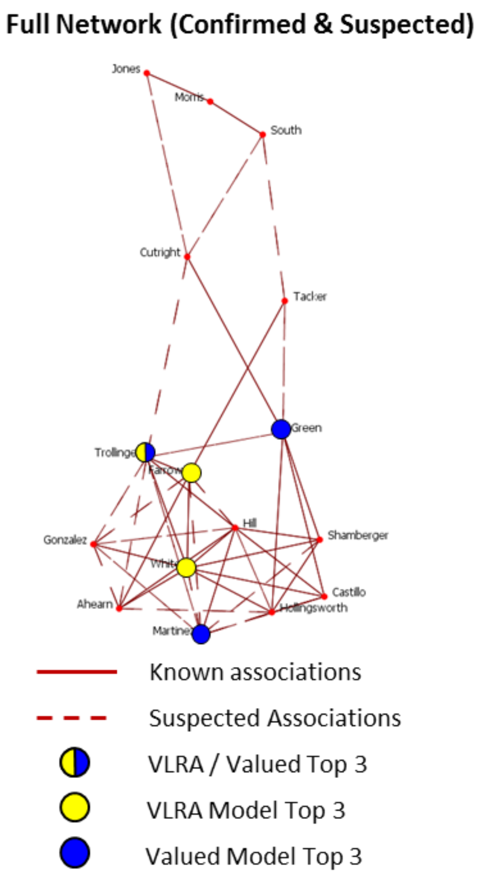


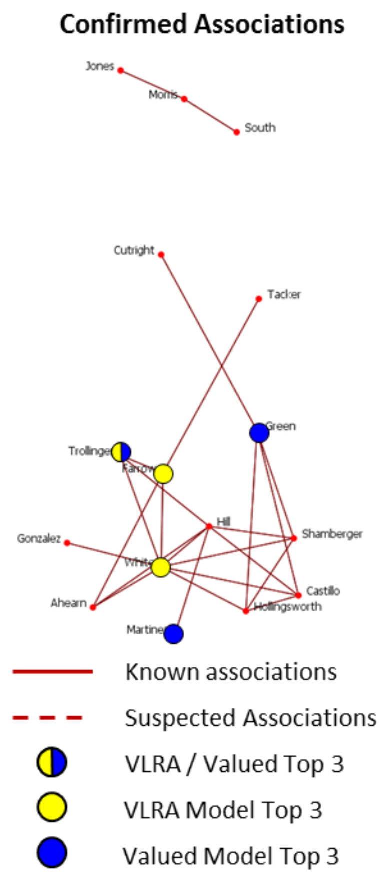
**Histograms of Degree Centrality Rank Frequency**



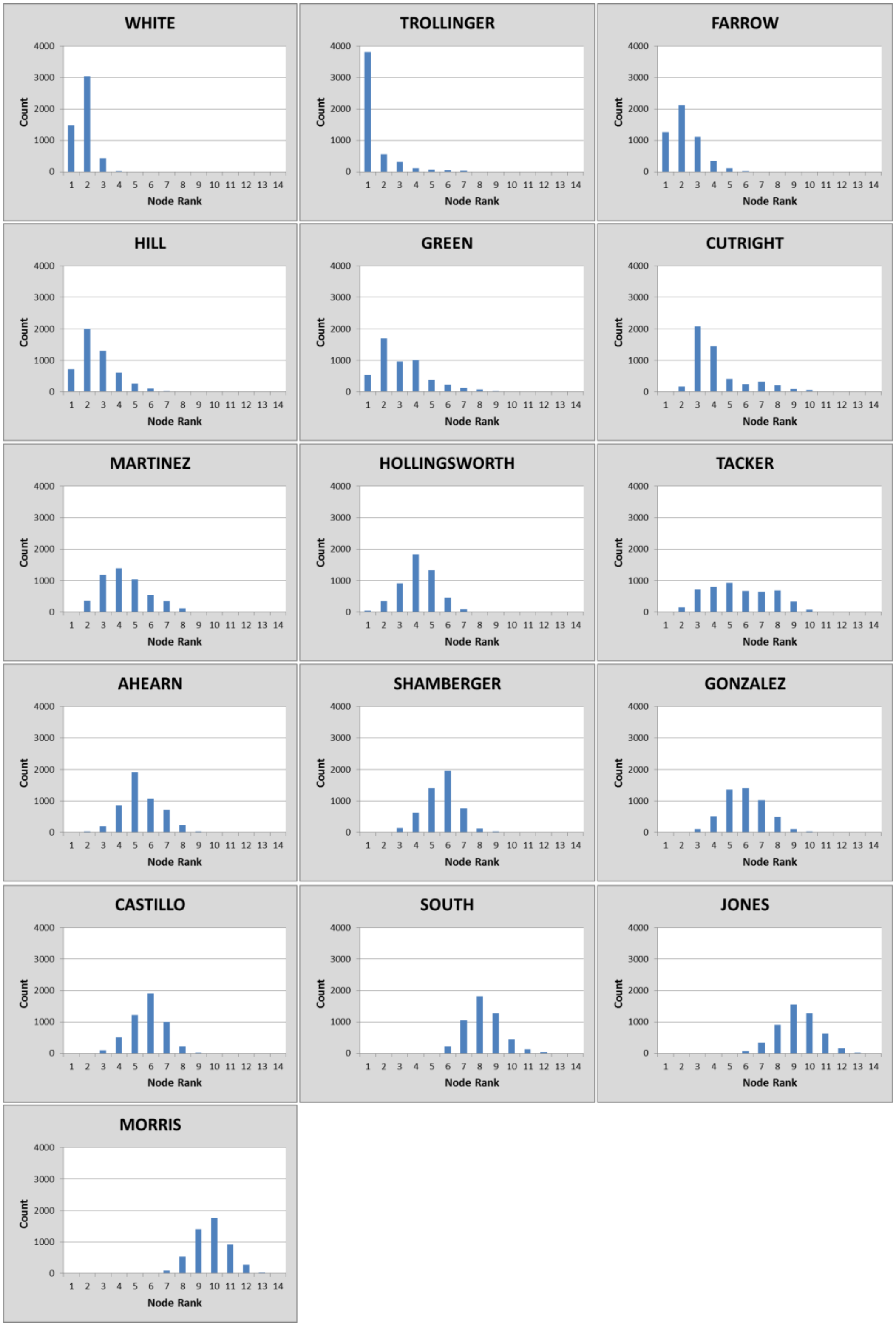
## D.2 CLOSENESS CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| Trollinger | 1 | Trollinger | 1 | White | 1.80 | 99.5% | 0.5% | 0.0% |
| White | 2 | Green | 2 | Trollinger | 1.50 | 94.0% | 6.0% | 0.0% |
| Farrow | 2 | Martinez | 3 | Farrow | 2.20 | 90.3% | 9.7% | 0.0% |
| Hill | 2 | Farrow | 4 | Hill | 2.62 | 80.2% | 19.8% | 0.0% |
| Green | 2 | Hill | 5 | Green | 3.14 | 63.6% | 36.2% | 0.2% |
| Cutright | 3 | Gonzalez | 6 | Cutright | 4.20 | 44.8% | 54.2% | 1.0% |
| Martinez | 3 | Ahearn | 7 | Martinez | 4.36 | 30.9% | 69.0% | 0.1% |
| Hollingsworth | 4 | White | 8 | Hollingsworth | 4.15 | 26.0% | 74.0% | 0.0% |
| Ahearn | 5 | Cutright | 9 | Tacker | 5.61 | 17.2% | 81.3% | 1.5% |
| Gonzalez | 5 | Hollingsworth | 9 | Ahearn | 5.38 | 4.3% | 95.7% | 0.0% |
| Tacker | 5 | Tacker | 10 | Shamberger | 5.60 | 2.8% | 97.2% | 0.0% |
| Castillo | 6 | Castillo | 11 | Gonzalez | 5.92 | 2.3% | 97.3% | 0.3% |
| Shamberger | 6 | Shamberger | 12 | Castillo | 5.79 | 2.1% | 97.8% | 0.0% |
| South | 7 | South | 13 | South | 8.24 | 0.0% | 87.6% | 12.4% |
| Jones | 8 | Jones | 14 | Jones | 9.28 | 0.0% | 57.6% | 42.4% |
| Morris | 9 | Morris | 15 | Morris | 9.76 | 0.0% | 40.6% | 59.4% |



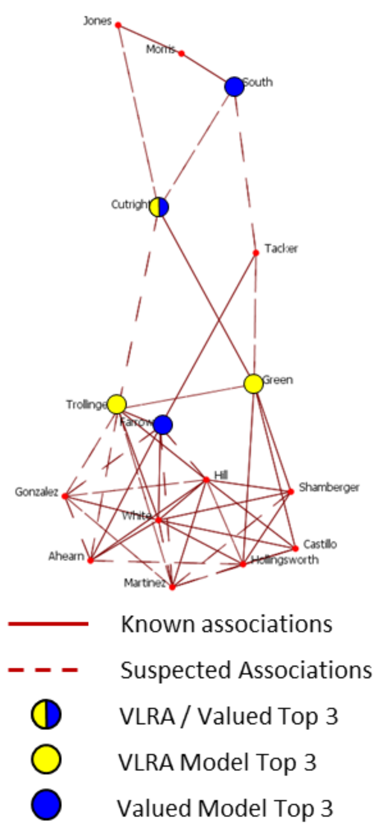
**Histograms of Closeness Centrality Rank Frequency**



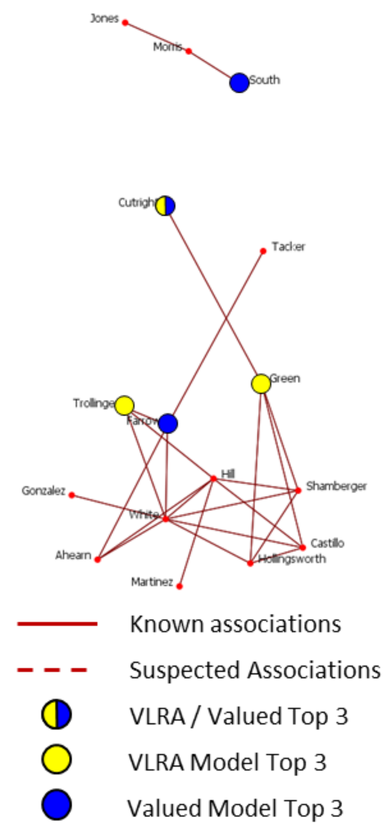
## D.3 BETWEENNESS CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| Cutright | 1 | Cutright | 1 | Cutright | 2.81 | 78.2% | 14.6% | 7.2% |
| Trollinger | 2 | Farrow | 2 | Trollinger | 3.45 | 67.1% | 30.2% | 2.7% |
| Farrow | 3 | South | 3 | Green | 3.39 | 60.5% | 39.3% | 0.2% |
| Green | 4 | Trollinger | 4 | Farrow | 4.05 | 36.0% | 64.0% | 0.0% |
| South | 5 | White | 5 | South | 4.57 | 29.0% | 68.4% | 2.6% |
| Tacker | 6 | Green | 6 | Tacker | 5.24 | 28.5% | 62.6% | 8.9% |
| White | 7 | Tacker | 7 | Jones | 9.71 | 1.9% | 65.3% | 32.8% |
| Hill | 7 | Hill | 8 | White | 6.85 | 0.7% | 99.3% | 0.0% |
| Jones | 8 | Hollingsworth | 9 | Hollingsworth | 9.43 | 0.2% | 79.1% | 20.7% |
| Hollingsworth | 9 | Morris | 10 | Hill | 7.99 | 0.2% | 95.3% | 4.5% |
| Martinez | 10 | Shamberger | 11 | Morris | 10.22 | 0.0% | 40.2% | 59.8% |
| Morris | 11 | Castillo | 12 | Martinez | 10.38 | 0.0% | 57.8% | 42.2% |
| Shamberger | 12 | Martinez | 13 | Shamberger | 11.68 | 0.0% | 21.0% | 79.0% |
| Ahearn | 12 | Jones | 13 | Castillo | 12.11 | 0.0% | 15.5% | 84.5% |
| Gonzalez | 13 | Ahearn | 13 | Ahearn | 12.16 | 0.0% | 11.3% | 88.7% |
| Castillo | 14 | Gonzalez | 13 | Gonzalez | 13.75 | 0.0% | 2.7% | 97.3% |

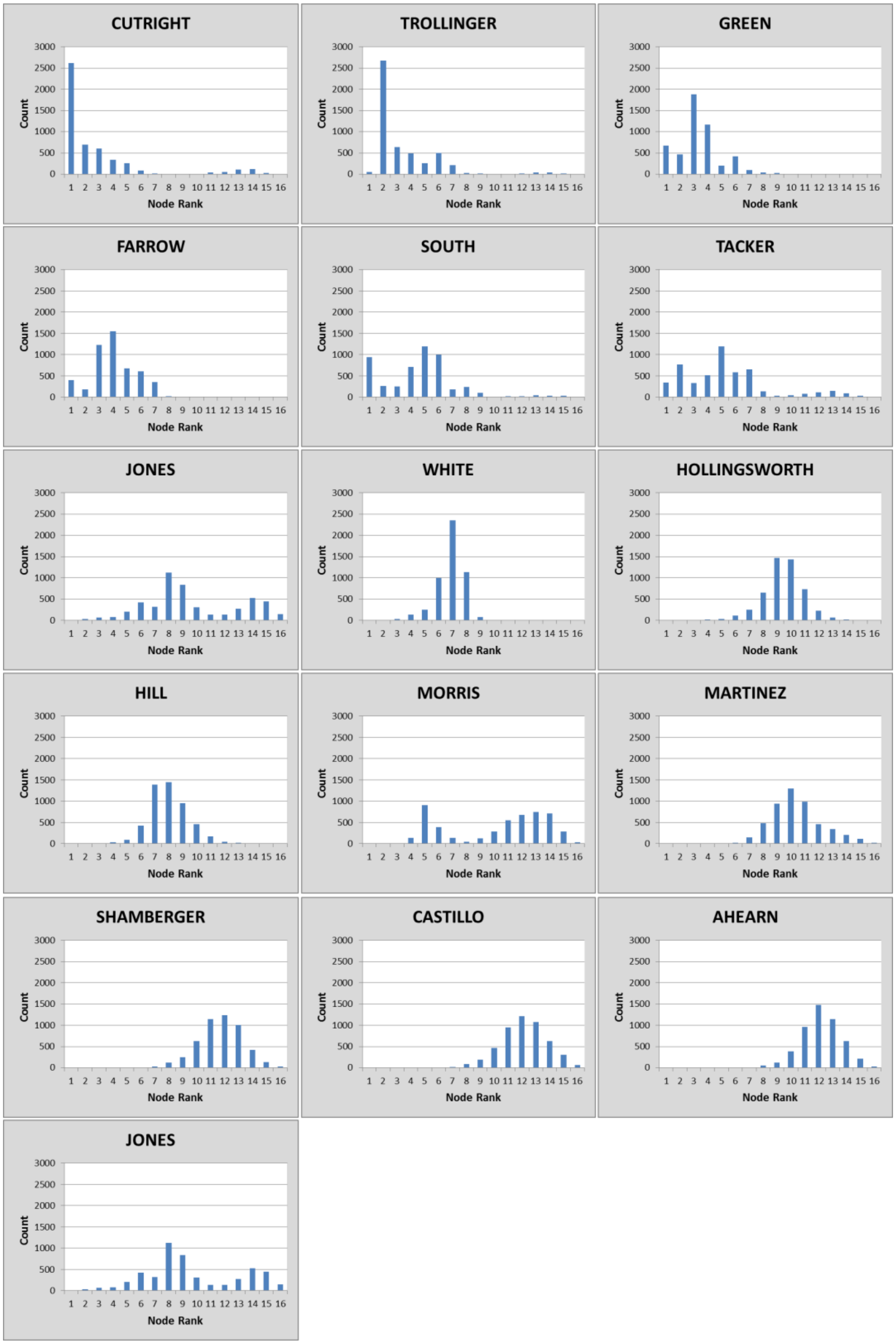
**Full Network (Confirmed & Suspected)**




**Confirmed Associations**

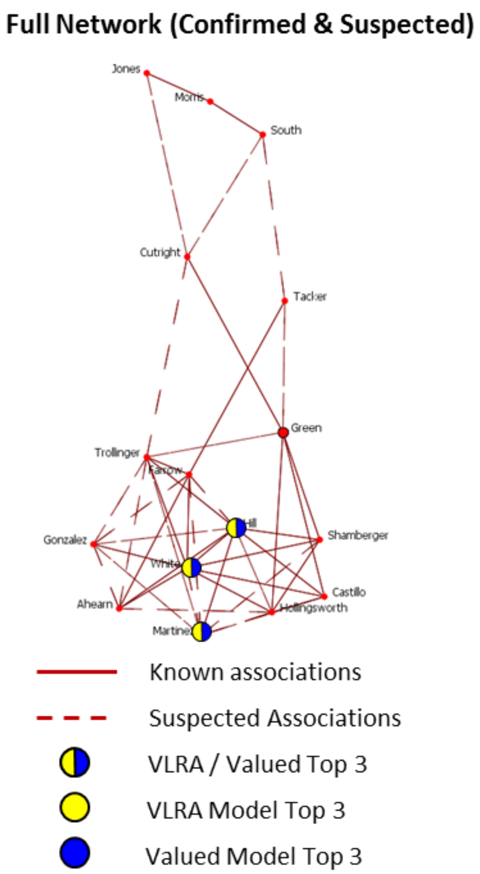


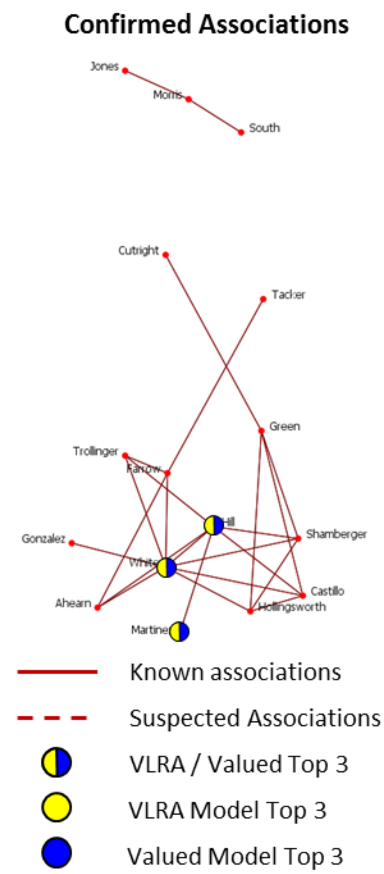
**Histograms of Closeness Centrality Rank Frequency**



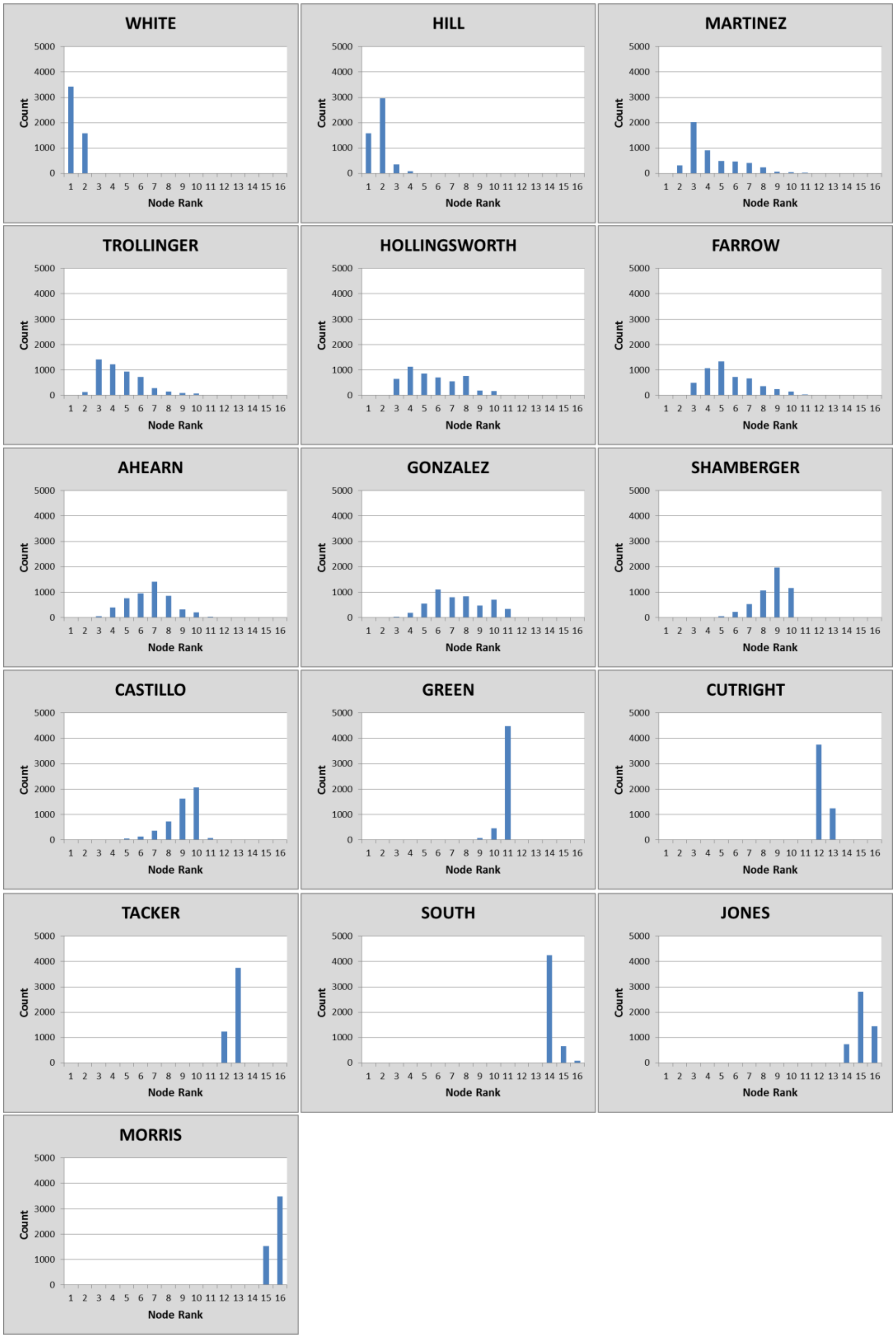
## D.4 EIGENVECTOR CENTRALITY ANALYSIS

| Certainty Model | | Valued Network Model | | VLRA Model | | | | |
|---|---|---|---|---|---|---|---|---|
| Node ID | Rank | Node ID | Rank | Node ID | Average Rank | P(top 3) | P(middle) | P(bottom ⅓) |
| Hill | 1 | White | 1 | White | 1.32 | 100.0% | 0.0% | 0.0% |
| White | 1 | Hill | 2 | Hill | 1.80 | 98.1% | 1.9% | 0.0% |
| Martinez | 2 | Martinez | 3 | Martinez | 4.34 | 46.8% | 52.8% | 0.4% |
| Trollinger | 3 | Trollinger | 4 | Trollinger | 4.60 | 30.7% | 69.3% | 0.0% |
| Farrow | 4 | Hollingsworth | 5 | Hollingsworth | 5.67 | 12.8% | 86.9% | 0.3% |
| Hollingsworth | 5 | Farrow | 6 | Farrow | 5.58 | 9.8% | 89.7% | 0.6% |
| Gonzalez | 6 | Ahearn | 7 | Ahearn | 6.67 | 1.2% | 98.1% | 0.7% |
| Ahearn | 7 | Gonzalez | 8 | Gonzalez | 7.48 | 0.6% | 92.6% | 6.8% |
| Shamberger | 8 | Shamberger | 9 | Shamberger | 8.64 | 0.0% | 99.7% | 0.3% |
| Castillo | 8 | Castillo | 10 | Castillo | 9.04 | 0.0% | 98.7% | 1.3% |
| Green | 9 | Green | 11 | Green | 10.88 | 0.0% | 10.4% | 89.6% |
| Cutright | 10 | Cutright | 12 | Cutright | 12.25 | 0.0% | 0.0% | 100.0% |
| Tacker | 11 | Tacker | 13 | Tacker | 12.75 | 0.0% | 0.0% | 100.0% |
| South | 12 | South | 14 | South | 14.17 | 0.0% | 0.0% | 100.0% |
| Jones | 13 | Jones | 15 | Jones | 15.14 | 0.0% | 0.0% | 100.0% |
| Morris | 14 | Morris | 16 | Morris | 15.69 | 0.0% | 0.0% | 100.0% |



**Histograms of Eigenvector Centrality Rank Frequency**



**APPENDIX E: MATLAB CODE**

**E.1 MAIN EXECUTION SCRIPT**

```
%% Main Execution Script
clear all; clc;

% FM 3-24 Network Properties
p = .25;      % Estimated reliability of suspected information for ULRA
n = 16;       % number of nodes

% Incidence Matrix of known associations
kI = [1 1  1  1  1  1  2  2  4  5  6  6  6  6  6  6  6  8  8  11  11  11  12  13]';
kJ = [6 9  10  12  13  16  3  15  14  8  7  8  9  10  11  12  13  9  10  12  13  14  14  14]';
kS = ones(size(kI));

kwnLinks = [kI kJ kS];  % I = source node, J = target node, S = edge value

% Incidence Matrix of suspected associations
sI = [3 4  6  10  7  7  9  12  1  7  8  11  1  1  3  4  9  5  7  9  13  1  2  5  5  5  5  8  1  5  10  3
4  5  5  6  8  8  9  14  1  2  4  5  7  9  10  5  10]';
sJ = [5 9  16  11  8  9  16  16  8  10  16  16  7  11  4  15  14  14  16  10  16  4  5  9  11  12  15  12  14
6  14  15  5  7  10  14  11  14  11  16  5  4  7  13  12  12  16  16  12]';
sS = ones(size(sI));

susLinks = [sI sJ sS];  % I = source node, J = target node, S = edge value

edge_likelihood = [0.9  0.9 0.9 0.9 0.85    0.85    0.85    0.85    0.8 0.8 0.8 0.8 0.75    0.75    0.75    0.7
0.7 0.65    0.65    0.65    0.65    0.6 0.6 0.6 0.6 0.6 0.6 0.6 0.55    0.55    0.55    0.5 0.5 0.5 0.5 0.5 0.5 0.5
0.5 0.5 0.45    0.45    0.45    0.45    0.45    0.45    0.45    0.4 0.4];

% Replication Parameters
r = round((1-p)*size(susLinks,1));  % number of edges to remove for ULRA
repcount = 5000;                    % number of replicates

%% Execute ULRA or VLRA (1 = Degree, 2 = Closeness, 3 = Betweenness, or 4 = Eigenvector)
%[rnkcnt rnkfrac nodeRnks] = ULRA(kwnLinks,susLinks,r,n,repcount,1);
[rnkcnt rnkfrac nodeRnks] = VLRA(kwnLinks,susLinks,edge_likelihood,n,repcount,1);
disp(' ')

for i = 1:n
  avgrnk(i) = mean(nodeRnks(i,:));
end
```

```
% Display table of results
[meanrnk idx] = sort(avgrnk,'ascend');
header = {'NodeID' 'Avg Rank'};
sortmean = [idx' meanrnk'];
disp(' ')
ReplicationResults = [header; num2cell(sortmean)]
top10 = idx(1:10);

% Create histogram of node rankings
rnkdist = zeros(n,10);
for i = 1:10
   rnkdist(:,i) = rnkcnt(:,top10(i))./sum(rnkcnt(:,top10(i)),1);
   subplot(2,5,i)
   bar(1:n,rnkdist(:,i))
   axis([0.5 16 0 1.1])
   set(gca,'XTick',[1:n])
   string = sprintf('Node %.f',top10(i));
   title(string,'FontWeight','bold')
   ylabel('Rank Occurence')
   xlabel('Rank')
   hold off
end
```

**E.2 ULRA FUNCTION**

```
function [rnkcnt rnkfrac nodeRnks] = ULRA(kwnLinks,susLinks,r,n,repcount,CentralityType)
%Universal Likelihood Replication Analysis (ULRA) returns the distribution of centrality rankings for all nodes based on randomly iterating the removal of "r" suspected edges from the network.

%   Inputs:
%       kwnLinks,       Incidence list of known associations
%       susLinks,       Incidence list of suspected associations
%       r,              Number suspected edges to remove in each iteration
%       n,              Total number of nodes in the network
%       repcount,       Number of random iterations to be performed
%       CentralityType  1 = Degree, 2 = Closeness, 3 = Betweenness, or 4 = Eigenvector

%   Outputs:
%       rnkcnt,         Matrix of the number of times a node is ranked in each position
%       rnkfrac,        Matrix of the percentage of time a node is ranked in each position
%       nodeRnks,       Matrix of the node rankings for every iteration of node removal
%                       where column (i) represents the results of iteration (i)

k_ijs = sparse(kwnLinks(:,1),kwnLinks(:,2),kwnLinks(:,3),n,n);
kM = full(k_ijs) + full(k_ijs)';    %Adjacency Matrix of known associations

s_ijs = sparse(susLinks(:,1),susLinks(:,2),susLinks(:,3),n,n);
sM = full(s_ijs) + full(s_ijs)';    %Adjacency Matrix of suspected associations

estimate = kM + sM;             %Adjacency Matrix of known and suspected associations

% DEGREE CENTRALITY
if CentralityType == 1

initial_degree_rnk = degree_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount);idxmat = zeros(n,repcount);
for j=1:repcount
   ijs = ULRA_removal(susLinks(:,1),susLinks(:,2),r,n,susLinks(:,3));
   M = kM + full(ijs) + full(ijs)';    %Updated adjacency matrix with "r" edges removed
   degcent = sum(M)./(n-1);
   [srt, idxChg] = sort(degcent,'descend');
   idxRepeat = [false diff(srt) == 0];
   ChgRnk = cumsum(~idxRepeat)';
   rnkmat(:,j) = ChgRnk;
   idxmat(:,j) = idxChg;
end
```

```
nodeRnks = zeros(n,1);
for j=1:repcount
   for i=1:n
      nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
   end
end

ranks = [1:n];
for j=1:n
   for i=ranks
      rnkcnt = sum(nodeRnks(j,:)==ranks(i));
      rnkcntmat(i,j) = rnkcnt;
   end
end

rnkcnt = rnkcntmat;                 %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                            %Matrix of node rankings for every iteration of node removal

% CLOSENESS CENTRALITY
elseif CentralityType == 2

initial_closeness_rnk= closeness_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount); idxmat = zeros(n,repcount);
for j=1:repcount
   ijs = ULRA_removal(susLinks(:,1),susLinks(:,2),r,n,susLinks(:,3));
   M = kM + full(ijs) + full(ijs)';      %Updated adjacency matrix with "r" edges removed
   ccent = (n-1).*closeness(M);
   [srt, idxChg] = sort(ccent,'descend');
   idxRepeat = [false diff(srt)' == 0];
   ChgRnk = cumsum(~idxRepeat)';
   rnkmat(:,j) = ChgRnk;
   idxmat(:,j) = idxChg;
end

nodeRnks = zeros(n,1);
for j=1:repcount
   for i=1:n
      nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
   end
end

ranks = [1:n];
```

```
for j=1:n
   for i=ranks
      rnkcnt = sum(nodeRnks(j,:)==ranks(i));
      rnkcntmat(i,j) = rnkcnt;
   end
end

rnkcnt = rnkcntmat;                %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                          %Matrix of node rankings for every iteration of node removal

% BETWEENESS CENTRALITY
elseif CentralityType == 3

initial_betweenness_rnk = btwn_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount); idxmat = zeros(n,repcount);
for j=1:repcount
   ijs = ULRA_removal(susLinks(:,1),susLinks(:,2),r,n,susLinks(:,3));
   M = kM + full(ijs) + full(ijs)';      %Updated adjacency matrix with "r" edges removed
   BC = betweenness_wei(M);
   btwncent = BC/((n-1)*(n-2));
   [srt, idxChg] = sort(btwncent,'descend');
   idxRepeat = [false diff(srt)' == 0];
   ChgRnk = cumsum(~idxRepeat)';
   rnkmat(:,j) = ChgRnk;
   idxmat(:,j) = idxChg;
end

nodeRnks = zeros(n,1);
for j=1:repcount
   for i=1:n
      nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
end
end

ranks = [1:n];

for j=1:n
   for i=ranks
      rnkcnt = sum(nodeRnks(j,:)==ranks(i));
      rnkcntmat(i,j) = rnkcnt;
   end
end
```

```
rnkcnt = rnkcntmat;                %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                          %Matrix of node rankings for every iteration of node removal

% EIGENVECTOR CENTRALITY
elseif CentralityType == 4

initial_eigenvector_rnk = eig_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount); idxmat = zeros(n,repcount);
for j=1:repcount
  ijs = ULRA_removal(susLinks(:,1),susLinks(:,2),r,n,susLinks(:,3));
  M = kM + full(ijs) + full(ijs)';      %Updated adjacency matrix with "r" edges removed
  [V,D] = eig(M);
  [maxValue,index] = max(diag(D));  %# The maximum eigenvalue and its index
  maxVector = V(:,index);            %# The associated eigenvector in V
  eigcent = abs(maxVector)/sqrt(.5);
  [srt, idxChg] = sort(eigcent,'descend');
  idxRepeat = [false diff(srt)' == 0];
  ChgRnk = cumsum(~idxRepeat)';
  rnkmat(:,j) = ChgRnk;
  idxmat(:,j) = idxChg;
end

nodeRnks = zeros(n,1);
for j=1:repcount
  for i=1:n
    nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
end
end

ranks = [1:n];
for j=1:n
  for i=ranks
    rnkcnt = sum(nodeRnks(j,:)==ranks(i));
    rnkcntmat(i,j) = rnkcnt;
  end
end

rnkcnt = rnkcntmat;                %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                          %Matrix of node rankings for every iteration of node removal
end
end
```

**E.3 VLRA FUNCTION**

```
function [rnkcnt rnkfrac nodeRnks] = VLRA(kwnLinks,susLinks,edge_likelihood,n,repcount,Centrality)
%Variable Likelihood Replication Analysis (VLRA) returns the distribution of centrality rankings for all
nodes where suspected edges are randomly removed from the network in each replicate based on
individually assessed edge likelihood estimates.

%   Inputs:
%         kwnLinks,         Incidence list of known associations
%         susLinks,         Incidence list of suspected associations
%         edge_weights,     List of the estimated likelihood that an edge exists between two nodes
%         n,                Total number of nodes in the network
%         repcount,         Number of random iterations to be performed
%         CentralityType    1 = Degree, 2 = Closeness, 3 = Betweenness, or 4 = Eigenvector

%   Outputs:
%         rnkcnt,           Matrix of the number of times a node is ranked in each position
%         rnkfrac,          Matrix of the percentage of time a node is ranked in each position
%         nodeRnks,         Matrix of the node rankings for every iteration of node removal
%                           where column (i) represents the results of iteration (i)

k_ijs = sparse(kwnLinks(:,1),kwnLinks(:,2),kwnLinks(:,3),n,n);
kM = full(k_ijs) + full(k_ijs)';    %Adjacency Matrix of known associations

s_ijs = sparse(susLinks(:,1),susLinks(:,2),susLinks(:,3),n,n);
sM = full(s_ijs) + full(s_ijs)';    %Adjacency Matrix of suspected associations

estimate = kM + sM;              %Adjacency Matrix of known and suspected associations

% DEGREE CENTRALITY
if CentralityType == 1
initial_degree_rnk = degree_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount);idxmat = zeros(n,repcount);
for j=1:repcount
   ijs = VLRA_removal(susLinks(:,1),susLinks(:,2),edge_likelihood,n,susLinks(:,3));
   M = kM + full(ijs) + full(ijs)';      %Updated adjacency matrix with "r" edges removed
   degcent = sum(M);
   [srt, idxChg] = sort(degcent,'descend');
   idxRepeat = [false diff(srt) == 0];
   ChgRnk = cumsum(~idxRepeat)';
   rnkmat(:,j) = ChgRnk;
   idxmat(:,j) = idxChg;
end
```

```
nodeRnks = zeros(n,1);
for j=1:repcount
   for i=1:n
      nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
   end
end

ranks = [1:n];
for j=1:n
   for i=ranks
      rnkcnt = sum(nodeRnks(j,:)==ranks(i));
      rnkcntmat(i,j) = rnkcnt;
   end
end
rnkcnt = rnkcntmat;                    %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                               %Matrix of node rankings for every iteration of node removal

% CLOSENESS CENTRALITY
elseif CentralityType == 2
initial_closeness_rnk= closeness_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount); idxmat = zeros(n,repcount);
for j=1:repcount
   ijs = VLRA_removal(susLinks(:,1),susLinks(:,2),edge_likelihood,n,susLinks(:,3));
   M = kM + full(ijs) + full(ijs)';       %Updated adjacency matrix with "r" edges removed
   ccent = (n-1).*closeness(M);
   [srt, idxChg] = sort(ccent,'descend');
   idxRepeat = [false diff(srt)' == 0];
   ChgRnk = cumsum(~idxRepeat)';
   rnkmat(:,j) = ChgRnk;
   idxmat(:,j) = idxChg;
end

nodeRnks = zeros(n,1);
for j=1:repcount
   for i=1:n
      nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
end
end
```

```
ranks = [1:n];
for j=1:n
  for i=ranks
    rnkcnt = sum(nodeRnks(j,:)==ranks(i));
    rnkcntmat(i,j) = rnkcnt;
  end
end

rnkcnt = rnkcntmat;               %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                          %Matrix of node rankings for every iteration of node removal

% BETWEENESS CENTRALITY
elseif CentralityType == 3
initial_betweenness_rnk = btwn_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount); idxmat = zeros(n,repcount);
for j=1:repcount
  ijs = VLRA_removal(susLinks(:,1),susLinks(:,2),edge_likelihood,n,susLinks(:,3));
  M = kM + full(ijs) + full(ijs)';      %Updated adjacency matrix with "r" edges removed
  BC = betweenness_wei(M);
  btwncent = BC/((n-1)*(n-2));
  [srt, idxChg] = sort(btwncent,'descend');
  idxRepeat = [false diff(srt)' == 0];
  ChgRnk = cumsum(~idxRepeat)';
  rnkmat(:,j) = ChgRnk;
  idxmat(:,j) = idxChg;
end

nodeRnks = zeros(n,1);
for j=1:repcount
  for i=1:n
    nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
end
end

ranks = [1:n];

for j=1:n
  for i=ranks
    rnkcnt = sum(nodeRnks(j,:)==ranks(i));
    rnkcntmat(i,j) = rnkcnt;
  end
end
```

```
rnkcnt = rnkcntmat;                 %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                             %Matrix of node rankings for every iteration of node removal

% EIGENVECTOR CENTRALITY
elseif CentralityType == 4
initial_eigenvector_rnk = eig_rank_centrality(kwnLinks,susLinks,n)

rnkmat = zeros(n,repcount); idxmat = zeros(n,repcount);
for j=1:repcount
   ijs = VLRA_removal(susLinks(:,1),susLinks(:,2),edge_likelihood,n,susLinks(:,3));
   M = kM + full(ijs) + full(ijs)';      %Updated adjacency matrix with "r" edges removed
   [V,D] = eig(M);
   [maxValue,index] = max(diag(D));  %# The maximum eigenvalue and its index
   maxVector = V(:,index);               %# The associated eigenvector in V
   eigcent = abs(maxVector)/sqrt(.5);
   [srt, idxChg] = sort(eigcent,'descend');
   idxRepeat = [false diff(srt)' == 0];
   ChgRnk = cumsum(~idxRepeat)';
   rnkmat(:,j) = ChgRnk;
   idxmat(:,j) = idxChg;
end

nodeRnks = zeros(n,1);
for j=1:repcount
   for i=1:n
      nodeRnks(i,j) = rnkmat((find(idxmat(:,j)==i)),j); %Matrix of node rankings for every iteration
end
end

ranks = [1:n];
for j=1:n
   for i=ranks
      rnkcnt = sum(nodeRnks(j,:)==ranks(i));
      rnkcntmat(i,j) = rnkcnt;
   end
end
 rnkcnt = rnkcntmat;                 %Matrix of the number of times a node is ranked in each position
rnkfrac = rnkcntmat./repcount;  %Matrix of the percentage of time a node is ranked in each position
nodeRnks;                             %Matrix of node rankings for every iteration of node removal
end
end
```

### E.4 ULRA EDGE REMOVAL FUNCTION

```
function ijs = ULRA_removal(I,J,r,n,sS)
%ULRA_removal - Randomly removes a specified percentage of suspected edges

% Inputs:
%     I,      source nodes from the incidence list of suspected associations
%     J,      source nodes from the incidence list of suspected associations
%     r,      specified number of edges to remove
%     n,      total number of nodes in the network
%     sS,     edge values for suspected associations (%assume binary, but could use a valued)

% Output:
%     ijs,    updated incident list of suspected edges

idxperm = randperm(size(sS,1));  % permute a random sequence of index numbers for the incidence
matrix

sS(idxperm(1:r)) = deal(0);         % remove (i.e. change edge value to zero) for the first "r" nodes of
the random index

ijs = sparse(I,J,sS,n,n);
end
```

### E.5 VLRA EDGE REMOVAL FUNCTION

```
function ijs = VLRA_removal(I,J,edge_likelihood,n,S)
%VLRA_REMOVAL - Removes edges from the network based on the estimated likelihood that they
exist

% Inputs:
%     I,                Source nodes from an association incidence list
%     J,                Target nodes from an association incidence list
%     edge_likelihood,  Estimates for the likelihood that an association exists between nodes i and j.
%     S,                Edge values

% Outputs:
%       ijs,            Updated adjacency matrix after edges are removed

rand_weights = rand(size(edge_likelihood,2),1);

Sn = zeros(size(edge_likelihood,2),1);
in = [1:size(edge_likelihood,2)];
for i = in
  if rand_weights(i) < edge_likelihood(i)
  Sn(i) = S(i);
  else
  Sn(i) = 0;
  end
end
Sn;

ijs = sparse(I,J,Sn',n,n);
end
```

**E.6 DEGREE CENTRALITY RANK FUNCTION**

```
function [cent_rnk] = degree_rank_centrality(kwnLinks,susLinks,n)
%DEGREE_RANK_CENTRALITY - returns a cell array of degree centrality values and rankings for all nodes.

% Doctrinal Definition: Degree centrality describes how active an individual is in the network and is measured using the concept of degrees – the number of direct connections a node has (FM 3-24, December 2006)

%   Inputs:
%         kwnLinks,     Incidence list of known associations
%         susLinks,     Incidence list of suspected associations
%         n,            Total number of nodes in the network

%   Outputs:
%         cent_rnk,     Cell array of degree centrality values and rankings

% References:
% FM 3-24, Counterinsurgency

k_ijs = sparse(kwnLinks(:,1),kwnLinks(:,2),kwnLinks(:,3),n,n);
kM = full(k_ijs) + full(k_ijs)';    %Adjacency Matrix of known associations

s_ijs = sparse(susLinks(:,1),susLinks(:,2),susLinks(:,3),n,n);
sM = full(s_ijs) + full(s_ijs)';    %Adjacency Matrix of suspected associations

estimate = kM + sM;                 %Adjacency Matrix of known and suspected associations

basedeg = sum(estimate)./(n-1);     %Normalize by dividing by (n-1)
[initsrt, idxSrt] = sort(basedeg,'descend');
idxRepeat = [false diff(initsrt) == 0];
baseRnk = cumsum(~idxRepeat)';      %Rankings where equal value are given same rank
nodeID = idxSrt';                   %Ordered listing of nodes
init_rnk = [nodeID baseRnk initsrt'];
header = {'NodeID' 'Rank' 'Degree'};
cent_rnk = [header;num2cell(init_rnk)];
end
```

**E.7 CLOSENESS CENTRALITY RANK FUNCTION**

```
function [cent_rnk] = closeness_rank_centrality(kwnLinks,susLinks,n)
%CLOSENESS_RANK_CENTRALITY - returns a cell array of closeness centrality values and rankings for
all nodes.

%Definition: Closeness centrality is a measure of how close an actor is to all other actors in a set of
actors (p183) (Wasserman, 1994).

%   Inputs:
%        kwnLinks,      Incidence list of known associations
%        susLinks,      Incidence list of suspected associations
%        n,             Total number of nodes in the network

%   Outputs:
%        cent_rnk,      Cell array of degree centrality values and rankings

% Other routines used: closeness.m

k_ijs = sparse(kwnLinks(:,1),kwnLinks(:,2),kwnLinks(:,3),n,n);
kM = full(k_ijs) + full(k_ijs)';    %Adjacency Matrix of known associations

s_ijs = sparse(susLinks(:,1),susLinks(:,2),susLinks(:,3),n,n);
sM = full(s_ijs) + full(s_ijs)';    %Adjacency Matrix of suspected associations

estimate = kM + sM;                 %Adjacency Matrix of known and suspected associations

basecloseness = (n-1).*closeness(estimate);     %calculate normalized closeness
[initsrt, idxSrt] = sort(basecloseness,'descend');
idxRepeat = [false diff(initsrt)' == 0];
baseRnk = cumsum(~idxRepeat)';      %Rankings where equal value are given same rank
nodeID = idxSrt;                    %Ordered listing of nodes
init_rnk = [nodeID baseRnk initsrt];
header = {'NodeID' 'Rank' 'Closeness'};
cent_rnk = [header;num2cell(init_rnk)];
 end
```

### E.7.1 CLOSENESS CENTRALITY FUNCTION

NOTE: Code available at http://strategic.mit.edu/docs/matlab_networks/closeness.m

```
function C=closeness(adj)
% Computes the closeness centrality for every vertex: 1/sum (distance to all other nodes)
% Other routines used: simple_dijkstra.m

 C=zeros(length(adj),1);  % initialize closeness vector

for i=1:length(adj);
   d = simple_dijkstra(adj,i);
   d(find(d==inf))= length(adj);  % NOTE: Code modified so the distance between two disconnected
                                    nodes is equal to "n"
   C(i)= 1/sum(d);
end
```

### E.7.2 SIMPLE DIJKSTRA FUNCTION

NOTE: Code available at http://strategic.mit.edu/docs/matlab_networks/simple_dijkstra.m

```
function d = simple_dijkstra(adj,s)
% Implements a simple version of the Dijkstra shortest path algorithm. Returns the distance from a
single vertex to all others, doesn't save the path
% INPUTS: adjacency matrix (adj), start node (s)
% OUTPUTS: shortest path length from start node to all other nodes

n=length(adj);
d = inf*ones(1,n); % distance s-all nodes
d(s) = 0;    % s-s distance
T = 1:n;    % node set with shortest paths not found

while not(isempty(T))
   [dmin,ind] = min(d(T));
   for j=1:length(T)
      if adj(T(ind),T(j))>0 & d(T(j))>d(T(ind))+adj(T(ind),T(j))
         d(T(j))=d(T(ind))+adj(T(ind),T(j));
      end
   end
   T = setdiff(T,T(ind));
end
```

## E.8 BETWEENNESS CENTRALITY RANK FUNCTION

```
function [cent_rnk] = btwn_rank_centrality(kwnLinks,susLinks,n)
%BTWN_RANK_CENTRALITY - returns a cell array of betweenness centrality values and rankings for all nodes.

% Doctrinal Definition: Betweenness centrality indicates the extent to which an individual lies between other individuals in the network, serving as an intermediary, liaison, or bridge. A node with high "betweenness" has great influence over what flows in the network.  Depending on position, a person with high betweenness plays a "broker" role in the network.  A major opportunity exists for counterinsurgents if the high betweenness centrality person  is also a single point of failure which, if removed, would fragment the organization.

%   Inputs:
%        kwnLinks,    Incidence list of known associations
%        susLinks,    Incidence list of suspected associations
%        n,           Total number of nodes in the network
%
%   Outputs:
%        cent_rnk,    Cell array of betweenness centrality values and rankings

% Other routines used: betweenness_wei.m

k_ijs = sparse(kwnLinks(:,1),kwnLinks(:,2),kwnLinks(:,3),n,n);
kM = full(k_ijs) + full(k_ijs)';    %Adjacency Matrix of known associations

s_ijs = sparse(susLinks(:,1),susLinks(:,2),susLinks(:,3),n,n);
sM = full(s_ijs) + full(s_ijs)';    %Adjacency Matrix of suspected associations

estimate = kM + sM;                 %Adjacency Matrix of known and suspected associations

BC = betweenness_wei(estimate);     %Betweenness MATLAB function available on googlecode.com
   basebtwn = BC/((n-1)*(n-2));
[initsrt, idxSrt] = sort(basebtwn,'descend');
idxRepeat = [false diff(initsrt)' == 0];
baseRnk = cumsum(~idxRepeat)';      %Rankings where equal value are given same rank
nodeID = idxSrt;                    %Ordered listing of nodes
init_rnk = [nodeID baseRnk initsrt];
header = {'NodeID' 'Rank' 'Betweenness'};
cent_rnk = [header;num2cell(init_rnk)];
end
```

### E.8.1 BETWEENNESS FUNCTION

NOTE: Code available at
http://visualconnectome.googlecode.com/svn-history/r2/trunk/Plugins/BCT/betweenness_wei.m

```
function BC=betweenness_wei(G)
%BETWEENNESS_WEI    Node betweenness centrality

%   Node betweenness centrality is the fraction of all shortest paths in the network that contain a
given node. Nodes with high values of betweenness centrality participate in a large number of
shortest paths.

%   Input:      L,     Directed/undirected connection-length matrix.
%   Output:    BC,     node betweenness centrality vector.

%   Notes: The input matrix must be a connection-length matrix, typically obtained via a mapping
from weight to length. For instance, in a weighted correlation network higher correlations are more
naturally interpreted as shorter distances and the input matrix should consequently be some inverse
of the connectivity matrix. Betweenness centrality may be normalised to the range [0,1] as BC/[(N-
1)(N-2)], where N is the number of nodes in the network.

%   Reference: Brandes (2001) J Math Sociol 25:163-177.
%   Mika Rubinov, UNSW, 2007-2012

n=length(G);
E=find(G); G(E)=1./G(E);                %invert weights
BC=zeros(n,1);                          %vertex betweenness

for u=1:n
  D=inf(1,n); D(u)=0;                   %distance from u
  NP=zeros(1,n); NP(u)=1;               %number of paths from u
  S=true(1,n);                          %distance permanence (true is temporary)
  P=false(n);                           %predecessors
  Q=zeros(1,n); q=n;                    %order of non-increasing distance

  G1=G;
  V=u;
  while 1
    S(V)=0;                             %distance u->V is now permanent
    G1(:,V)=0;                          %no in-edges as already shortest
    for v=V
      Q(q)=v; q=q-1;
      W=find(G1(v,:));                  %neighbors of v
      for w=W
        Duw=D(v)+G1(v,w);               %path length to be tested
```

```
          if Duw<D(w)                    %if new u->w shorter than old
            D(w)=Duw;
            NP(w)=NP(v);                 %NP(u->w) = NP of new path
            P(w,:)=0;
            P(w,v)=1;                    %v is the only predecessor
          elseif Duw==D(w)               %if new u->w equal to old
            NP(w)=NP(w)+NP(v);           %NP(u->w) sum of old and new
            P(w,v)=1;                    %v is also a predecessor
          end
        end
      end

      minD=min(D(S));
      if isempty(minD), break            %all nodes reached, or
      elseif isinf(minD),                %...some cannot be reached:
        Q(1:q)=find(isinf(D)); break     %...these are first-in-line
      end
      V=find(D==minD);
    end

    DP=zeros(n,1);                       %dependency
    for w=Q(1:n-1)
      BC(w)=BC(w)+DP(w);
      for v=find(P(w,:))
        DP(v)=DP(v)+(1+DP(w)).*NP(v)./NP(w);
      end
    end
  end
```

### E.9 EIGENVECTOR CENTRALITY RANK FUNCTION

```
function [cent_rnk] = eig_rank_centrality(kwnLinks,susLinks,n)
%EIG_RANK_CENTRALITY - returns a cell array of eigenvector centrality values and rankings for all
nodes.

%Definition:  Eigenvector centrality is a measure of popularity in the sense that a node with high
eigenvector centrality is connected to nodes that are themselves well connected (Borgatti, Martin,
& Johnson, Analyzing Social Networks, 2013).

%   Inputs:
%          kwnLinks,      Incidence list of known associations
%          susLinks,      Incidence list of suspected associations
%          n,             Total number of nodes in the network

%   Outputs:
%          cent_rnk,      Cell array of eigenvector centrality values and rankings

k_ijs = sparse(kwnLinks(:,1),kwnLinks(:,2),kwnLinks(:,3),n,n);
kM = full(k_ijs) + full(k_ijs)';    %Adjacency Matrix of known associations

s_ijs = sparse(susLinks(:,1),susLinks(:,2),susLinks(:,3),n,n);
sM = full(s_ijs) + full(s_ijs)';    %Adjacency Matrix of suspected associations

estimate = kM + sM;              %Adjacency Matrix of known and suspected associations

[V,D] = eig(estimate);
[maxValue,index] = max(diag(D));  %# The maximum eigenvalue and its index
 maxVector = V(:,index);           %# The associated eigenvector in V
   baseeig = abs(maxVector)/sqrt(.5);
[initsrt, idxSrt] = sort(baseeig,'descend');
idxRepeat = [false diff(initsrt)' == 0];
baseRnk = cumsum(~idxRepeat)';      %Rankings where equal value are given same rank
nodeID = idxSrt;                    %Ordered listing of nodes
init_rnk = [nodeID baseRnk initsrt];
header = {'NodeID' 'Rank' 'Eigenvector'};
cent_rnk = [header;num2cell(init_rnk)];

end
```