A QUANTITATIVE METHODOLOGY FOR VETTING "DARK NETWORK" INTELLIGENCE SOURCES FOR SOCIAL NETWORK ANALYSIS

DISSERTATION

James F. Morris

AFIT/DS/ENS/12-05

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

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

AFIT/DS/ENS/12-05

A QUANTITATIVE METHODOLOGY FOR VETTING "DARK NETWORK" INTELLIGENCE SOURCES FOR SOCIAL NETWORK ANALYSIS

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

James F. Morris, BS, MS

Department of the Air Force

June 2012

AFIT/DS/ENS/12-05

A QUANTITATIVE METHODOLOGY FOR VETTING "DARK NETWORK" INTELLIGENCE SOURCES FOR SOCIAL NETWORK ANALYSIS

James F. Morris, BS, MS
Department of the Air Force

Approved:

_____//signed//_____	31 May 2012
Richard F. Deckro, DBA (Chairman)	Date
_____//signed//_____	31 May 2012
Darryl K. Ahner, PhD (Member)	Date
_____//signed//_____	31 May 2012
Dursun A. Bulutoglu, PhD (Member)	Date
_____//signed//_____	31 May 2012
Lt Col Jonathan Todd Hamill, PhD (Member)	Date

Accepted:

_____//signed//_____	20 June 2012
Dr. Marlin U. Thomas Dean, Graduate School of Engineering and Management	Date

Abstract

Social network analysis (SNA) is used by the DoD to describe and analyze social networks, leading to recommendations for operational decisions. However, social network models are constructed from various information sources of indeterminate reliability. Inclusion of unreliable information can lead to incorrect models resulting in flawed analysis and decisions. This research develops a methodology to assist the analyst by quantitatively identifying and categorizing information sources so that determinations on including or excluding provided data can be made.

This research pursued three main thrusts. It consolidated binary similarity measures to determine social network information sources' concordance and developed a methodology to select suitable measures dependent upon application considerations. A methodology was developed to assess the validity of individual sources of social network data. This methodology utilized source pairwise comparisons to measure information sources' concordance and a weighting schema to account for sources' unique perspectives of the underlying social network. Finally, the developed methodology was tested over a variety of generated networks with varying parameters in a design of experiments paradigm (DOE). Various factors relevant to conditions faced by SNA analysts potentially employing this methodology were examined. The DOE was comprised of a 2^4 full factorial design augmented with a nearly orthogonal Latin hypercube. A linear model was constructed using quantile regression to mitigate the non-normality of the error terms.

To my wife and sons

Acknowledgements

My education pursuit was possible thanks to several gracious benefactors, the DAGSI, SMART, and Bonder Scholarships. My employer, the National Air & Space Intelligence Center, led by Mr. O'Connell's, Mr. Fuell's, Mr. Rowland's, and Rob's advocacy, permitted full time study—absolutely essential to this success. I hold sincere appreciation for the numerous coworkers and others who provided constant encouragement. Mr. J, Hal and the Mick provided the intellectual stimulation for me to begin this journey and I thank you.

In the course of this journey, the support of my fellow students at AFIT and Nick were indispensable. My committee and the AFIT faculty constantly challenged me to expand my capabilities. Their commitment to excellence is inspiring. This dissertation would not have been possible without the dedication, direction, questions, and insight of my advisor, Dr. Deckro.

Finally, I wish to thank my family for being so understanding and supportive throughout this experience. My kids' affections, despite Dad's absences, were the best way to end a long day. My completion would not have been realized if hadn't been for my wife's devotion and support. Her repeated sacrifices and consistent encouragement enabled me to finally finish this endeavor.

James F. Morris

Table of Contents

List of Figures

List of Tables

Page

A QUANTITATIVE METHODOLOGY FOR VETTING "DARK NETWORK" INTELLIGENCE SOURCES FOR SOCIAL NETWORK ANALYSIS

I. Introduction

The initial decade of the 21st century has been characterized by the United States in direct conflict with terrorist organizations and insurgent groups, while attempting to mitigate the effects of organized criminal enterprises, drug cartels, human trafficking, piracy, and cyber crime. These entities utilize support networks composed of money laundering, weapons smuggling, illegal technology proliferation and other illicit activities. Dealing with this myriad of interconnected organizations and activities has led to the development of nontraditional analytic techniques in support of strategies addressing these threats to national security. One such analytic technique brought to bear on this problem set is Social Network Analysis (SNA), not necessarily a new technique, but novel in its relatively recent application to the national security arena. As such, the Department of Defense's (DoD) initial unfamiliarity with Social Network Analysis has now transitioned to various instantiations in levels of application and expertise in numerous DoD organizations. The DoD's use of SNA as a military tool against an array of organizations has delivered successes and failures in providing useful analysis on the target subject's inner workings and identifying strategies to inhibit these targets.

Social Network Analysis is a quantitative methodology to model networked actors' behavior. SNA focuses on the relationships among actors and the implications on both collective and individual behavior resulting from the structure of the network and

the patterns in the relationships. The network structure portrays the pattern of relationships among the actors and models organized collective behavior. This structure can affect, promote, and constrain individual actor behavior (Wasserman & Faust, 1994, pp. 3-4). The quantitative nature of SNA methodology enables the characterization of the network structure and its implications upon collective and individual behavior, identification of actors significantly involved in organized behavior, and the determination of groups of actors contained within the structure. Additionally, the quantitative basis enables detection of changes over time in network structure, actor prominence, and group formulation or dissolution (Wasserman & Faust, 1994, pp. 9-10).

The network models used in this type of analysis are dependent upon the veracity of the information sources providing social network data. Social network information sources may provide unreliable information leading to inaccurate conclusions from the model. The information sources may confirm or discredit reports from other sources, leaving the SNA analysts to arbitrate what data is used in the social network model. This research addresses the lack of suitable quantitative methodological approaches to aid SNA analysts facing this complex problem.

SNA methodologies have predominantly evolved from research conducted on open networks such as businesses, governmental organizations, social groups, and activities where data is voluntary provided or permissibly collected. In contrast, adversarial organizations considered threats to U.S. national security are structured and have mechanisms emplaced minimizing the effectiveness of traditional social science SNA data collection techniques. These organizational mechanisms present additional challenges in applying SNA to these problem sets. To address the associated

implications on modeling, a conceptual understanding of these organizations and associated mechanisms is required.

1.1 Dark Networks

The inherent complexity of dealing with the wide range of organizations subject to interest by the DoD is beginning to be deciphered by several characterization and generalization efforts initialized in the academic realm which show promising utility in addressing the problem set. One such notion, coined "dark networks," serves as a basis for a conceptual framework suitable to categorize and address the range of potential organizational adversaries faced by the DoD.

Raab and Milward (2003) introduced and defined "dark networks" as actors and organizations that cooperate in activities that are both covert and illegal, in contrast to "bright networks", formally defined as "a legal and overt governance form that is supposed to create benefits for the participating actors and to advance the common good and does not—at least intentionally—harm people"; dark networks' illegal activities are not meant to be visible (Raab & Milward, 2003, p. 419; Milward & Raab, 2006, p. 334). The preponderance of related academic literature examining group and organizational behaviors is derived from research centered in characterized social interactions and improving organizational efficiency and effectiveness of bright networks. Efforts extending research findings derived from this academic literature for application against dark networks has grown dramatically since September 11, 2001.

Dark networks as a framework appear to describe and address the litany of adversaries faced by the United States. Applying organizational theory to identify and

model vulnerabilities in dark network organizations may enable more efficient utilization of limited military and other agencies' resources used in reducing or eliminating various dark networks' capabilities. The difficulty lies in accurately distilling generalizations, the conditions for which they are applicable, and their implications from the spectrum of organizations characterized as dark networks.

Dark networks differ from overt "bright" networks in several dimensions. Stemming from dark network members' desires for their operations to generally remain undetected, their structures contain specific aspects and characteristics ensuring a sustainable amount of security and organizational resilience. Some of these structural characteristics are intentional designs, while others are a function of how the networks form and evolve over their lifespan. The security needs of dark networks manifest themselves intra-organizationally, in the relationships among members; inter-organizationally, between various organizations involved with dark network activities; and externally, as dark networks interface with the general population.

Intra-organizationally, members' relations are defined by trust due to the risks they incur for participation in illegal activities, or the mere association of being members of the illicit organization (Erickson, 1981, p. 195). Binding the organization together, "integration is primarily based on trust relations between individual persons and their complementary interest (Raab & Milward, 2003, p. 432)." These trust relations can be established and reinforced within the organization via ideological commitments, indoctrination, joint participation in activities, common fate sentiment, and other socialization processes. Due to the high need for security and its resultant demand for trust among an organization's members, "risk enforces recruitment along lines of trust

and, thus, through preexisting networks of relationships, which set the limits of the secret society's structure (Erickson, 1981, p. 188)." Dark network organizations' growth and ability to reconstitute its personnel is dependent upon its members' other social networks, which are not necessarily based in illegal behavior, but may stem from familial, educational, geographical, and other social contexts.

Members of dark network organizations, who are generally included due to preexisting social connections, create an environment which reinforces continued affiliation and activity with the organization. "Because risk is such a big factor, professional and personal lives are intermeshed and almost indistinguishable (Raab & Milward, 2003, p. 431)." Despite these personal connections potentially existing among members, the overall "structure of covert networks will tend to be as sparse as possible to achieve the goals of the participating actors (Raab & Milward, 2003, p. 433)." Dark network organizations structure themselves in a manner to insulate and limit possible damage due to individual defections, arrest, capture, or compromise. For these organizations, security is paramount and crucial for them to continue to conduct operations in pursuit of their goals, and their organizational structure reflects that concern.

Dark networks possess loose connections that drive interactions, mutually beneficial behavior and cooperation among organizations engaged in illegal activity. Organizations conducting illicit activities require resources; organizational specialization and market segmentation has occurred similar to overt networks. Illegal narcotic production, smuggling and distribution organizations may rely upon weapons smuggling organizations to supply arms. Organizations engaged in kidnapping may utilize

organizations specializing in money laundering to handle that aspect of the operation. Enabling these transactions are "actors who function as brokers between these different networks (Raab & Milward, 2003, p. 431)." However, these interactions among dark networks and their composite organizations do not obviate the need for security. Similarly to specific organizations structuring themselves to promote security, the transactions, interconnectivity, alliances and cooperation, occurring within dark networks are structured under a security conscious paradigm. "Dark networks try to function with as few ties as possible (Milward & Raab, 2006, p. 353)." Minimizing inter-organizational connections insulates individual organizations from potential compromise or exposure from necessary transactions and interactions.

Dark networks' preoccupation with security and the nature of their activities present a unique challenge to inter-organizational transactions. "Dark networks cannot rely on formal institutions or the legal systems for dispute resolution (Raab & Milward, 2003, p. 430)." As such, "persuasion, exchange, and negotiation are the central mechanisms for management and conflict resolution in overt networks, coercion and physical force are the distinctive characteristic of covert networks (Raab & Milward, 2003, p. 432)." Due to this, "transaction costs in covert networks are higher than those in overt networks (Raab & Milward, 2003, p. 432)." Despite these deterrents to inter-organizational interaction within dark networks, it occurs due to organizations' necessity of acquiring resources, sometimes unique resources, in order to achieve their aims.

The United States' National Security establishment has in some instances dealt with dark networks over a substantial amount of time. Milward and Raab (2006, p. 336, 351) note this "resilience of these [dark] networks in the face of massive efforts to control

them," defining resilience as "the ability to recover from or to resist being affected by some shock, insult, or disturbance." In the dark networks' organizational paradigm, resilience is the ability of the organization "to avoid disintegration when coming under stress (Milward & Raab, 2006, p. 351)." Despite numerous governmental programs and efforts to inhibit dark networks, it appears to be an interminable problem.

1.2 SNA's Utility in Analyzing Dark Networks

The desire to understand dark network organizations via identifying roles and their associated individuals brings forth the need for a methodological approach. One methodology displaying high promise for fulfilling this analytic gap is Social Network Analysis (SNA). Several studies have examined and proposed utilizing SNA to study dark networks, covering the gamut from organized criminal networks to insurgencies and terrorist groups (Sparrow, 1991; Coles, 2001; Reed, 2006; Ressler, 2006; van der Hulst, 2009) with several introducing new methodologies and algorithms specifically adapted to these problem sets (Renfro, 2001; Sterling, 2004; Clark, 2005; Hamill, 2006; Farley, 2007; Geffre, 2007; Herbranson, 2007; Seder, 2007; Leinart, 2008; Kennedy, 2009). Acquired knowledge of dark networks' and their organizations' structures, operations' processes and mechanisms, coupled with discerning roles and associated individuals has the potential to grant understanding into inherent vulnerabilities that are susceptible to exploitation.

Social Network Analysis is an analytic methodology encompassing social network data collection methodologies through mathematical analysis and visualization. Social network data is collected, compiled and filtered to create a social network model

representing underlying social dimensions of interactions among actors scoped to answer specific questions. A variety of mathematical techniques are then applied to analyze the social network model to generate insights into phenomena of operational significance. Generally, Social Network Analysis is employed to: model actors and their relationships to depict the structure of the network, analyze the impact of this structure on the function of the network, analyze the impact of this structure on individuals within the network, and assess changes over time (Wasserman & Faust, 1994, pp. 3, 9-10).

Of significance, the social network model is constructed under a specific context. The data is collected and the model is generally constructed for a specific analytic purpose, a set of "real world" questions on which the subsequent analysis is attempting to shed light. This context and the associated questions drive the analyst's selection of SNA analytic techniques to apply to the social network model. As these analytic techniques are mathematically based, they each have associated assumptions. These assumptions precipitate the social network data requirements in order to satisfy the mathematical requirements of the chosen SNA analytic techniques. The conclusions and interpretations derived from the SNA analytic techniques are dependent upon the contextual considerations of the social network model as well as the collected social network data.

1.3 Problem Context

In order to fully and accurately characterize a dark network organization, investigations must expand beyond purely "professional" relationships within an organizational context and explore other relation types, such as familial and social, to

ensure acquiring all relevant and pertinent information. This requires a host of sources to provide the necessary data, of which several may be nontraditional or on the forefront of technical collection capabilities, burdened with additional encumbrances due to issues of data availability and/or legal authority to collect. Expanding to personal social contexts introduces the problem of determining when the boundary of the network has been reached.

Data collection is further confounded as dark network organizations are purposely attempting to remain opaque. Data collection is conducted possibly in the face of adversarial active denial and deception measures. Social Network Analysis of dark networks will surely be applied in an imperfect data situation, with certain data elements missing and others being corrupted or inaccurate.

Social Network Analysis suffers from a lack of standardized adequacy criteria for data collection. It is indeterminate whether enough data is present and appropriate to conduct a SNA or rely on its results in decision making. This lack of criteria has left SNA analysts to self-determine when sufficient data has been collected to perform a SNA with limited intuition or guidelines of the corollary impact upon analytic results. Analytical conclusions are drawn and presented on data sets that may provide erroneous results due to an indeterminate significant amount of missing data or data corruption. These errors may be significant as they could produce analytic results that are counter to the true situation, leading to misappropriation of resources, improper strategy adoption, and erroneous targeting.

Methods exist within academic literature for imputing missing data (Leinart, 2008) and for attempting to probabilistically classify collected data as valid or invalid

(Butts, 2003). However, there is no benchmark of when these methods should be applied in SNA. A method currently does not exist to determine if a sufficient amount of data has been collected to perform SNA appropriately, inevitably resulting in incorrect conclusions being drawn from misapplication. While undesirable, this may be acceptable if one is conducting a sociological study; however, it can be disastrous if applied to the national security arena and influence operations (law enforcement, military or civil.)

In the arena of national security, erroneous results stemming from Social Network Analysis may produce substantial unintended consequences. Decisions derived from the analysis may expend critical resources without generating any associated progress towards achieving national security objectives. More damaging than inefficient utilization of resources, erroneous results precipitating unacceptable collateral damage, such as actions against innocent individuals and organizations, could hinder and restrain current and future operations. Such mishaps may drive previously neutral individuals into dark network organizations or at least increase their support of the groups' goals.

Particularly when dealing with dark networks, improper strategy or poor targeting could potentially provide a strategic advantage to targeted organizations; conducting "organizational Darwinism" by removing non-effective actors from an organization may strengthen the organization while improving its overall effectiveness and ability to accomplish its objectives. A possible example is large-scale arrests of low-level incompetent criminals in organized crime, leading to an organization comprised predominantly of very effective and efficient actors. Dark network organizations are comprised of human beings, and as such, successful ones respond, adapt, and exhibit creativity when presented with adversity. Government forces applying improper

strategies or targeting solutions could lead to adversary organizational alterations generating significant unintended consequences. Hypothetically, removing adversary leadership personnel may lead to their replacement with individuals who are characterized as even less desirable, i.e. more violent, intelligent, dedicated, inspirational, and so forth.

1.4 Problem Statement

Social network analysis is used by the DoD and other government agencies to describe and analyze social networks, leading to recommendations for operational decisions. However, social network models are constructed from various information sources of indeterminate reliability. Inclusion of unreliable information can lead to incorrect models resulting in flawed analysis and decisions.

It is critical for continued use of SNA by the DoD that methods are developed to determine if information sources reporting social network data are of sufficient reliability to allow the use of SNA techniques generating solutions within acceptable operational risk. Formalized SNA includes various methodologies for social network data collection. The collected data must be contextually relevant and of sufficient quality and quantity. The contextual relevancy is a function of the specific problem set and the instantiated social network model. Absent in Social Network Analysis are methods to determine if information sources are reporting data of sufficient quality and quantity to conduct SNA appropriately. This gap is currently addressed via the Social Network analyst's intuition on whether sufficient reliable data exists and is available to conduct the analysis or further data collection is warranted, or data imputation techniques need to be employed.

Analogous to confidence intervals in statistics, presenting SNA conclusions without an associated degree of confidence does not fully answer and address the questions precipitating the analysis.

Current practice assumes that all social network analytic techniques are applicable under all data conditions despite growing evidence in the academic literature to the contrary. Additionally, traditional application of SNA treats all obtained information on the social network equally, regardless of the acquisition means or source, depicted in **Error! Reference source not found.**. Adopting a term from the computer science discipline, data provenance refers to the origins of a piece of data and the process by which it was obtained (Buneman, Khanna, & Tan, 2000, p. 88). It is presumed that some information sources providing social network data are more reliable in that their data is generally a more accurate representation of the social network under observation. Information sources established as being reliable should be considered differently than unreliable or untested sources in the construction of the social network model, portrayed in Figure I-2, particularly in the case of conflicting reporting. The DoD utilizes SNA on problem sets, such as dark networks, with social network data acquired through various collection means. Due to the nature of the problem set, SNA applied in this context has additional concerns of reliable, unreliable, and deceptive sources of information.

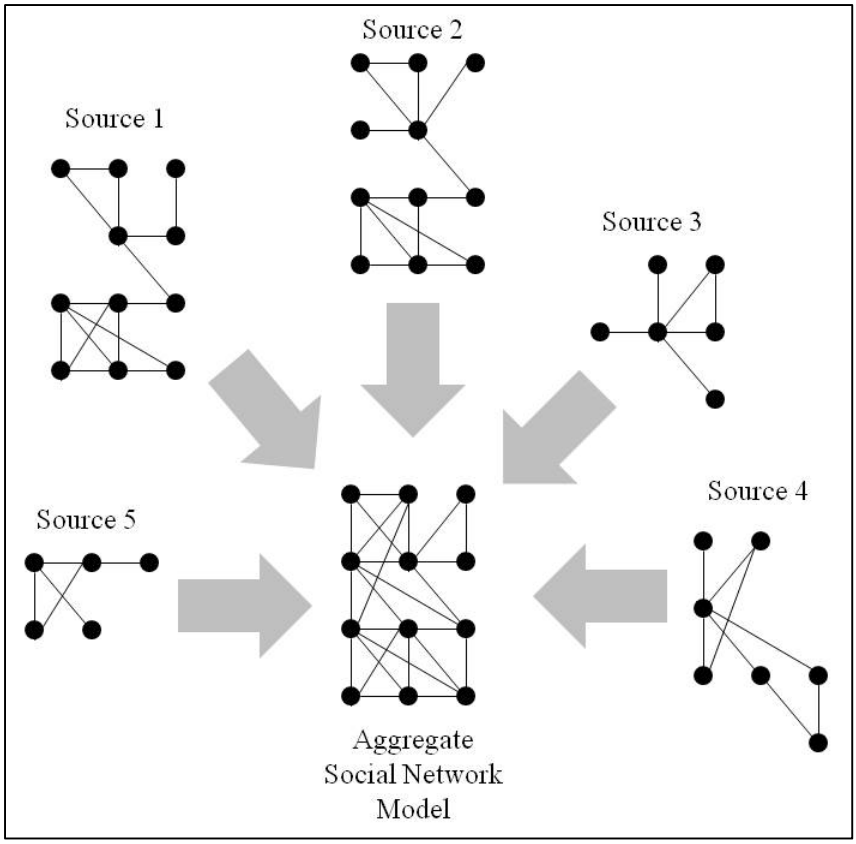

Figure I-1 SNA Model Construction – Current Practice

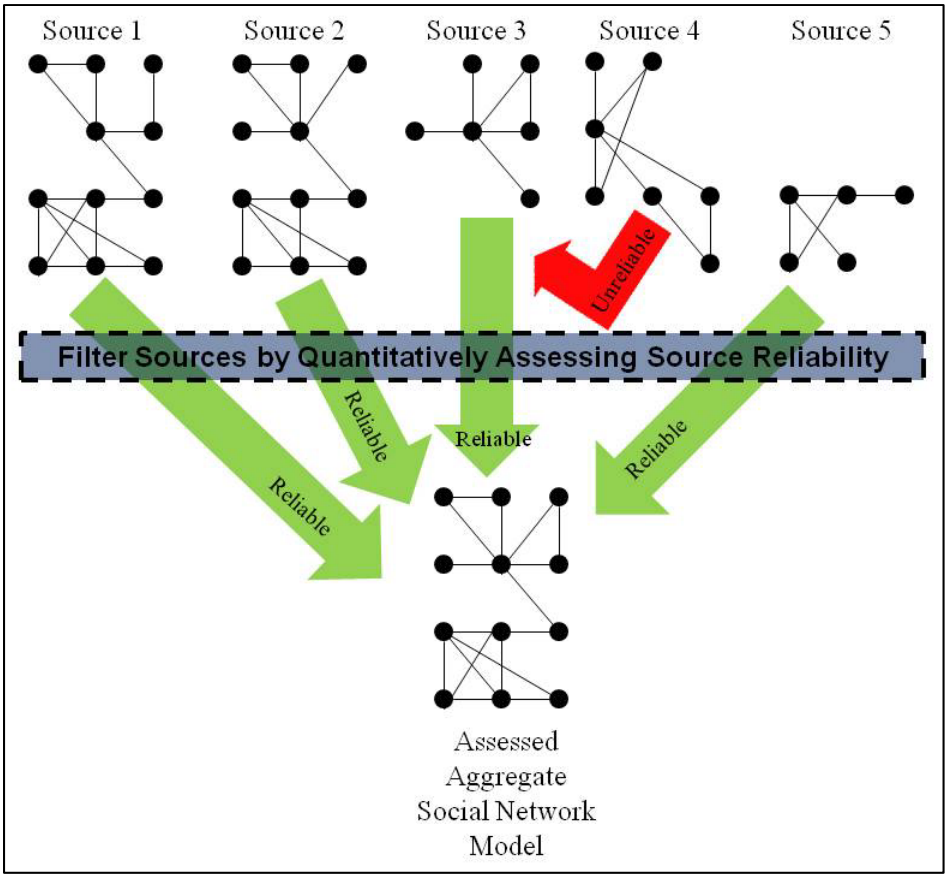

Figure I-2 Constructing SNA Models Considering Source Reliability

1.5 Research Objectives

This dissertation conducts a line of research investigating the construction of a social network model in the face of reliable and unreliable information sources. Along those lines, ***this research developed a methodology to quantitatively identify and categorize information sources so that determinations on including or excluding their provided data can be made***. This research proceeded with the following goals:

- ***Consolidate similarity measures to determine social network information sources' concordance and develop a methodology to select suitable measures dependent upon application considerations.*** Many of these similarity measures

have been introduced across various disciplines and no complete consolidated listing is available. Despite numerous similarity measures existing in the literature, no guidelines exist for selecting suitable measures for a given application. Developing a methodology to select similarity measures for social network information source comparisons will enable quantitative means to evaluate confirmations and dissentions among sources.

- ***Develop a methodology to assess validity of individual sources of social network data.*** One method of constructing a social network data set is compiling information from various sources. When examining dark networks, it is imperative to consider that due to the nature of organizations involved some sources will be delivering only limited network perspectives, i.e. imperfect data, as well as professing corrupt data, either intentionally or unintentionally. Evaluating and verifying various sources will provide a means to construct a social network data set, hopefully, minimizing the impact of imperfect data by appropriately weighting sources that provide verified information.
- ***Test the methodology over a variety of generated networks with varying parameters in a design of experiments paradigm.*** Dark networks may appear in various regimes of network parameter space. Identifying network parameter subspaces and examining SNA applicability for networks contained within those subspaces will enable assessment of the appropriateness of specific SNA techniques on dark network organizations.

1.6 Dissertation Overview

This dissertation is organized as follows: Chapter II provides a literature review of Social Network Analysis with a focus on the impact of imperfect data, which includes missing and corrupt data elements. Chapter III discusses the methodological approach to address the social network information source assessment problem introduced in this introduction. Additionally, Chapter III provides the experimental design to be employed to assess the methodology's performance. Chapters IV and V discuss in detail components of the methodology and present analytical results of the experimentation. Chapter VI employs the developed methodology in a case study format for demonstration purposes. Chapter VII reviews the contributions of this research, discusses assumptions and limitations, and indicates future research threads to explore.

II. Literature Review

This chapter begins with an introduction to the modeling aspects of social network analysis applicable to dark networks, followed by a brief overview of several SNA measures referenced in imperfect data literature. Next, a discussion of the social science underpinnings and the relation to social network modeling is presented. A detailed review of the academic literature describes current efforts to date on the impact of imperfect data on social network analysis is then provided. Methodologies in the literature to address the issues associated with imperfect data in SNA, namely consensus structure aggregation and a Bayesian approach, are presented with discussions of their limitations. Following the social science underpinnings and the impact of imperfect data, a discussion of methods and techniques necessary for the methodology presented in Chapter III is discussed. Social network data sources are explored, followed by statistical methods to measure source agreement in reporting. Next, classifier performance metrics are described. Finally, statistical analysis techniques employed in Chapter IV are described.

2.1 Social Network Analysis

Social network analysis focuses on relationships among entities. This allows inferences to be drawn from patterns of relationships or the implications of certain structures upon actor behavior as well as the impact of actors upon other entities in a social network structure (Wasserman & Faust, 1994, p. 3). Social network analysis models social interactions among entities as a network with mathematical formulization, enabling algorithms, procedures and computations based in social science theory.

Presented here is a brief overview of social network analysis, some of its components, and commonly used algorithms and pertinent calculations in understanding the impact of imperfect data on SNA results.

2.1.1 Actors.

Various social entities can be envisioned as actors within a social network. Countries, organizations, groups, social units, or individuals can be modeled as actors in social network analysis. If the social network analysis is constrained to one type of actor, for example only deals with individuals, it is defined as a one-mode network. If the analysis contains two actor types, such as individuals and their affiliations with organizations, the model is defined as a two-mode network (Wasserman & Faust, 1994, p. 17). Predominantly, social network analysis is applied to one-mode networks. They are modeled as nodes, or vertices, in a network graph representation.

2.1.2 Relationships.

Actors are connected via relationships, also referred to as relational ties. A relationship defines the linkage between actors. There are many different kinds of relationships reflected in social network analysis, which are categorized by relation type summarized in Table II-1. A relationship between two actors creates a structure termed a dyad. The relationships among three actors are referred to as a triad. Of note, there are several combinations of relationship pairings that could constitute a triad, i.e. only a subset of the potential relationships exist between all pairing within the triad (Wasserman & Faust, 1994, p. 18). Relationships are modeled as arcs or edges in a network graph representation.

Table II-1 Relation Types

Relation Type	Description (Examples)
Individual evaluations	Measurements of positive or negative affect for another actor; sometimes referenced as sentiment. (Ex: friendship, respect)
Transfer of material resources	Transfer of goods, specific forms of social support. (Ex: exchanges of gifts)
Transfer of non-material resources	Communication, sending/receiving information. (Ex: sending a message)
Interaction	Physical interaction of actors; presence at the same place at the same time. (Ex: sitting next to another actor, two actors attending the same meeting)
Movement	Physical movement; social movement. (Ex: changing location, change in social status)
Formal roles	Power and authority between actors. (Ex: boss/employee relationship)
Kinship	Familial and marriage relationships. (Ex: parent, spouse)

(Wasserman & Faust, 1994, pp. 37-38)

2.1.3 Relations.

Social relations among actors may be based upon a perception of a relationship; referred to in the literature as the cognitive network (Wasserman & Faust, 1994, p. 51). Relations that are determined by perception have significant implications upon the relationships among actors that are ascertained for an analysis. Perceived ties may be more appropriate for analysis conducted on phenomenon such as influence, attitude or opinion development and propagation through a network's actors. On the contrary, Marsden suggests that relations defined by interactions or transfers of goods or information may be more appropriate for analysis on diffusion of material through a network (Marsden, 1990, p. 437).

A temporal aspect to relations may exist within a social network. Relationships may be episodic, transient, or a single-occurrence between actors, or be based upon recurrent interactions or exchanges (Marsden, 1990, p. 437). This temporal nature of relationships may necessitate a scoping of the analysis to consider only a specified time frame of activity. Additionally, thresholds, based upon transaction or interaction frequency or intensity, may need to be established to confirm the presence of a relationship at a significant level between two actors for inclusion into the analysis.

2.1.3.1 Directed Relations.

Some relation types may imply directionality. As many of the relation types involve a transfer of a resource or information, a direction of the relationship is defined by the sender and the receiver of the resource. Additionally, non-transfer relations can involve a direction. For example, a boss giving orders to a subordinate implies a direction of the relationship, in this case the boss exerting authority over the subordinate (Wasserman & Faust, 1994, pp. 121-122).

2.1.3.2 Asymmetric Relations.

As a result of directed relations, it is conceivable that a relationship between one actor and another is not reciprocated. An example is a relation type that involves choice. If actors choose the relationship with another actor, the other actor may choose to not respond with the same relationship back to the sender. For relation types that are based upon transfer, a sender could pass resources to a recipient and the recipient may not transfer resources back to the original sender. For non-transfer relations, such as

affection, it is possible for an actor to express a sentiment, such as love or respect, for another actor that is not returned (Wasserman & Faust, 1994, p. 122).

2.1.3.3 Valued Relations.

A relationship between actors in some instances can be valued to describe the strength of the relationship. The measurement generally attempts to reflect the intensity of the relationship, as detected by proxies such as amount or frequencies of interactions (Wasserman & Faust, 1994, p. 140).

2.1.3.4 Multiple Relations.

If multiple relations are used to model a social network, multiple relationships among two actors could be present (Wasserman & Faust, 1994, p. 146). An example could be co-workers, denoting a formal role, who are also friends, representing an individual evaluation.

2.1.4 Data Representation.

The predominant data structure used to represent social network data is the sociomatrix, more typically referred to as an adjacency matrix in operations research. It is a square matrix where each row or column represents an actor within the network. The common convention maintains that the actors are in identical order for the rows and columns. If a relationship is present between actors i and j, the matrix element of the i^{th} row and j^{th} column, x_{ij}, of matrix X is set to one, as exemplified in Table II-2 and its associated social network graph displayed in Figure II-1. Additionally, if the relation is undirected, the matrix element of the j^{th} row and i^{th} column, x_{ji}, of matrix X is set to one, ensuring the resultant matrix is symmetric. If a relationship is not present, x_{ij} is set to

Table II-2 Notional Adjacency Matrix (Directed Graph)

Actor	A	B	C	D	E
A	-	1	0	1	1
B	1	-	1	0	1
C	0	0	-	0	1
D	1	1	0	-	0
E	0	0	1	1	-

zero and x_{ji} is set to zero for symmetric relations. Thus, for undirected graphs the adjacency matrix will be symmetric, and for directed graphs may be asymmetric (Wasserman & Faust, 1994, pp. 150-151). For valued relations, matrix element x_{ij}, describing the relationship between node i and node j, is set to the value of the relationship, commonly referred to as a weight reflecting the strength or intensity of the relationship, as opposed to being set to one. Multiple relations may involve multiple adjacency matrices, each one representing a single relation.

2.2 SNA Measures

Marsden (1990) highlights the importance of the purpose of utilizing measures to characterize the relationship between the measure and its foundational sociological or psychological underpinnings. Utilizing measures to provide a precise description of the social ties that compose a network requires a different level of accuracy than from using measures as indicators of differences between individuals within a network or between networks' structural properties.

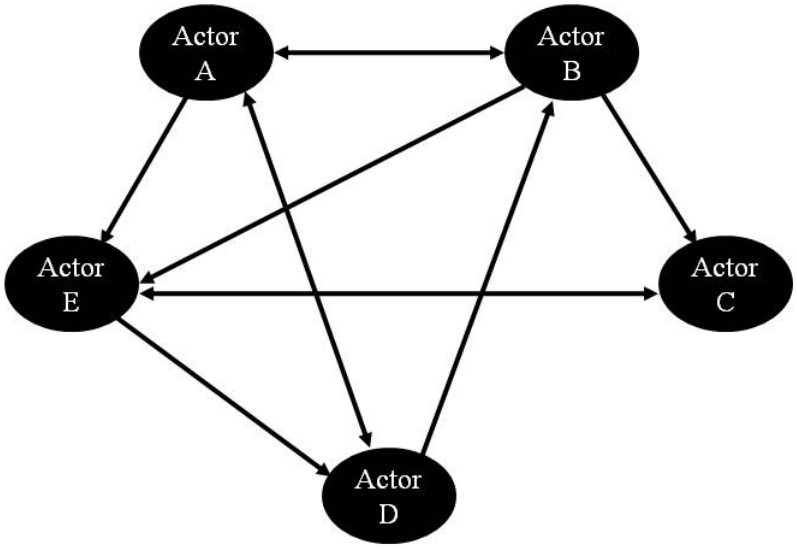

Figure II-1 Social Network Graph Based upon Table II-2

2.2.1 SNA Nodal Measures.

Many questions, in particular when examining dark networks, focus upon the relative importance of individual actors within the network. A host of measures derived from calculations based upon the network structure are available to characterize individual actor relative importance. Presented here is a subset of available SNA nodal measures, selected due to their common usage within SNA, specifically in literature addressing imperfect social network data.

2.2.1.1 Degree.

Nodal degree is the number of direct relationships with other actors possessed by an actor. It is simply the total number of edges incident to a node. For a network with undirected dichotomous edges, the degree of node v, $C_D(v)$, is the number of immediate neighbors node v possesses and can be computed by summing the corresponding row or

column of the adjacency matrix, A, as either will produce the same result as shown in Equation 1) (Wasserman & Faust, 1994, pp. 100, 163). The degree can be normalized, $C_D'(v)$, by dividing the degree of each node by the maximum possible value of $n - 1$, with n denoting the number of nodes in the network, as displayed in Equation (2.2) (Wasserman & Faust, 1994, pp. 178-179).

$$C_D(v) = \sum_i A_{iv} = \sum_j A_{vj} \tag{2.1}$$

$$C_D'(v) = \frac{C_D(v)}{n-1} \tag{2.2}$$

Degree centrality is a reflection of an actor's potential involvement in communication. Actors with high degree centrality are considered to be "in the thick of things" and "focal point[s] for communication" (Freeman, 1978/1979, pp. 219-220). However, when applied to dark networks, actors with high degree centrality may more accurately reflect "who you know most about, rather than who is central or pivotal in any structural sense (Sparrow, 1991, p. 256)."

*2.2.1.2 **In-Degree.***

Nodal in-degree is the number of relationships that are directed from other actors into an actor. It is the total number in incoming edges incident to a node. For a network with directed dichotomous edges, the in-degree of node v, $C_{ID}(v)$, is the number of neighbors with directed arcs to node v. It can be computed by summing the corresponding column of the adjacency matrix A (Wasserman & Faust, 1994, pp. 126-127).

$$C_{ID}(v) = \sum_{j} A_{vj} \tag{2.3}$$

2.2.1.3 Out-degree.

Nodal out-degree is the number of relationships that are directed from an actor to other actors. It is the total number in outgoing edges incident to a node. For a network with directed dichotomous edges, the out-degree of node v, $C_{OD}(v)$, is the number of directed arcs emanating from node v. It can be computed by summing the corresponding row of the adjacency matrix A (Wasserman & Faust, 1994, pp. 126-127).

$$C_{OD}(v) = \sum_{i} A_{iv} \tag{2.4}$$

In-degree and out-degree centralities reflect prestige among actors, measured by being the object of a number of ties, in a sense the amount of times other actors choose a particular actor. Dependent upon the specific relationship being modeled, being selected by other nodes may be an indication of power or influence over other actors. For example, if the relationship under consideration is a form of popularity, actors who are chosen more often theoretically have more influence over others, which would reflect mathematically in a high in-degree centrality. A hypothetical example reflecting this could be requests for co-authorship for academic publications, distinguished individuals may be sought out by others attempting to improve their status. If the relationship is a form of power, for example gives orders, actors exhibiting high out-degree centrality are considered influential (Wasserman & Faust, 1994, pp. 174-175).

2.2.1.4 Betweenness Centrality.

Betweenness centrality is a measure intended to identify actors who can control information flow in a network. It is based upon actors being located upon the shortest paths (geodesic) connecting other actors and this position allows them to exert interpersonal influence. As the actor is on the path between others in the network, they exhibit the potential to control occurring communication. They have the opportunity to prevent, delay, withhold, or alter information or materials passing through them. Betweenness centrality, $C_B(v)$, does implicitly assume flow occurs along the shortest paths between actors in the networks. It is computed by calculating the number of shortest paths existing between actors *j* and *k* that include distinct actor *v*, denoted as $g_{jk}(v)$, divided by the number of shortest paths existing between actors *j* and *k*, denoted as g_{jk}. The measure is sometimes standardized, $C_B'(v)$, by dividing $C_B(v)$ by the maximum achievable value possible for the center of a star graph (Freeman, 1977, pp. 35-38; Freeman, 1978/1979, p. 221; Wasserman & Faust, 1994, pp. 189-190).

$$C_B(v) = \sum_{j<k} g_{jk}(v)/g_{jk} \tag{2.5}$$

$$C_B'(v) = \frac{2C_B(v)}{(n-1)(n-2)} \tag{2.6}$$

Betweenness centrality identifies individual actors who lie on communication paths between other actors. As such, an actor acting as an intermediary on a communication path between two actors "exhibits a potential for control of their communication." Betweenness centrality reflects actors who coordinate group processes as a function of their role in maintaining communication between others. An actor with

high betweenness centrality could "influence the group by withholding or distorting information in transmission (Freeman, 1978/1979, p. 221)."

2.2.1.5 Closeness Centrality.

Closeness centrality measures an actor's centrality by examining their shortest path distance from all other actors in the network, resulting in a measure more associated with the center of a network from a graph theoretic perspective. Due to the requirement of a path existing between nodes, reachability, the measure is only appropriate for strongly connected graphs, which in practice precludes directed graphs. Similarly to betweenness centrality, closeness centrality assumes flow occurs along shortest paths. For an actor v, the length of the shortest path, $d(v,i)$, between actor v and all other actors i is summed and the inverse is computed. An actor's closeness centrality, $C_C(v)$, can be standardized, $C'_C(v)$, so the maximum value is one, by multiplying by the number of nodes in the network, n, minus one (Wasserman & Faust, 1994, pp. 184-185; Sabidussi, 1966, pp. 597, 602). Closeness centrality has also been extended to incorporate disconnected and/or directed graphs, by considering only reachable nodes from a given actor. Adjusted closeness centrality, $C_{CA}(v)$, for actor v, incorporates the number of actors reachable from v, R_v, and for nodes unreachable from v sets $d(v,i)$ to zero (Wasserman & Faust, 1994, pp. 200-201).

$$C_C(v) = \left[\sum_i d(v,i)\right]^{-1} \tag{2.7}$$

$$C'_C(v) = (n-1)C_C(v) \tag{2.8}$$

$$C_{CA}(v) = \frac{R_v/(n-1)}{\sum_i d(v,i)/R_v} \tag{2.9}$$

Closeness centrality is interpreted as the extent that an actor can avoid the potential control of other actors. Actors with high closeness centrality can seek information from throughout the network, and thus are not as dependent upon intermediaries for maintaining communication. Their central position is also an indication of their capability to propagate a message through the network with minimum cost or time (Freeman, 1978/1979, pp. 224-225).

2.2.1.6 Eigenvector Centrality.

Eigenvector centrality is based upon an actor's status as a function of the status of the actors with whom they possess direct or indirect relationships. Computationally, it is a weighted sum of direct and indirect associations across all paths, though it is sensitive to differences in degree among the actors (Bonacich, 2007, pp. 555, 564). An individual actor's status is computed as the results of a weighted linear combination of all actors' status' scores. For *n* actors in a network, this leads to a set of *n* equations and *n* unknowns, one equation and one unknown status score for each actor, though these equations are not guaranteed to possess a non-zero solution. Various methods have been constructed to reflect this social phenomenon and provide slight modifications to enable computations for actors' statuses evaluation, though generally not widely used due to the requirement of input parameters without appropriate establishing guidelines (Wasserman & Faust, 1994, pp. 205-210; Katz, 1953; Hubbell, 1965; Mizruchi, Mariolis, Schwartz, & Mintz, 1986; Bonacich, 1987; Bonacich & Lloyd, 2001). The predominant measure in practice, proposed by Bonacich (1972), establishes the actors' status scores by the

eigenvector, x, associated with the largest eigenvalue, λ, of a symmetric adjacency matrix, A, with values restricted between zero and one inclusive. These conditions ensure the largest eigenvalue will be positive and its associated eigenvector will be composed of nonnegative elements (Bonacich, 1972, pp. 113, 119).

$$x = \frac{1}{\lambda} Ax \tag{2.10}$$

2.2.1.7 ***Integration.***

Integration is a measure of how well connected an actor is within a network. It is conceptually similar to closeness centrality, though adapted for directed networks. Integration is a nodal measure based upon existing shortest paths between all other actors and the actor of interest. Since it is applicable to directed networks, a path may not exist. Computing integration, $I(v)$, for actor v, involves the reverse distance. The reverse distance, RD(i,v), involves calculating the length of the shortest path, $d(i,v)$ beginning with actor i and terminating at actor v. If a shortest path between actors i and v does not exist, $d(i,v)$ is set to zero. For directed networks it must be noted that the length of the geodesic starting at actor i and terminating at actor j may differ from the geodesic beginning at actor j and ending at actor i. To compute RD(i,v) as shown in Equation (2.11), the shortest path length is subtracted from the diameter of the network, d, plus one, with diameter defined as the longest existing shortest path in the graph between any two nodes. The reverse distances are summed and divided by the total number of nodes in the network, (n), minus one, (n – 1) as displayed in Equation (2.12). The measure can be normalized, Equation (2.13), to produce a relative score by dividing each actor v's

score by the longest shortest path terminating at actor *v* (Valente & Foreman, 1998, pp. 90-93).

$$RD(i,v) = d - d(i,v) + 1 \tag{2.11}$$

$$I(v) = \frac{\sum_{i \neq v} RD(i,v)}{n-1} \tag{2.12}$$

$$I'(v) = \frac{C_I(v)}{\max_{i,v}\,[d(i,v)]} \tag{2.13}$$

*2.2.1.8 **Radiality.***

Radiality is the obverse of integration. While integration is based upon incoming arcs to a node, radiality is dependent upon edges emanating from a node. In contrast to integration, radiality measures an actor's reachability into a network. Computing radiality, *R*(*v*), for actor *v*, again involves the reverse distance, though focuses on the paths emanating from node *v*. The reverse distance, *RD*(*v*,*i*), involves calculating the length of the shortest path, *d*(*v*,*i*) beginning with actor *v* and terminating at actor *i*. Similar to integration, if a shortest path between actors *v* and *i* does not exist, *d*(*v*,*i*) is set to zero. To compute *RD*(*v*,*i*), the shortest path length is subtracted from the diameter of the network, *d*, plus one, and is shown in Equation (2.14). The reverse distances are summed and divided by the total number of nodes in the network, *n*, minus one, as computed by Equation (2.15). The measure can be normalized, Equation (2.16), to produce a relative score by dividing each actor *v*'s score by the longest shortest path beginning at actor *v* (Valente & Foreman, 1998, pp. 90-93).

$$RD(v,i) = d - d(v,i) + 1 \tag{2.14}$$

$$R(v) = \frac{\sum_{i \neq v} RD(v,i)}{n-1} \tag{2.15}$$

$$R'(v) = \frac{C_I(v)}{\max_{v,i}\,[d(v,i)]} \tag{2.16}$$

Care must be taken when interpreting integration or radiality. Dependent upon the relationship being modeled the meanings to these two measures may reverse. For example, if the directed relationship in a social network is giving orders, possessing a high radiality score indicates power, perhaps representing a general in a military hierarchy, while a high integration score is not indicative of status. If the hypothetical directed relationship is gives information, a high integration score would indicate an individual with abilities to collect information from throughout the network, possibly increasing status if knowledge is power applies. In contrast, a high radiality score reflects an individual's capacity to rapidly disseminate information across the social network.

2.2.1.9 Clustering Coefficient.

The clustering coefficient for a given actor measures the number of connections among its neighbors, and is related to the transitivity concept in the social network literature. Transitivity is based upon the observance that "a friend of a friend is a friend (Wasserman & Faust, 1994, p. 150)." The clustering coefficient is a measurement of this social phenomenon of a person's friends also being friends of each other. It is a local measure as for a given actor it only needs its immediate neighbors and their interrelationships for calculation. It is defined as the proportion of interrelationships among neighbors compared against the potential links that could exist, which differ for undirected versus directed graphs. Given an undirected graph and a node v with k neighbors with m edges connecting neighbors of v, the clustering coefficient $C(v)$ of node

v is defined as in Equation (2.17). For directed graphs, the equation is multiplied by one-half, Equation (2.18) (Watts & Strogatz, 1998, p. 441).

$$C_{\text{undirected}}(v) = \frac{2m}{k(k-1)} \tag{2.17}$$

$$C_{\text{directed}}(v) = \frac{m}{k(k-1)} \tag{2.18}$$

2.2.2 SNA Network Measures.

One objective of SNA is to characterize a network, in terms of efficiency and effectiveness, or conduct a comparison to other networks. A host of measures derived from calculations based upon the network structure are available to characterize a network. Presented here is a subset of available SNA network measures, selected due to their usage within SNA, specifically in literature addressing imperfect social network data. Difficulties arise in the interpretation of these measures due to a lack of knowledge of what is an appropriate or optimal value in the context of the specific social network under investigation. Comparisons among various social networks with network measures are also suspect due to the impact of network size upon several of the calculations.

2.2.2.1 Density.

The density of a graph is the number of edges in the network, *m*, divided by the maximum possible number of edges (Wasserman & Faust, 1994, p. 101). If the graph is a directed network, the maximum number of possible edges is doubled (Wasserman & Faust, 1994, p. 129).

$$D_{\text{undirected}} = \frac{m}{n(n-1)/2} = \frac{2m}{n(n-1)} \tag{2.19}$$

$$D_{\text{directed}} = \frac{m}{n(n-1)} \tag{2.20}$$

2.2.2.2 Average Degree and Degree Distribution.

A commonly presented network measure is the average degree of a graph, simply defined the average of all the individual nodes' degrees (Watts, 1999, pp. 26-27). More descriptive than a simple mean, the degree distribution characterizes the variation among individual nodes' degrees. Significant to social networks, a power law degree distribution, or a closely related distribution such as a power law with cutoff, appears to be prevalent in empirical studies of social networks (Barabási & Bonabeau, 2003, pp. 63-64). A degree distribution follows a power law distribution if the probability of a given nodal degree, $p(x)$, is drawn from Equation (2.21), which is characterized by its exponent or scaling parameter α:

$$p(x) \propto x^{-\alpha} \tag{2.21}$$

Due to the tail behavior of a power law, empirically accurately estimating the distribution's parameters, the scaling parameter and the normalization constant, is difficult. Clauset *et al* (2009) introduced a statistical procedure to compute the parameters, calculate the goodness-of-fit, and compare against other potential distributions, though only their method to compute the parameters is discussed here. Their power law degree distribution, $p(x)$, is a function of the scaling parameter, α, and the minimum degree for which the power law is appropriate, x_{min}. Effectively, nodes with degrees below x_{min} are ignored in the computations estimating α. x_{min} is determined by iteratively investigating every potential x_{min}, estimating α, and comparing the corresponding model's degree distribution against the node degrees found in the

complete network of n nodes via the Kolmogorov-Smirnov statistic, a common statistical method of comparing distributions. The α and $x_{\min}$ are selected from the respective model that fits the data set best. Equations (2.22) and (2.23) summarize the approximations for the power law parameters (Clauset, Shalizi, & Newman, 2009).

$$p(x) = \frac{(\alpha - 1)x_{\min}^{\alpha-1}}{\sum_{i=0}^{\infty}(i + x_{\min})^{-\alpha}} \tag{2.22}$$

$$\hat{\alpha} \cong 1 + n\left[\sum_{i=1}^{n} \ln\frac{x_i}{x_{min} - .5}\right]^{-1} \tag{2.23}$$

2.2.2.3 <u>Degree Correlation (Assortativity).</u>

Degree correlation, also referred to as assortativity, is the Pearson correlation coefficient of nodal degrees (Newman, 2002). A network displaying assortative mixing, or positive correlation among nodal degrees, will have high-degree nodes connected directly to other high-degree nodes. The converse, a network with disassortative mixing, or negative correlation among nodal degrees, will possess high-degree nodes directly connected to low-degree nodes. Many social networks appear to possess assortative mixing, positive degree correlation (Newman & Park, 2003, pp. 036122-2), with high-degree nodes interconnected within a network core. Assortativity, r, is calculated via the following equation, where M is the total number of edges in the network, and j_i, k_i are the respective degrees of the vertices at the endpoints of the i^{th} arc (Newman, 2002):

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2} \tag{2.24}$$

2.2.2.4 Average Clustering Coefficient.

The average clustering coefficient, *C*, "measures the cliquishness" of a network, simply calculated by computing the average over all nodes' clustering coefficients given *n* nodes in the graph and ranges (Watts & Strogatz, 1998, p. 441). An average clustering coefficient for a network equal to zero implies that for all nodes in the graph, no neighbors of any node *v* is adjacent to any other neighbor of node *v* (Watts, 1999, p. 33).

$$C = \frac{\sum_v C(v)}{n} \tag{2.25}$$

Alternative clustering network measures exist, though they all maintain the same range of [0,1]. One alternative averages the clustering coefficient of only nodes with degree greater than one (Soffer & Vázquez, 2005).

$$C_d = \frac{\sum_{v|d(v)>1} C(v)}{\sum_{v|d(v)>1} 1} \tag{2.26}$$

Newman, Strogatz and Watts (2001) introduced an alternative definition of a network measure of clustering involving the number of triangles present on the graph compared against the number of connected triples of nodes. "'Triangles' are trios of vertices each of which is connected to both of the others, and 'connected triples' are trios in which at least one is connected to both the others (Newman, Strogatz, & Watts, 2001, pp. 026118-12)." The numerator is multiplied by three to account for triangles are composed of three connected triples of nodes.

$$C_t = \frac{3(\text{number of triangles on the graph})}{\text{number of connected triples of nodes}} \tag{2.27}$$

This definition of the clustering coefficient enables the comparison of a social network against a random model. Utilizing the average degree, $\bar{d}$, the average of the

squared nodal degrees, $\overline{d^2}$, and the size of the network, n, the clustering coefficient value for a network with no structure assumptions, C_n, can be computed. Real world social network data sets display higher clustering coefficients than their corresponding null structure configuration random model would suggest (Newman & Park, 2003, pp. 036122-3 : 036122-4).

$$C_n = \frac{\left(\overline{d^2} - \bar{d}^2\right)}{n\bar{d}^3} \tag{2.28}$$

An additional network measure based upon clustering coefficient involves a ratio of the number of neighbor interrelationships for each node compared against the sum of the possible edges for each set of neighbors in the graph. This can be interpreted as the three times the number of triangles present in the graph divided by the number of pairs of adjacent edges. Given an undirected graph, for each node v with $d(v)$ neighbors with m_v edges connecting neighbors of v, the ratio clustering coefficient is defined as follows (for directed graphs the equation is multiplied by ½) (Bollobás & Riordan, 2003, p. 18; Soffer & Vázquez, 2005):

$$C_r = \frac{\sum_v m_v}{\sum_v \binom{d(v)}{2}} \tag{2.29}$$

2.2.2.5 Fractional Size of Largest Component.

When modeled, social networks can form disconnected graphs, and often do in the case of dark networks. This phenomenon causes numerous difficulties in applying various SNA measures due to their assumptions of connected graphs. In practice, SNA measures are applied against each component, i.e. connected sub-graph, individually, or in some cases calculated for the largest component only. One network measure that

investigates this phenomenon is the fractional size of the largest component, simply defined as the number of nodes in the largest component divided by the total number of vertices in the graph.

2.2.2.6 Mean Path Length and Characteristic Path Length.

The mean path length is the average length of the shortest paths between all nodes contained within the largest component (Kossinets, 2006, p. 254). Existing within the literature is also the characteristic path length defined as the median of the means of the shortest path lengths between all nodes. Originally defined for undirected single component graphs, computing the characteristic path length involves calculating the average of the shortest paths for a given node to all other vertices. The characteristic path length is the median of this set of averages, i.e. one average per node (Watts, 1999, p. 29).

2.2.3 Bipartite Affiliation Networks in SNA.

Affiliation networks are two-mode networks describing the relationships existing between actors and events. Actors are connected only to the second mode, the events, and events are only connected to actors. This results in a bipartite graph with the subsets of actors and events with all arcs spanning between the two subsets and not within a subset. This bipartite graph can be projected into a one mode actor network by assuming actors connected to the same event possess, or are more likely to possess, a direct tie between them and other actors linked to the same event. Similarly, the bipartite graph can be projected into a one mode event network with the events interconnected through actor linkages (Wasserman & Faust, 1994, pp. 291-312). The projection of an affiliation

network, represented as a two mode bipartite network, into a single mode network appears in the literature as *unipartite projection* (Kossinets, 2006, pp. 250-251).

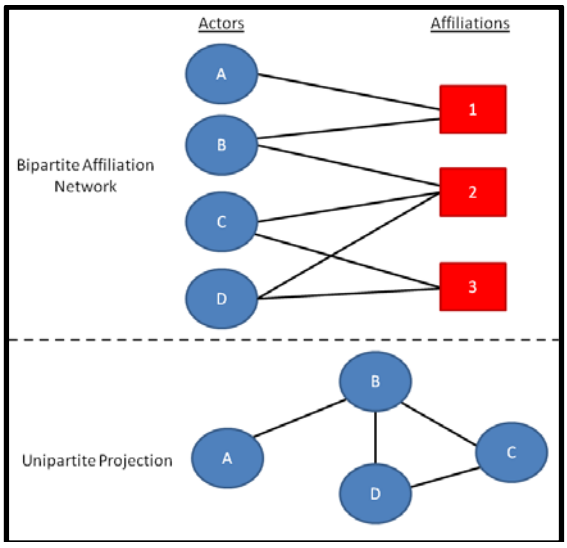

Figure II-2 Bipartite Affiliation Network and Associated Unipartite Projection (Adapted from Kossinets, 2006, p. 253)

2.2.4 SNA Bipartite Affiliation Network Measures.

2.2.4.1 Redundancy.

Kossinets (2006) introduced a measure to gauge the average importance of an affiliation within a bipartite graph, referred to as redundancy, β. Redundancy is calculated via the following equation where μ is defined as the average number of affiliations per actor, v is the average size of an affiliation, and z is the mean actor degree in the unipartite projection of the actors (Kossinets, 2006, p. 257).

$$\beta = \frac{\mu v - z}{\mu v} = 1 - \frac{z}{\mu v} \tag{2.30}$$

2.2.5 SNA Measures Overview.

Social network analysts utilize SNA measures discussed in Section 2.2, along with others, to detect changes in the social network's behaviors, changes in individual actors' behaviors, and to make decisions on how to affect dark network organizations. The SNA nodal measures can be used to determine which individuals or sets of actors to target for cooption, prosecution, message insemination, monitoring to gain more information on the organization and potential removal from the network. Some of these decisions constitute single opportunity events for execution or involve extensive resource commitments. As a result, fully understanding the impact of conducting these decisions using SNA measures that are conducting calculations on underlying imperfect social network data is paramount to adequately and appropriately assess risks and opportunities to the decision-maker.

2.3 Causes of Imperfect Social Network Data

Imperfect social network data stems from various sources. One source, boundary specification, is ever present as a social network is a model, an abstraction. As such, the modeler makes decisions regarding the inclusion and exclusion of specific data elements. In the particular case of social network analysis, a boundary specification problem arises. The modeler must decide which actors and which relations are to be included in the data set, in effect, determining the network's context. The challenge is further compounded when selecting the associated variables to collect on each actor and each relation. Difficulty in acquiring specific nodal or edge information, such as actors' demographic

data or the ability to measure relationships' intensities may drive boundary specification decisions.

Additional challenges and sources of imperfect data arise in the data collection portion of the analysis. An improper data collection design, inherent inaccuracies generated by the specific data providers, or a lack of information—which may be the intention of the subject network as in the case of dark networks—may introduce extraneous, spurious, or inaccurate data. These factors also potentially prevent the comprehensive collection of essential elements which can significantly impact the subsequent analysis and results. Inaccuracies in the collected data coupled with missing observations potentially lead to social network analysis being conducted in an environment of imperfect data.

2.3.1 Boundary Specification Problem.

The boundary specification problem involves the inclusion and exclusion of actors and the inclusion and exclusion of relations. The actors and corresponding relations that are included in the analysis—that define the network—are a subset of all existing actors and relations—from many potential networks. Rules are established to define an actor's inclusion into the network of study. Actors may be included or excluded from the social network based upon actor characteristics, their affiliations, or other specifications. Additionally, specific relation types are identified for inclusion in the network from the set of all relations. The appropriateness of the resultant reduction in actors or relations is dependent upon the analysis being conducted (Wasserman & Faust, 1994, pp. 30-39). Dark network actors often intermesh their professional and personal

lives, infusing difficulty in clearly delineating where illicit organizations and operations end and legitimate transactions and activities begin, in effect creating a fuzzy boundary which the social network analyst must arbitrate (Sparrow, 1991, p. 262).

Two different approaches are presented in Laumann, Marsden, and Prensky (1983) to address the boundary specification problem: realist and nominalist, though in application, a combination of the approaches may be employed for a particular study. Each boundary specification approach couples inclusion and exclusion rules that can be applied to collect data for modeling the target subject of interest (Laumann, Marsden, & Prensky, 1983, pp. 20-21).

The realist approach defines the boundary by assuming "that a social entity exists as a collectively shared subjective awareness of all, or at least most, of the actors who are members (Laumann, Marsden, & Prensky, 1983, p. 21)." This approach is somewhat circular argumentation in effect as the social network is defined by those who compose the social network. For formal organizations with clear membership this assumption is benign; however, when dealing with informal groups, such as collections of friends or criminal networks, this assumption creates a fluidity of the boundary of the social network. It has the potential to create a paradox where an individual actor may not consider themselves part of the social network, while members of the social network consider the actor as part of the collective. The obverse of the paradox could just as easily occur. From a modeling perspective, this inconsistency of the appropriateness of including the individual actor in the social network makes delineating the boundary difficult.

In the nominalist approach, the "analyst self-consciously imposes a conceptual framework constructed to serve his own analytic purposes (Laumann, Marsden, & Prensky, 1983, p. 21)." The social network is defined by arbitrary criteria that serve the analyst's lines of inquiry. In opposition to the realist approach, the social network's self-defined boundary is no longer an assumption, but an empirical question of how it compares against the analyst's defined boundary (Laumann, Marsden, & Prensky, 1983, p. 22). The arbitrary boundary selection by the analyst if applied inappropriately could significantly alter the social network analysis results. Conversely, if properly accomplished, this could distill the data requirements to essential elements required to satisfactorily analyze the question at hand, while concurrently eliminating extraneous data that could distort the results.

The data collected for a SNA study is generally either actors, relations, events, affiliations or a combination of the four. The inclusion and exclusion rules determine which elements of the four data types are incorporated into the social network model. Various inclusion and exclusion rules can be applied exclusively or in combination to determine which social data elements, specifically which actors, relations, events or affiliations, are incorporated into the social network model and subsequent analysis (Laumann, Marsden, & Prensky, 1983, p. 22).

2.3.1.1 Boundary Specification of Actors.

A network is partially defined by the actors to be represented as nodes. Laumann *et al* (1983) identify two types of actor boundary specification inclusion and exclusion rules, with the potential of generating rules combining the types. Positional rules test

actors' attributes for inclusion into the social network. The actor attribute could be fulfilling a specific position or role within an organization, hence the category name. The other type, a reputational rule, "utilizes the judgments of knowledgeable informants in delimiting participant actors (Laumann, Marsden, & Prensky, 1983, p. 23)." Hybrid inclusion and exclusion rules generated from elements of both types are sometimes found in practice (Laumann, Marsden, & Prensky, 1983, p. 23).

Applying these rule categories to real world problems generates a wide range of options to discriminate actors for potential incorporation into the social network model under investigation. Stemming from the three rule categories defined by Laumann *et al* (1983), an actor's inclusion and exclusion may be based upon membership with particular organizations, positional specification, demographic data or other actor attributes, involvement with specific relation types, event attendance, identification of inclusion by other actors, or a combination of these factors (Kossinets, 2008, p. 5). Though not a comprehensive categorization of actor inclusion and exclusion rules, a brief discussion of several rules follows.

The network of interest could be a formal organization in which actors are identified as members. If the organization's internal transactions are of interest, limiting the network to include only those who are members of the organization may be appropriate and enhance the accuracy, in terms of representation and interpretation, of computed social network analysis measures. Examples may include business corporations in which there may be a number of relations with suppliers and customers, but to accurately describe internal processes the social network may need to be limited to only employees of the organization (Marsden, 1990, p. 439).

Dependent upon the network of interest, positional specification may be used to define the actors within the network. Positional specification limits the actors in the network to those who occupy positions of rank in a formally constituted group. A military social network example may only include actors who are in command of a unit (Kossinets, 2008, p. 5).

Attributes of an actor could determine their appropriateness for inclusion into the network model. These attributes can include demographic data on the individual, such as gender, age, or rank. Utilizing actor attributes enables a reduction in extraneous nodes which may alter the analytic results by limiting the network to actors of significant value. A notional example may include investigating familial relationships and their impact on an organization. It may be prudent to remove children under a set age to minimize the impact of the relations on the social network analysis measures. A possible real world network where this is applicable may be beneficial is familial based organized crime, such as the mafia, or certain terrorist or insurgent organizations based primarily on familial connections.

Dependent upon the analytic goals of the social network study, the actors may be limited to those who only possess specific types of relations (Marsden, 1990, p. 439). If particular relation types are of specific interest, perhaps due to their impact upon the network, only actors who possess those types of relations may be included within the network. An epidemiological example could involve the tracking of a disease. Only actors with sexual relations may be pertinent in determining the social network in the case of a sexually transmitted disease.

In some real world networks there is not a clear delineation of actors belonging to specific organizations or groups. In these cases, it is possible that actor inclusion in the network is defined by those actors within the network. As this is based upon the individual perspectives of actors within the network, it is subject to biases. Kossinets (2008, p. 5) notes that, "Actors may disagree in their perception of social structure; they may be attributing different weights to certain other actors, relationship or types of relationships." Others' perceptions of an actor's activities determine whether the actor is considered as part of the network. A paradox can exist in which an actor believes he is part of a network, while the other actors do not include him as part of the network. Examples include collective movements on issues without political party affiliation or several party affiliations represented. Each individual within the network may construe a differing threshold of activity for inclusion with the network. Some may view voting in a specific manner, or donating funds and resources as justification for inclusion, while others may set the threshold higher as in actively protesting, and so forth. A real world example of this phenomenon is the environmental movement. There is no definition for someone being green, which could be interpreted as someone recycling to being a member of Greenpeace. Each potential member of the network can define the network inclusion criteria differently.

2.3.1.2 Boundary Specification of Relations.

Relational rules only allow actors possessing specific, defined relationship types into the social network model (Laumann, Marsden, & Prensky, 1983, p. 23). Relations for a given social network are chosen to encompass and represent specific types of actor

interaction. The exclusion of extraneous relationship types focuses the social network to represent a phenomenon of interest of the actors' interactions. The inclusion of relationships types that have no bearing on the analytical question at hand may negatively affect the accuracy of the resultant SNA measures. Likewise, ignoring relationships among actors of certain types that significantly describe and impact actors' behavior may negatively affect the accuracy of the SNA measures as well. Incorrectly bounding the relationships that are included in a social network generates extraneous links or, in effect, removes links that are present in the "true network" under investigation.

2.3.1.3 Boundary Specification of Events and Affiliations.

Actors and their associated relationships derived from participation in specific events or defined affiliations can provide a basis for inclusion and exclusion rules. An event or activity is specified by the analyst as being relevant to the social network. Only actors and the inter-relationships derived from participating in the event or activity are included in the social network model (Laumann, Marsden, & Prensky, 1983, p. 24).

A similar boundary specification can be extended to affiliations. In some instances, affiliation data between actors is generated by event attendance, but can also be derived through membership to multiple organizations and groups. All actors who attend a particular set of events are included as part of the network (Marsden, 1990, p. 439). Kossinets (2008, p. 5) warns that event attendance "is particular error-prone and is best described as convenience sampling." His reasoning is based upon the self selection of event attendance by the actors. Self selection could result from actors who attend a meeting, though numerous circumstances could affect those who did not attend and thus

preclude them as members of the network. Examples such as subordinates directed to meetings in place of their superiors, virtual participation via telecommuting, meeting conflicts inadvertently altering the membership, can substantially impact which actors are included in the social network model.

2.3.2 Data Collection Effects on Imperfect Data.

The technique used to acquire social network data may introduce sources of error and biases. The type of social network of interest may contribute or even compound these errors or biases. For example, data collection on bright networks is commonly accomplished through surveys and data freely provided by actors within the network. In contrast, with dark networks, the actors within the network may purposely inhibit data collection or encourage the collection of spurious data through a variety of methods and means.

2.3.2.1 Disambiguation of Actor Aliases.

One issue inhibiting the collection of accurate social network data on dark networks is disambiguation of actor aliases. Some actors within a dark network operate by using a series of aliases, which could take the form of alternate names, *noms de guerre*, redundant email accounts, multiple IP addresses, or several cell phones. Actors use their various identities when interacting with other actors to conceal the scope of their activities and provide a level of protection if a portion of the network is compromised. This effect can also be present in social network data not by actors' design or intentions, but as a failure to disambiguate actor information properly. Causes include trivial

mistakes such as typological errors, misspellings, poor transliteration, failure to reconcile nicknames, and so forth.

This phenomenon presents a challenge in the collection of social network data. Aliases generate false actors in the network analysis while masking the full spectrum of relationships of a single individual. With most SNA nodal measures, the removal of relationships to an actor diminishes the actor's computed importance. An actor spreading their relationships over several aliases can significantly diminish their appeared importance within the network. The challenge in SNA data collection when confronted with this issue is successful disambiguation and correct aggregation of aliased actors and their collective relationships into a proper single actor.

2.3.2.2 Respondent Inaccuracies.

As social network data can be reported by actors within or outside of the network, there is an inherent human error mechanism in accurately reporting network information. Research has shown that "people are incapable of reporting accurately on transactions that take place within highly specific time frames (Marsden, 1990, p. 447)." Respondents are biased towards reporting the routine, typical network structure. This is exemplified by event attendance, where actors have a tendency to attribute an actor's attendance to a specific event occurrence based upon the actor's attendance in general (Marsden, 1990, p. 447).

Several studies (Killworth & Bernard, 1976, 1979; Bernard & Killworth, 1977; Bernard, Killworth, & Sailer, 1979/1980, 1982; Romney & Faust, 1982; Bernard, Killworth, Kronenfeld, & Sailer, 1984; Romney & Weller, 1984) have investigated the

accuracy of informants providing information on social networks. Generally, the studies involve voluntary participants reporting their own social contacts within given time periods. The results from these investigations paint a foreboding picture in terms of the amount of potential error stemming from respondent inaccuracies for any given social network study.

The initial investigation into this phenomenon tersely summarized informants as "extremely inaccurate (Killworth & Bernard, 1976, p. 269)" and "people simply do not know, with any degree of accuracy, with whom they communicate (p. 283)." In this particular study, respondents were reporting those with whom they communicated and the relative rankings of their direct communication partners. This is the most simplistic case in terms of social network data collection, as an individual is only reporting their own direct communication. However, the results showed that individuals display relatively poor performance in accurately reporting their own communication patterns. Individuals did not even characterize more than one-third of their most frequent contacts (Killworth & Bernard, 1976, p. 280). In their initial study, Killworth and Bernard (1976, p. 281) stated "people only seem capable of ranking their most frequent communicator with any accuracy—and then only half the time!" Responding to critiques about their methodology, additional studies were conducted utilizing more expansive data sets and improved data collection means designed to improve the rigor. These refined studies (Bernard & Killworth, 1977) further solidified their initial assertions; "at best, people can recall or predict (on average) less than half their communication (either amount or frequency (p. 10)." Individuals' inability to accurately report their own communication

patterns casts aspersion upon an informant's ability to correctly disclose the inner workings of a real world social network, particularly dark networks.

The only bright spots in terms of informant accuracy identified in Bernard's, Killworth's, and Sailer's extensive work are the general trend of "errors of omission are more severe than those of commission (Bernard, Killworth, & Sailer, 1982, p. 53)" and "although individual people did know with whom they communicated, people *en masse* seemed to know certain broad facts about the communication pattern (Bernard, Killworth, & Sailer, 1982, p. 62)." The first quote indicates that social network data sets, in this case they were examining cooperative bright or open society networks, suffer more from missing data as opposed to spurious data. The second observation provides optimism that although actors may not identify their local structure accurately, they may possess insights into the global structure of the social network. As several social network measures are intended to quantify social interactions based upon prestige, influence, social status, and communication activity, informants' characterizations of relationships existing within the entire social network may prove to be accurate reflections of the network's structure. Informant reporting of the entire social network structure may generate data of sufficient quality and quantity that individuals and subgroups identified from SNA measures are of utility and appropriately characterize the real world social network.

Killworth's and Bernard's (1976) initial study utilized teletype communication logs among a deaf community in the Washington, D.C. area. Members of the network were asked to rank order those with whom they communicated. The subjects' responses were compared against the teletype logs to assess informant accuracy. Their subsequent

studies (Bernard & Killworth, 1977; Killworth & Bernard, 1979/1980;) drew their conclusions from four data sets: another replication of the deaf community experiment, amateur radio operator communications, observed interactions among office personnel, and observed verbal communications among faculty, graduate students, and staff in a graduate program at a university. The deaf community experiment replication again had teletype logs to compare network members' recall of communications against. The amateur radio operators' communications were monitored to generate the accurate network against which the informants' recall could be measured. Bernard, Killworth, and Sailer (1979/1980) replaced the deaf community data set with observed interactions of a college fraternity. Their methodology consistently involved comparing the social network data provided by participant sources, i.e. informant reported, against the empirically collected network representation, obtained through technical collection, such as the teletype logs or the monitoring of radio communications, or by observed interactions.

Unfortunately, somewhat contradictory findings to those of Bernard, Killworth and Sailer are also present in the literature. Romney and Faust (1982, p. 300) stated their ability to "capture the structure of a communications network from recalled data." Using one of Bernard's, Killworth's and Sailer's data sets, Romney and Faust demonstrated that structure could be extracted, though they truncated and normalized the original data. The structure they detected was one of interactions, and displayed that "the more similar two people judge the communication patterns of others, the more they interact with each other (Romney & Faust, 1982, pp. 297-300)."

adams and Moody (2007) investigated respondent concordance on sexual, drug, and social relationships. Individuals were interviewed up to five times over various time periods and questioned on relationships they were involved with and those of other actors in the network. Their results indicated that "respondents appear reliable for both who they are connected to and when they were connected (adams & Moody, 2007, p. 55)." However, respondents "are not as reliable reporting non-contact ties [relationships in which they are not a participant] as they are for their own ties (adams & Moody, 2007, p. 55)." Furthermore, there is a bias as "respondents are assuming that since they know two people, those two people must in turn know each other (adams & Moody, 2007, p. 56)."

2.3.2.3 Fixed Choice Effect.

Surveys are a common approach to collecting social network data on bright networks. One limitation that occasionally appears in surveys is termed the fixed choice effect, also referred to in the literature as right-censoring by vertex degree (Kossinets, 2006, pp. 252-253). Survey respondents are only permitted to nominate up to a certain number of actors with whom they possess relationships. The fixed choice effect generates imperfect data errors dependent upon the number of permitted choices in comparison with the underlying true structure. If the true structure of the social network for an individual possesses fewer connections than the number of choices, participants may feel compelled to nominate other choices, disregarding whether they accurately reflect reality. If the true structure contains more connections than permitted by the number of choices, the resulting data collection produces a truncated social network model (Holland & Leinhardt, 1973, pp. 88, 90).

2.3.2.4 Non-Responsiveness and Non-Detection.

The actors and relations define the social network, and collecting all of the data may not be feasible, so the resultant network for analysis may only be a sample of the true underlying network. This missing data can be absent due to a variety of mechanisms and thus may range from random to a stochastic process possibly correlated with some of the network's parameters. In bright networks, actors may choose not to respond to the collection method, such as a survey. For dark networks, specific actors or relationships may not be susceptible to the data collection means due to dealing with non-cooperative entities. Dark networks might exhibit a correlated missing data stochastic process as the actors within the social network conduct activities to purposely promote non-detection of actors within the network and the relationships that exist among them. Non-detection, in effect, produces a similar effect as non-responsiveness. Correlation of missing data to network parameters could potentially stem from important nodes within a dark network intentionally attempting to appear as non-important, such as hiding relationships with other important actors. The impact and effectiveness of these activities may vary significantly dependent upon the SNA measure in use, the structure of the network, and other factors.

Dark network activities that frustrate efforts to obtain social network data are analogous to the Department of Defense's concept of operations security (OPSEC) that promotes the protection of unclassified information that may be advantageous to an adversary. OPSEC is defined as:

> a process which includes identifying actions that can be observed by adversary intelligence systems; determine indicators that adversary intelligence systems might obtain that could be interpreted or pieced

> together to derive critical information in time to be useful to adversaries; and select and execute measures that eliminate or reduce to an acceptable level the vulnerabilities of friendly actions to adversary exploitation (Department of Defense, 2006, pp. GL-4).

Non-responsiveness of members of a social network has been investigated in the literature. Generally, these studies focused on response rates to social network data collection surveys in which participants voluntarily identify the relationships they possess. Identifying, evaluating, and developing survey strategies which increase the response rate are the main thrust of the academic research in this area. Its direct applicability to investigating dark networks is questionable as members of a dark network sometimes take active measures to preclude their identification and position in the social structure. However, several insights derived from the impact of non-responsiveness upon Social Network Analysis may prove beneficial to the investigation of dark networks.

Stork and Richards (1992) categorized social network information obtained in the presence of non-responsive participants. For a given complete network of n actors and a response percentage rate of R, three categories of information emerge: complete information, partial information, and no information. Complete information is the category of which both members of the reported relationship were collected, i.e. both responded to the survey, and reflects the upper left quadrant in Figure II-3. For an undirected relationship, both members confirmed either the existence of the relationship or the null relationship. For directed relationships, if collected properly, the existence of the potential relationships directed from actor A to actor B, actor B to actor A, reciprocation, or null relationships is in effect confirmed by both parties. For example, if the relationship is advises, simply querying an actor who they advise delivers one set of

the relationships. However, by additionally querying from whom the actor receives advice delivers the remainder of the relationships. Proceeding in this manner enables confirmatory nominations from both parties on the existence or absence of directed relationships.

In some instances, the potential exists for a responsive participant in the social network to provide information on their relationships involving a non-responsive participant. This single nomination case does provide partial information on the sociomatrix, as depicted in the upper right and lower left quadrants in Figure II-3. Single nomination identifies potential candidate actors for inclusion into the social network model, though the appropriateness of incorporating them must be assessed as the candidate in question did not confirm the relationships drawing them into the network. For directed relationships, the participating respondents must be queried on both directions of potential relationships to fully encompass all potential non-respondent network members. Inquiring upon a responsive participant's outgoing directed relationships, as well as the incoming directed relationships, identifies the maximum obtainable number of non-respondents in the social network.

When examining dark networks, there will probably exist network members whose social relationships will not be captured by the data collection technique. Additionally, these non-respondents may interact among themselves; however, these relationships will be opaque to the social network analyst. This phenomenon is represented in the lower right quadrant in Figure II-3, where no social network information exists.

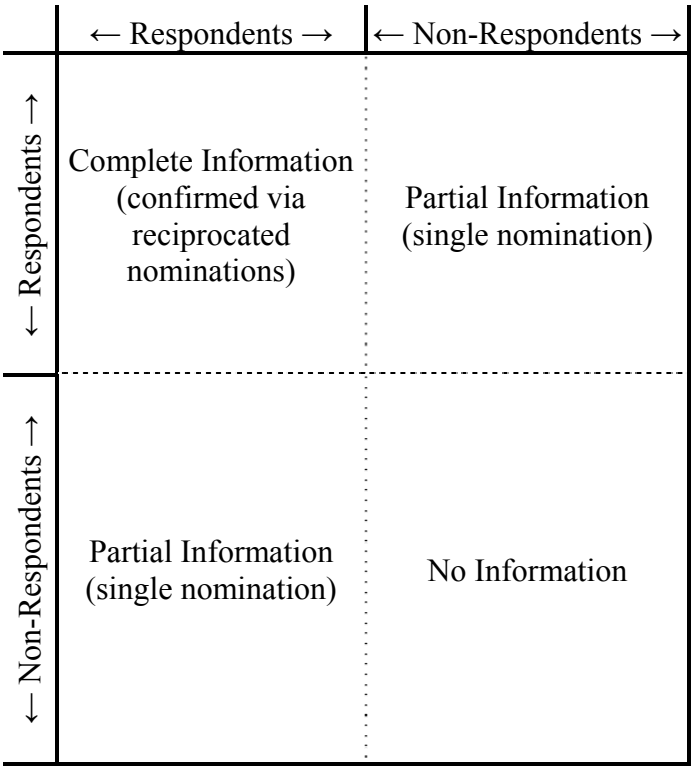

Figure II-3 Notional Sociomatrix with Responsiveness Rate

Sorting actors by their response status in a sociomatrix highlights the categories of information as displayed in Figure II-3. Modeling options as identified by Little and Rubin (1989) are composed of: complete-case analysis, available-case analysis, and imputation and are overlaid in Figure II-4. Complete-case analysis only utilizes relationships obtained from confirmed nominations with both actors reporting the existence or absence of the relationship, with the basis of this decision grounded by the assumption "reciprocated reports are substantially more likely to match observed interactions than are unreciprocated reports (Marsden, 1990, p. 447)." Available-case analysis extends the data set to include relationships that are identified by only one member of the dyad, also referenced as reconstruction (Stork & Richards, 1992, p. 197).

The data for available-case analysis can be extended via imputation. "Imputation replaces missing values by suitable estimates and then applies standard complete-data methods to the filled-in data (Little & Rubin, 1989/1990, p. 294)." However when applied, imputation may introduce imperfect social network data into the model that could significantly affect the SNA results.

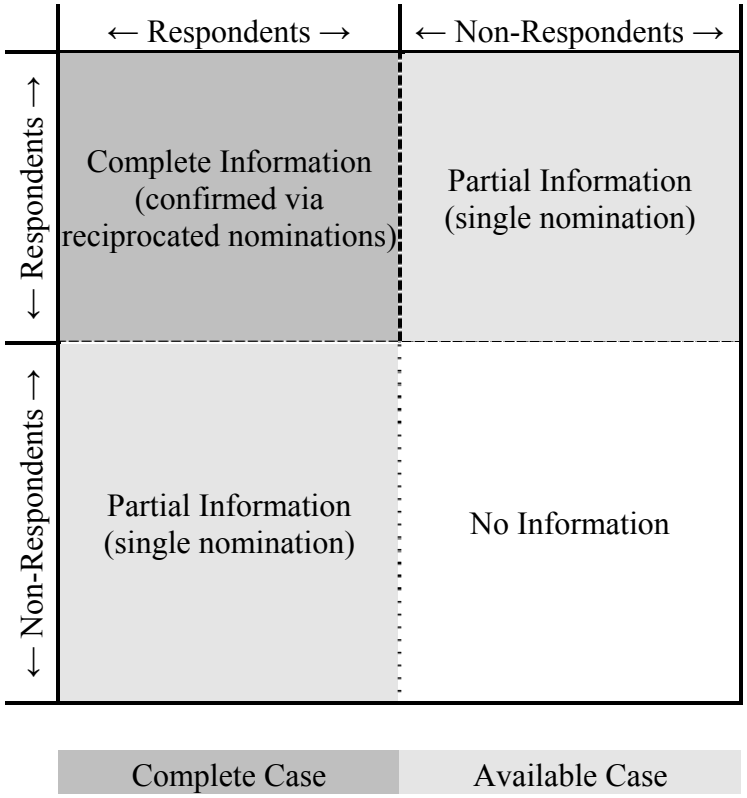

Figure II-4 Information Cases with Responsiveness Rate

Figure II-5 displays the number of directed relationships for each quadrant of the sociomatrix for notional values of the response rate *R*, and the number of nodes in the complete network, *n*. Of note, the objective of a data collection activity could be either to

maximize the amount of "complete information" to support complete-analysis or to minimize the amount of "no information" to support available-case analysis and imputation. Either strategy has substantial implications upon the resulting social network analysis.

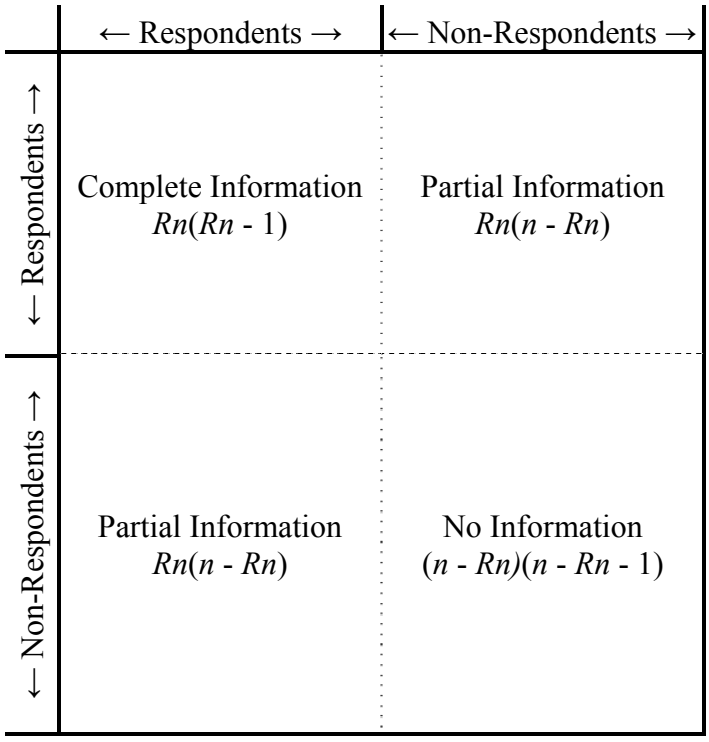

Figure II-5 Number of Relationships in Each Information Category

2.3.2.5 <u>Snowballing Data Collection Technique.</u>

Due to the difficulty of constructing social networks, while somewhat alleviated due to electronic social phenomenon (such as online social networking sites, texting, and so forth), several data collection techniques exist. One prominent technique in the literature is snowball sampling. The process involves beginning with a seed actor or

actors, collecting a specified number of their contacts and adding those actors to the network. A specified number of the contacts of the recently added actors are collected and those actors are added to the network. Additional contacts are added to the social network model in an iterative fashion for the desired number of repetitions. Thus, there are two parameters of the algorithm: the number of contacts identified by an individual actor and the number of iterations of the algorithm (Goodman, 1961).

There are numerous variations to snowball sampling: increasing the number of initial seed actors to begin the process, only investigating a subset of an actor's contacts to increase the network either from a random process or driven by local evaluations of SNA nodal measures (Tsvetovat & Carley, 2007; Illenberger, Flotterod, & Nagel, 2008; Bonneau, Anderson, Anderson, & Stajano, 2009; Zilli, Grippa, Gloor, & Laubacher, 2006). Clearly, snowball sampling may bias observed network structures.

2.4 Modeling Imperfect Data in Social Network Models

Currently several techniques exist to simulate imperfect data in social networks to assess the impact on Social Network Analysis measures. The techniques involve the addition of extraneous nodes and arcs to mimic actors and relationships that should not be present in the social network and the removal of nodes and arcs to mimic missing actors and relationships. The generated graphs from these techniques are compared against the original network, or "true" network, to evaluate the impact of the imperfect data upon some specified network or nodal measure. The conclusions of these assessments are generally ascribed as evaluating the robustness of the measure in the presence of imperfect data.

2.4.1 Statistical Missing Data Mechanisms.

There are three missing data mechanisms in the statistics literature: missing completely at random, missing at random, and not missing at random. Missing completely at random defines the case where the missing data mechanism, *M*, does not depend on the values of the data, *Y*, though there may be a pattern present to the missing data due to unknown parameters, ϕ (Little & Rubin, 2002, p. 12).

$$f(M|Y,\phi) = f(M|\phi) \text{ for all } Y,\phi \tag{2.31}$$

Missing at random defines the less restrictive situation where the missing data mechanism, *M*, depends on the data that is observed, Y_{obs}, and not on the data that is missing, Y_{mis} (Little & Rubin, 2002, p. 12).

$$f(M|Y,\phi) = f(M|Y_{obs}\phi) \text{ for all } Y_{mis},\phi \tag{2.32}$$

The least restrictive mechanism is not missing at random, where the missing data can be a function of the values of the missing and/or observed data. Not missing at random is the most difficult to deal with analytically (Little & Rubin, 2002, pp. 12, 17).

An example of missing completely at random in a social network application could be non-detection of a communication between two actors. Collecting data through the snowball technique may exemplify a data missing at random mechanism. An actor purposely taking action to minimize their appearance of importance within a network is an example of not missing at random. All three of the missing data mechanisms may be present when confronting social network data, particularly when applied against dark networks.

2.4.2 Comparing Networks.

A common approach (Costenbader & Valente, 2003; Borgatti, Carley, & Krackhardt, 2006; Kossinets, 2006; Kim & Jeong, 2007) for assessing imperfect data effects upon Social Network Analysis and its associated measures, generally proceeds as follows: given an original network, set it as the baseline, and then conduct comparisons against other networks, usually generated from altering, augmenting, or reducing the original network. These comparisons traditionally involve computing SNA measures for the original network and observing the same measure values recomputed on the generated networks. These comparisons attempt to assess the impact of imperfect data on determining the "true" network structure, measured via differences in the SNA network measures, and the effects of imperfect data upon "true" nodal prominence and positioning within the network, measured via changes in the SNA nodal measures.

2.4.2.1 Comparing Nodal Measures.

The effects of imperfect data on nodal measures are measured using several methods. Generally, the nodal measure under study is computed for the original complete network and then calculated for each sample of the network remaining after employing a node removal procedure. Several difficulties arise from this approach. The sampled networks are of smaller size, in terms of number of nodes, than the original network. This presents the situation of nodes possessing a node measure value in the original graph and being subsequently absent from the sampled network, potentially instilling a statistical bias into the resultant analysis dependent upon the specific node removal procedure utilized. This absence prohibits direct comparison of changes in

nodal measure values for all nodes. Another issue that arises is the inability to compute some nodal measures for all of the nodes in the sampled graph. The node removal process can create a sample network which is fractured into multiple components. Several SNA measures, such as closeness centrality or betweenness centrality, require a single connected component for calculation of the measure. The node removal procedure can separate the network into several disjoint networks prohibiting the computation, or altering the calculated values and/or interpretation, for some SNA nodal measures.

Table II-3 and Table II-4 summarize nodal measure comparison techniques found in the literature. With replications, the variance and standard deviation of the value derived from the comparison measure in use is often calculated within the study to assist in characterizing the robustness.

2.4.2.2 Comparing Network Measures.

Network measures are naturally suited for comparisons among various graphs. Since they are generally designed to characterize a network, a comparison of a generated graph against an original version from which it was created is ascribed as accurately assessing the effect of the generation mechanism upon the network measures. However, some of the network measures, such as characteristic path length, may prove to be incalculable if the generation mechanism creates multiple components. A general procedure employed in the literature if this phenomenon is encountered is to calculate the network measures only on the largest connected component in the graph (Kossinets, 2006, p. 254). Computing network measures utilizing only the largest component may

mischaracterize the effect of the generation procedure used to create the sample networks from the original network.

Table II-3 Statistical Based Nodal Measures Comparison Techniques

Comparison Technique	Description (Studies where Applied)
Pearson Correlation Coefficient	Correlation coefficient of the nodal measure values. Calculated for nodes occurring in both true and observed networks. (Costenbader & Valente, 2003, p. 290)
Square of the Pearson Correlation Coefficient (R^2)	Square of the correlation coefficient of the nodal measure values. Calculated for nodes occurring in both true and observed networks. Interpreted as the proportion of variance of true measure values accounted for by the observed network. (Borgatti, Carley, & Krackhardt, 2006, p. 127)
Kendall's Tau (τ)	Ordered similarity of ranked values. Isolated nodes are excluded. Probability p that an arbitrary pair is ordered similarly: $p = (\tau + 1)/2$ (Kim & Jeong, 2007, p. 110)
Probability Distribution Similarity (ρ)	Pearson correlation coefficient, ρ, between k_i and k_i^o for $i = 1,...,N$ for N sampled nodes, where k_i and k_i^o are the measure k 's i^{th} node value for the sampled and original network respectively, such that k_i^o satisfies $P_S(k_i) = P_O(k_i^o)$ where P_S and P_O are the cumulative distribution functions of the sampled and original networks respectively. Normalized to [0,1] by $(\rho - \rho_{th})/(1 - \rho_{th})$ where ρ_{th} is the Pearson correlation coefficient between k_i and k_i^o as if $P_S(k)$ is a linear function of k . Isolated nodes are excluded. (Kim & Jeong, 2007, p. 110)
Differences	Difference between the nodal measure values. Calculated for nodes occurring in both true and observed networks. (Costenbader & Valente, 2003, p. 290)

Table II-4 Proportion Based Nodal Measures Comparison Techniques

Comparison Technique	Description (Studies where Applied)
Top 1	Proportion of times that the most central node in the true network is the most central node in the observed network. (Borgatti, Carley, & Krackhardt, 2006, p. 127)
Top 3	Proportion of times that the most central node in the true network is among the top three most central nodes in the observed network. (Borgatti, Carley, & Krackhardt, 2006, p. 127)
Top 10%	Proportion of times that the most central node in the true network is among the top 10% most central nodes in the observed network. (Borgatti, Carley, & Krackhardt, 2006, p. 127)
Overlap	Number of nodes in both the top 10% of the true network and the top 10% of the observed network, divided by the number of nodes in either. (Borgatti, Carley, & Krackhardt, 2006, p. 127)

Table II-5 summarizes several comparison techniques currently employed in the literature to compare various network measures of a given original network against sample networks generated from the original through several network alteration techniques.

2.4.3 Sampling of Social Networks.

Sampling social networks to estimate network measures was introduced by Granovetter (1976) as a method to hurdle the requisite data collection efforts common to Social Network Analysis. His sampling method is dependent upon two parameters, the number of samples and the corresponding sample size (Granovetter, 1976, pp. 1290-1291). This introduction to the SNA field specifically provided a method to assess a network's density through sampling of the network, by examining only a subgraph, or multiple subgraphs, of the "true network", obviating the need to collect the entire network's structure. Granovetter's density estimation invoked a normality assumption of

Table II-5 Network Measures Comparison Techniques

Comparison Technique	Description (Studies where Applied)
Relative Error (ε)	$\varepsilon = \left\|\frac{q-q_0}{q_0}\right\|$, where q_0 is the value calculated from the true network and q is the value computed from the observed network. (Kossinets, 2006, p. 254)
Tolerable Fractional Amount of Missing Data	Maximum amount of missing data as a percentage of the total data where the relative error does not exceed a specified tolerance. (Kossinets, 2006, p. 254)
Kolmogorov-Smirnov *D*-statistic	$D = \max_x\{\|F'(x) - F(x)\|\}$, where x is over the range of the random variable, F and F' are two empirical cumulative distribution functions. (Leskovec & Faloutsos, 2006, pp. 2-3)
Normalized Hamming Distance ($D^{norm}_{\text{Hamming}}$)	The sum of difference between two graph structures with lower values equating to similarity. Normalized for networks of different sizes. $D^{norm}_{\text{Hamming}} = {D_{\text{Hamming}}}/{e}$, where e is the number of possible edges in a graph. (Tsvetovat & Carley, 2007, p. 68) [D_{Hamming}, or Hamming Distance, is the number of edges that must be altered to alter one graph into matching other (Hamming, 1950, p. 155).]

the subgraphs' density. Regardless of the veracity of this particular assumption, the methodology proposed by Granovetter is applicable to the imperfect data problem.

This section provides a survey of several social network sampling techniques employed within the literature. The general approach involves taking several samples of a given size and estimating characteristics of the population according to probability distribution assumptions. As illustrated by the variability and sometimes contradictory nature of the presented analytical studies' conclusions, the underlying assumptions

regarding distributions, network structures, and missing or imperfect data mechanisms are critical to accurate representation of imperfect data effects on SNA.

There are several methods present in the literature to generate sample graphs derived from an original network to represent imperfect data phenomenon. One such approach is to generate samples of the original network with data elements removed to represent missing data. The removal sampling approaches can be categorized into node removal and edge removal. An additional hybrid approach involving simultaneous removal of nodes and edges can be applied, but increases the difficulty of separating the confounding effects involved in the sample generation process. The process for the node or edge removal involves a mechanism to remove data, either nodes or edges, from the original network and then examine the subsequent sample of the original network, i.e., the newly generated network. Sampling an original network via a node or edge removal mechanism is predominantly ascribed as investigating the effects of the boundary specification problem for actors, the boundary specification problem for relations, and non-response.

2.4.3.1 Node Removal.

A common model to emulate the boundary specification problem for actors and non-response is the random removal of nodes within a network. Node removal as a missing data mechanism appears appropriate to modeling dark networks. Some members and participants in dark networks, actively attempt to obfuscate their memberships, roles, and/or connections to those outside of the dark network. Individuals hiding their identity or affiliation to a dark network are of reduced likelihood of inclusion in a social network

depiction of the network. Several studies (Costenbader & Valente, 2003; Borgatti, Carley, & Krackhardt, 2006; Kossinets, 2006; Kim & Jeong, 2007) have taken this approach to investigate the impact on nodal and network measures. These studies draw from numerous real world data sets and a wide range of graphs generated using various random network creation methods over a large parameter space. Generally, the node removal mechanism is conducted via deleting nodes using probabilities derived from a uniform distribution.

Several studies investigating imperfect data modeled by node removal via a uniform probability distribution have been conducted and presented in the literature (Costenbader & Valente, 2003; Borgatti, Carley, & Krackhardt, 2006; Kossinets, 2006; Kim & Jeong, 2007); the findings, however, are in some cases contradictory and highlight the difficulty of the imperfect data problem in Social Network Analysis. Differences in incorporated real world data sets, random network generation methods' appropriateness to modeling real world social networks, and implementation specifics of their respective node removal mechanisms may account for some of the discrepancies among the conclusions. Generally, nodes were selected for removal from an original social network according to a uniform probability distribution, i.e. each node is equally likely to be selected. Node removal, subject to a uniform probability distribution, implies a missing completely at random mechanism.

The applicability of results derived from analysis stemming from node removal utilizing a uniform probability distribution, in terms of implications on conducting Social Network Analysis on dark networks, is suspect. It is not unreasonable to assume that individuals participating in a dark network may take substantial actions to actively elude

detection by outside organizations. Individuals significantly invested in a dark network, whether due to familial or professional connections, resource control, or power derived from positions of authority in the network, are generally those that the Social Network Analyst is attempting to discover. The measures employed by significantly invested individuals are assumed to be more extensive than those less invested in the dark network. This implies a not missing at random data mechanism is more representative of the real world problem than a simplistic uniformly at random. None of the studies specifically address dark networks and the associated implications, and thus their conclusions' applicability to dark networks is an open question.

Borgatti, Carley, and Krackhardt (2006) investigated the impact of a node removal mechanism following a uniform probability distribution and documented its effects on the following four nodal centrality measures: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. They generated Erdös-Rényi random graphs for four different network sizes ranging from 10 to 100 nodes, varying eight density settings ranging from 1 to 90%. They removed either 1, 5, 10, 25, or 50% of the nodes in the original network, the Erdös-Rényi random graph, via a node removal mechanism operating uniformly at random. Their published results identified that the four investigated centrality measures "behave virtually identically (Borgatti, Carley, & Krackhardt, 2006, p. 128)" when in the presence of missing data. Table II-6 displays their generalizations for each of their nodal measure comparison techniques, each defined in Table II-3; unfortunately, the specific effects of random node removal are aggregated with the results of three other imperfect data modeling techniques: edge removal, node addition, and edge addition. This aggregation stemmed from their

surprising result of "different types of error had relatively similar effects on centrality robustness (Borgatti, Carley, & Krackhardt, 2006, p. 134)." However, the authors caveated this conclusion with the belief "that this result may be limited to random graphs (Borgatti, Carley, & Krackhardt, 2006, p. 134)." Additionally, linear regression was conducted to determine the effects of network size and density on the accuracy, in terms of each of the nodal measure comparison techniques. The regression coefficients indicated that "network size is only weakly related to accuracy (Borgatti, Carley, & Krackhardt, 2006, p. 133)" and density has virtually no effect on accuracy (Borgatti, Carley, & Krackhardt, 2006, pp. 128-134).

Table II-6 Imperfect Data Impact on Nodal Centrality Measures

Nodal Measure* Comparison Technique	**Imperfect Data Effect**
R^2	Decreases linearly
Top 1	Decreases (similar to exponential decay)
Top 3	Decreases linearly
Top 10%	Decreases linearly (least sensitive)
Overlap	Decreases (similar to exponential decay)
*Nodal measures: degree, closeness, betweenness, and eigenvector centralities	

(Borgatti, Carley, & Krackhardt, 2006, p. 128)

Costenbader and Valente (2003; 2004) provided an alternative approach to node removal by taking repeated random samples of the network in a boot-strapping method. They sampled the original network by drawing nodes by randomly selecting rows in the adjacency matrix, uniformly at random, varying the number of nodes ultimately selected

from 10 to 80% in increments of 10%. The selection of actors by rows in the adjacency matrix does introduce an out-degree bias for non-symmetric directed networks, but was purposely chosen by the authors based upon the assumption that researchers accept non-reciprocal nominations (Costenbader & Valente, 2003, p. 289). Since they utilized 59 real world networks resulting from 8 different studies, the network data suffered from non-response, which ranged from 51 to 100% survey response by the actors, which in effect, already incorporates a level of missing data. The impact of non-response upon SNA nodal measure correlation to the original network was assessed via multiple multivariate linear regressions with mixed results on statistical significance.

Examining real world data, Costenbader and Valente (2003; 2004) investigated node removal effects on the following nodal measures: in-degree, out-degree, symmetrized degree, betweenness and closeness centralities on both directed and non-directional (symmetrized) networks, eigenvector centrality, integration and radiality. The comparison between the nodal measures of the original and sampled graphs was conducted with Pearson's correlation coefficient, ignoring actors who did not appear in both networks. Table II-7 displays summarized results for the investigated centrality measures. A select few of the measures and their responses under node removal are displayed in Figure II-6, Figure II-7, and Figure II-8, where sampling fraction is defined as one minus the percentage of nodes removed, i.e. the percentage remaining of the original network in terms of number of nodes. Figure II-6 and Figure II-7 display results that are generally reconcilable with Borgatti, Carley, and Krackhardt (2006), while Figure II-8 displays trends that are not congruent with the other measures.

Table II-7 Missing Real world Data Impact on Nodal Centrality Measures

Nodal Centrality Measures	Imperfect Data Effect
In-degree	Decreases linearly and exhibited highest correlation of all measures.
Out-degree	Decreases linearly.
Degree (symmetrized)	Decreases linearly and has lower correlations than in-degree and out-degree with some examples of wave-like pattern.
Betweenness	Decreases linearly and faster than in-degree and out-degree centralities.
Betweenness (symmetrized)	Decreases linearly and has lower correlations than in-degree, out-degree and betweenness centralities.
Closeness	Decreases linearly with some examples of a wave-like pattern.
Closeness (symmetrized)	Decreases linearly with lower correlations than closeness
Eigenvector	Highly unstable with small changes in data, exhibits large swings in correlation.
Integration	Decreases linearly with correlation, nearly as high as in-degree.
Radiality	Decreases linearly with greatest variation of all investigated measures.

(Costenbader & Valente, 2003, pp. 290-299)

It is apparent that there are several large discrepancies between the conclusions drawn in Costenbader & Valente (2003; 2004) and Borgatti, Carley, & Krackhardt (2006). This dissonance is most apparent for eigenvector centrality, observing the radical behavior for eigenvector centrality correlation on real world data in Figure II-8 compared against Borgatti's, Carley's, and Krackhardt's conclusion of R^2 decreases linearly, implying that correlation decreases slower than linearly. Numerous factors could be contributing to this. As identified in their paper, Borgatti, Carley, & Krackhardt (2006, p. 125), real world data may contain systematic sampling errors whose patterns are

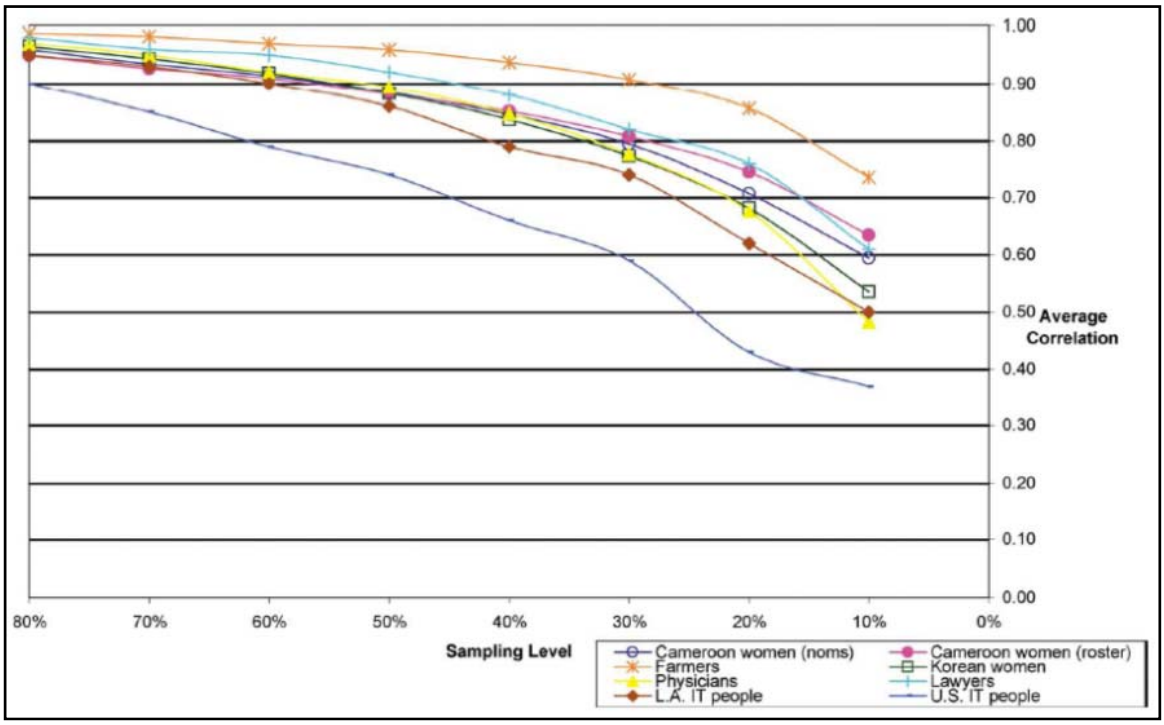

Figure II-6 Average Correlation for In-Degree on Real world Data (Costenbader & Valente, 2003, p. 293)

unknown. Leading to their design of experiments statistical approach, they randomly generated numerous graphs under varying parameter settings to conduct their analysis. The large divergence between their findings and those of Costenbader & Valente (2003; 2004) could be attributed to differences in the underlying structure of the networks. The generated networks were Erdös-Rényi random graphs, whose degree distributions, clustering coefficients and assortative mixings may have varied greatly from the real world data sets used by Costenbader and Valente (2003; 2004).

Instead of utilizing the raw SNA measure scores, Kim and Jeong (2007) conducted a comparison on the rank ordering of nodes within the largest component based upon the measures between the original and sampled graphs. They utilized three real world networks and randomly generated Barabási-Albert scale-free graphs to

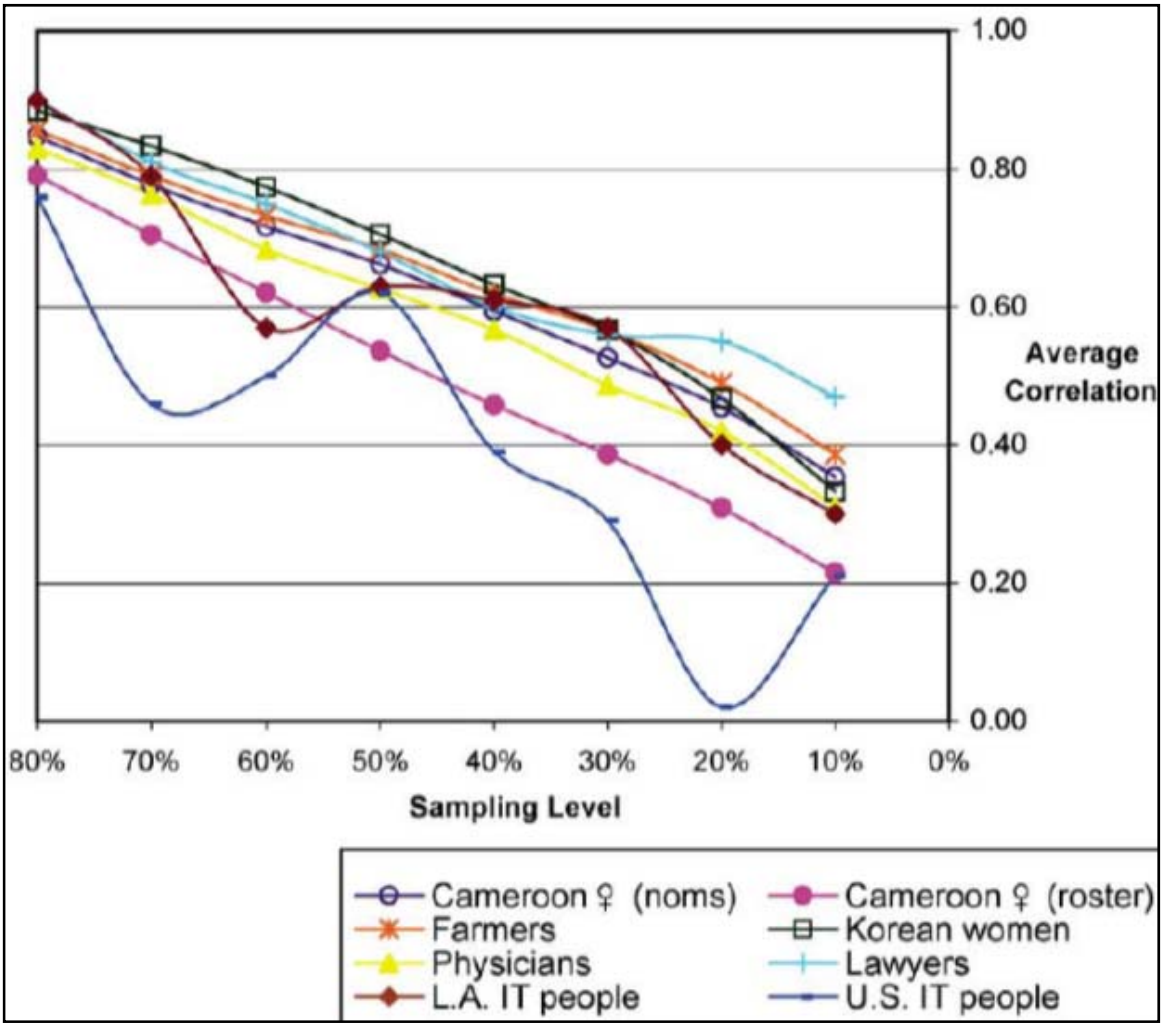

Figure II-7 Average Correlation for Closeness Centrality on Real world Data (Costenbader & Valente, 2003, p. 295)

investigate node-removal effects on SNA nodal centrality measures, specifically: degree, betweenness and closeness centralities. Kim's and Jeong's (2007) examination determined that random node removal had significant negative effects upon sampled network measure correlation with the original network using Kendall's Tau, τ, comparisons of actor rankings. Betweenness and closeness centralities performed significantly better, i.e. possessed a larger τ, than degree centrality on comparisons of

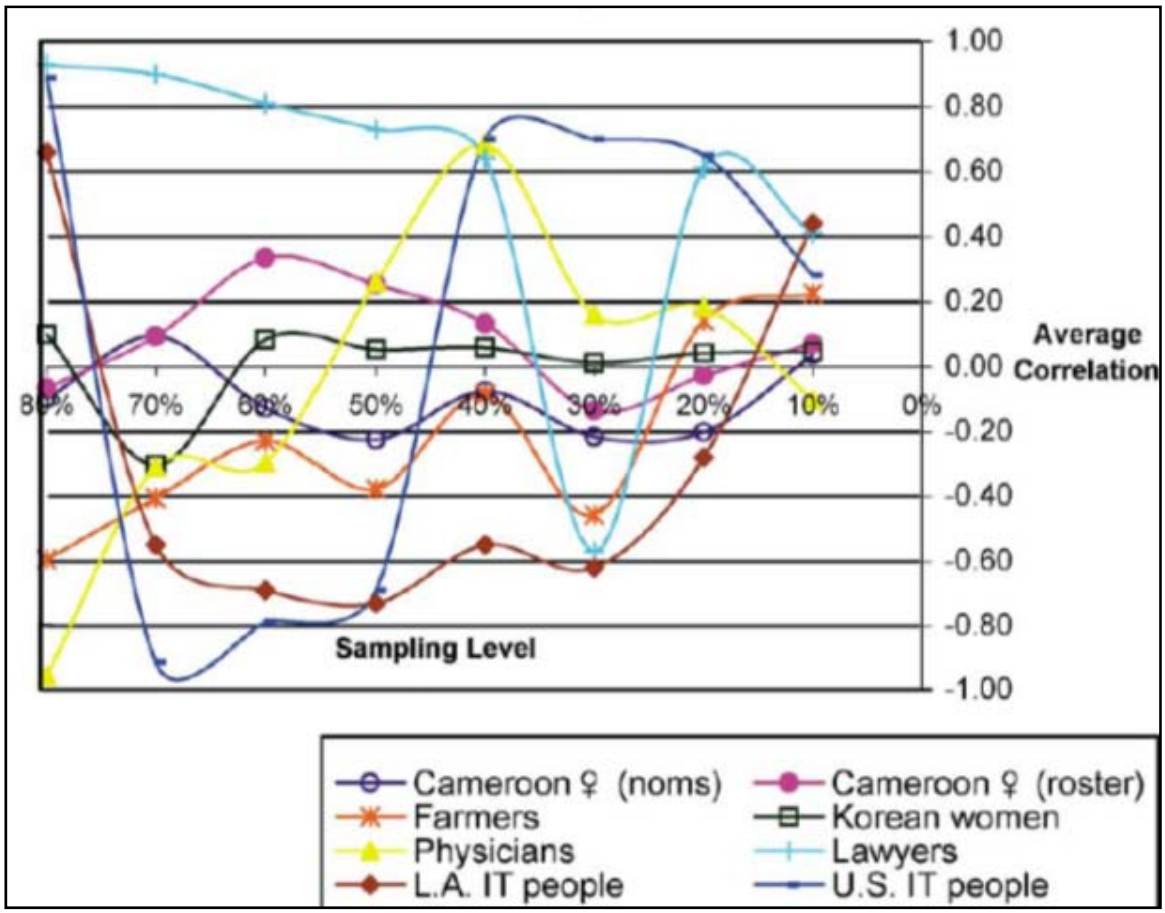

Figure II-8 Average Correlation for Eigenvector Centrality on Real world Data (Costenbader & Valente, 2003, p. 295)

actor rankings for the Barabási-Albert randomly generated graphs. On the real world networks, it was noted that "τ obtains its minimum value in an intermediate range of the sampling fraction." Figure II-9 displays the three measures' behaviors on real world networks as a function of the sampling fraction, which is one minus the percentage of removed nodes. This result contradicts those presented in Costenbader & Valente (2003; 2004) and Borgatti, Carley, & Krackhardt (2006), where, generally, they assessed a linear decline in correlation or accuracy.

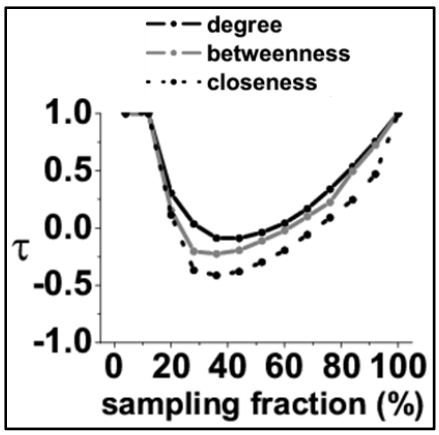

Figure II-9 τ Behavior on Real world Networks' with Node Removal (Kim & Jeong, 2007, p. 113)

By categorizing node rankings into deciles for a given measure and computing τ in a slightly different form for the Barabási-Albert randomly generated graphs, Kim and Jeong (2007) determined the contribution of each ranking interval of actors to τ. Thus, their results illustrated that higher ranking nodes, in terms of a given measure, are responsible for a larger τ. This implies that the highest ranking actors for a given measure in a sample provide a better representation of their associated rankings in the original network.

Kim's and Jeong's (2007) analysis of three real world scale-free data sets provided similar results with a few noted differences. They observed that betweenness centrality had significantly lower τ across the range of sampling fractions, percentage of nodes removed, for the real world networks than the randomly generated scale-free networks. They assessed this as a reflection of the modular structures located within the real world networks. They observed "the presence of nodes with small degree and large betweenness…indicates the existence of loose connections between tightly-knit

modules", referenced as the "modularity effect" (Kim & Jeong, 2007, p. 112). To explore this phenomenon, they conducted node removal by targeting actors with small correlation between degree and betweenness centrality scores as opposed to random sampling. This correction for modularity produced results for betweenness centrality's performance on the real world networks consistent with the results for randomly generated scale-free graphs.

For degree and betweenness centralities, Kim and Jeong (2007) also examined how their respective cumulative probability distributions, ρ, changed as the amount of missing data, removed nodes, is increased on Barabási-Albert randomly generated graphs. The cumulative probability distributions of degree and betweenness centralities appear to be resilient to random node removal as the sample networks maintain a similar distribution as the original network until withdrawal of a significant number of nodes, i.e. 80% of nodes removed.

Kim's and Jeong's (2007) noted superiority of closeness centrality, in terms of higher values of τ when confronted with uniform at random node removal, is assessed as resulting from the global characteristic of its calculation and its resultant insensitivity to the node removal mechanisms. However, closeness centrality's high values of τ are dependent upon the inclusion of hubs in the sample networks. Kim and Jeong (2007) conducted node removal by each centrality measure's score in descending order. By initially removing nodes with the largest value for a given measure, the τ rank comparison to the original network showed dramatic decreases. The authors noted that "sufficient sampling size of social individuals must be assured when access to the central leadership is restricted" (Kim & Jeong, 2007, p. 113). This may have significant

implications when dealing with dark networks where important actors, as defined by a given measure, may purposely hide their presence within the network through various methods.

Additionally, Kim and Jeong (2007) introduce a potential method of generating an approximate lower bound on τ in the case where a "true" network is unavailable. By conducting a node removal on the available network data and calculating τ, this serves as a lower bound on τ for the complete network assuming the same proportional amount of missing data, i.e. removed nodes.

Kossinets (2006; 2008) explored the impact of random node removal on SNA network measures using a large real world data set (n = 16,726, $m \approx$ 95,171) and 100 randomly generated bipartite graphs with similar network characteristics. The random node removal process incorporated deletion of an actor, and all of their associated edges, according to a uniform probability distribution. The SNA network measures investigated included: mean vertex degree, average clustering coefficient computed as defined by Newman *et al* (2001), assortativity calculated on the unipartite projections of bipartite graphs, the fractional size of the largest component, and the mean shortest path length in the largest component. In particular, the average clustering coefficient demonstrated very slow degradation in comparison with the original networks' values, displaying less than 10% error with 50% of nodes removed (Kossinets, 2006, p. 264). The summarized results on the sampled network measures performance in comparison with the original networks are included in Table II-8.

Table II-8 Random Node Removal Impact on Network Measures

Network Measure	Missing Data Effect	% missing data resulting in 10% relative error in measure	
		Real world Data	Random Graphs
Mean vertex degree	Decreases linearly	10%	10%
Average Clustering (Newman *et al* 2001)	No change	-	-
Assortativity (unipartite projection)	Small increase	-	15%
Fractional size of largest component	Decreases	8%	10%
Mean shortest path length (in largest component)	Increases*	30%	25%
*Until fragmentation phase change at ≈75% missing data			

(Kossinets, 2008, pp. 17-25)

2.4.3.2 Edge Removal.

Similarly to the node removal approach, emulating the boundary specification problem for relations and non-response is achieved via the random removal of edges within an original network. It has been noted that sampled graphs derived from a uniformly distributed edge removal process "will be very sparsely connected and will thus have large diameter and will not respect community structure" (Leskovec & Faloutsos, 2006, p. 3).

Borgatti, Carley, and Krackhardt (2006) conducted edge removal to investigate the impact on four nodal SNA measures: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Generating Erdös-Rényi random graphs for four different network sizes ranging from 10 to 100 nodes, with eight various densities ranging from 1 to 90%, they removed 1, 5, 10, 25, or 50% of the edges in the

original network, which results in Erdös-Rényi random graphs. The edge removal was conducted via deleting edges from probabilities derived from a uniform distribution. Their edge removal results are aggregated with other imperfect data modeling techniques and their generalized conclusions are presented in Table II-6.

Sterling (2004) investigated the impact of missing data in the form of an edge removal mechanism upon the ability to detect subgroups, i.e. community detection, using an algorithm developed in her thesis. The approach involved comparing the proportion of sampled networks producing the same 2-plexes as the original real world complete network data set (Sterling, 2004, p. 133); a 2-plex is a subgraph where all members possess ties with at least all minus two other members (Wasserman & Faust, 1994, p. 265).

Four different edge removal mechanisms were examined in Sterling's 2004 study: uniform at random, uniform at random of edges connecting nodes both of lower than average degree, uniform at random of edges connecting nodes both of higher than average degree, and uniform at random of edges connecting individuals who both belong in multiple subgroups. Each edge removal mechanism was intended to capture a phenomenon that could potentially occur with dark networks. The first method is the same uniform at random edge removal seen in other studies representing a simplified assumption of each relationship among actors is equally likely to be missing. The second method is attempting to replicate low level actors' relationships that are more likely to be missed due to the collection efforts ignoring the involved nodes because of assumed insignificance. The third method assumes important actors, such as hubs with their associated high degrees, are purposely trying to hide their significant positions in the

network. The fourth method models liaisons among different groups within the networks attempting to mitigate collection attempts (Sterling, 2004, p. 135). Sterling's results of each edge removal mechanism are summarized in Table II-9.

Table II-9 Various Edge Removal Mechanisms' Impact on Community Detection

Edge Removal Mechanism	% edges removed resulting in $\hat{p} = 0$
Uniform Random	10%
Between lower than average degree nodes	7.5%
Between higher than average degree nodes	12.5%
Between nodes in multiple subgroups	>50%

(Sterling, 2004, pp. 137-146)

Kossinets (2006; 2008) also simulated a non-response effect through an edge removal mechanism by randomly selecting a portion of the actors within the network and then removing all edges interconnecting this node subset of the graph. The subset of actors is included in the network if they possess other edges besides those that were eliminated. Of importance, relationships did not need to possess a reciprocal nomination to be included within the network. This is a substantial departure from the traditional uniform at random removal processes and purposely designed to replicate a non-response effect. Kossinets' empirical results of the non-response effect on network measures are summarized in Table II-10.

Table II-10 Non-Response Edge Removal Impact on Network Measures

Network Measure	Missing Data Effect	% missing data resulting in 10% relative error in measure	
		Real world data	Random graphs
Mean vertex degree	Decreases	30%	30%
Average Clustering (Newman *et al*, 2001)	Decreases	35%	35%
Assortativity (unipartite projection)	Decreases	35%	20%
Fractional size of largest component	Decreases	30%	50%
Mean shortest path length (in largest component)	Increases	50%	50%

(Kossinets, 2008, pp. 17-25)

Kossinets (2006; 2008) utilized affiliation networks and the unipartite projection to one mode actor networks to investigate the boundary specification problem for relations. Noting that an actor network is a collection of overlapping cliques, an event in a bipartite network represents an affiliation or interaction context which contains a clique. Since relations can be derived from multiple reinforcing contexts, each clique belongs to one or more events. Justification for this technique derives from the general practice in SNA of collapsing multiple relationships between actors into a single linkage to calculate SNA measures. By grouping relationships by context, i.e. relationships grouped to a single event, the removal of an event and thus its collection of relationships can simulate a relations' boundary specification problem, namely, the negative impact of excluding a significant relationship context.

Kossinets (2006), utilizing bipartite graphs and the unipartite projection, investigated the effects of missing edges on network level measures via the random

removal of the corresponding bipartite affiliation of the actors. This technique simulates a boundary specification problem for relations, due to its removal of a context upon which the relationships were formed. An affiliation is deleted, removing all of its edges adjoining it to actors, thus severing the relationships between those actors for that context. This contextual edge removal, as opposed to random edge removal in comparison, provides a different removal mechanism to better simulate a boundary specification for relations error. One hundred random bipartite graphs were constructed, each with the same number of actors, affiliation groups (articles), and edges corresponding to a single real world data set, the collaboration graph resulting from authors and articles contained within the Condensed Matter section of the Los Alamos E-print Archive for the years 1995 through 1999. Of note, the degree distribution of the randomly generated graphs possessed a Poisson degree distribution, while the exemplar real world data followed a power law degree distribution. The real world data set showed greater disassortativity on the bipartite graph and positive assortativity on the unipartite projection than the randomly generated graphs. Kossinets' evaluated the effect of contextual edge removal on several network measures and the results are summarized in Table II-11.

2.4.4 Inclusion of Incorrect Social Network Data.

The previous discussion and reference material in the literature exhibited the reduction of data in a network to assess the impacts upon various measures. As discussed in the description of the boundary specification problems for actors and relations, it is quite conceivable to collect extraneous social network data and include it in the analysis.

Borgatti, Carley, and Krackhardt (2006) explored the effects of the inclusion of extraneous data on SNA nodal centrality measures.

Table II-11 Contextual Edge Removal Impact on Network Measures

Network Measure	Missing Data Effect	% missing data resulting in 10% relative error in measure	
		Real world data	Random graphs
Mean vertex degree	Decreases linearly	14%	10%
Average Clustering (Newman *et al* 2001)	Increases	25%	10%
Assortativity (unipartite projection)	Increases	30%	10%
Fractional size of largest component	Decreases	15%	35%
Mean shortest path length (in largest component)	Increases*	40%	20%
*Until fragmentation phase change at ≈90% missing data			

(Kossinets, 2008, pp. 17-25)

2.4.4.1 Extraneous Node Addition.

Borgatti, Carley, and Krackhardt (2006) investigated the impact of node addition on four nodal SNA measures: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Generating Erdös-Rényi random graphs for four different network sizes ranging from 10 to 100 nodes, with eight various densities ranging from 1 to 90%, they added 1, 5, 10, 25, or 50% additional nodes to the original network, which was also an Erdös-Rényi random graph. The node addition was accomplished by setting the node's degree to the same as a node in the network selected

at random from a uniform distribution. With the new node's degree set, the node was incorporated into the network by randomly assigning its edges to nodes already present in the network. The node addition results are aggregated with other imperfect data modeling techniques and the general conclusions are presented in Table II-6.

2.4.4.2 Extraneous Edge Addition.

Similarly to their extraneous node addition investigation, Borgatti, Carley, and Krackhardt (2006) investigated the impact of edge addition on four nodal SNA measures: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Generating Erdös-Rényi random graphs for four different network sizes ranging from 10 to 100 nodes, with eight various densities ranging from 1 to 90%, they added 1, 5, 10, 25, or 50% additional edges to the original network, which was also an Erdös-Rényi random graph. The edge addition procedure was not explicitly defined in the published results, but it can be inferred that two nodes were selected at random from a uniform probability distribution and an edge was established between them unless one already exists. The edge addition results are aggregated with other imperfect data modeling techniques and the general conclusions are presented in Table II-6.

2.4.5 Snowballing Data Collection Effects on Imperfect Data.

Snowballing is a generally accepted data collection technique for social networks. However, the mechanisms involved create biases and significantly alter some SNA measures in the networks generated via snowballing. Studies have been conducted to investigate these biases and attempt to provide statistical corrections (Illenberger, Flotterod, & Nagel, 2008). Others have been conducted to modify the snowball data

collection algorithm to purposely skew the sample networks measures in a desirable manner (Zilli, Grippa, Gloor, & Laubacher, 2006; Tsvetovat & Carley, 2007).

2.4.5.1 Snowballing Data Collection Bias Corrections.

The snowballing data collection technique possesses some inherit biases which artificially skew SNA measures when they are applied to the collected networks. Due to the algorithm's selection of a seed node(s) and then proceeding to collect all actors linked to the seed node(s), high degree nodes are more likely to be selected early in the algorithm than low degree nodes. This also extends to a higher likelihood of early selection of nodes that are connected to high degree nodes (Illenberger, Flotterod, & Nagel, 2008, p. 4). These two biases have direct effects on SNA measures when the snowball technique is applied to scale-free graphs, with their right-skewed degree distributions and to networks with assortative, or disassortative, mixing. Some bias corrections for snowball sampling are displayed in Table II-12.

2.4.5.2 Snowballing Data Collection with Fixed Choice Effect.

One variant to snowballing data collection involves incorporating the fixed choice effect on the sampled nodes. When an actor is investigated to determine their links, only a fixed number are discovered. If relationships do not have to be reciprocated by both actors for inclusion into the network model, dependent upon the number of links returned under the fixed choice effect, some network measures can be computed with surprising accuracy. Bonneau, Anderson, Anderson, and Stajano (2009) utilized the fixed choice effect generated from the public view of individuals' Facebook accounts, a social network website, to approximate SNA measures of a subset of Facebook users.

Table II-12 Network Measures' Bias Corrections for Snowball Sampling

Network Measure	Bias Correction
Mean vertex degree $\langle k \rangle$	$\langle k \rangle^{(i)} \approx \frac{\sum_{v \in V^{(0 \ldots i)}} k_v / P_{k_v}^{(i)}}{\sum_{v \in V^{(0 \ldots i)}} 1 / P_{k_v}^{(i)}}$ where $V^{(0 \ldots i)} = V^{(0)} \cup \ldots \cup V^{(i)}$ is the set of all nodes sampled before or in iteration *i* which is of size $n^{(0 \ldots i)}$, k_v is the degree of sampled node *v*, and $P_k^{(i)} = 1 - \left(1 - \frac{(n^{i-1})}{N}\right)^k$ with *N* being the size of the total network
Degree Exponent (power-law degree distribution) γ	$\gamma^{(i)} = 1 + \frac{1}{\frac{1}{\sum_{v \in V'^{(0 \ldots i)}} 1 / P_{k_v}^{(i)}} \sum_{v \in V'^{(0 \ldots 1)}} \left(\ln \frac{k_v}{k_{\min}}\right) / P_{k_v}^{(i)}}$ where $V'^{(0 \ldots i)}$ is the set of all vertices sampled before or in iteration *i* that follow the power-law distribution which is of size $n'^{(0 \ldots i)}$ and $k_{\min}$ is the smallest degree for which the power-law holds
Clustering Coefficient C_v	$C^{(i)} = \frac{\sum_{v \in V^{(0 \ldots i)}} C_v / P_{k_v}^{(i)}}{\sum_{v \in V^{(0 \ldots i)}} 1 / P_{k_v}^{(i)}}$ where $C_v = \frac{2 e_v}{k_v (k_v - 1)}$ and e_v denotes the number of edges between *v*'s neighbors

(Illenberger, Flotterod, & Nagel, 2008, pp. 4-7)

Facebook's public view of an individual only displays eight friends. They approximated two nodal measures: degree and betweenness centralities; and several network properties: shortest paths, dominating sets and community detection through modularity. Their conclusions illustrated that despite given only a fixed number of random links from each actor, accurate approximations of some measures can be constructed, though accuracy is dependent on the number of links identified for each node.

2.4.5.3 Snowballing Data Collection with Targeted Search.

Due to the significant application of SNA in determining important individuals within the network, the snowballing data collection technique presents itself as a very intensive effort to fully explore the complete network. Several studies have been conducted investigating a modification to snowballing. Instead of investigating all

recently included actors at each iteration, only a subset of actors is explored to discover their relationships. This subset is chosen via SNA measures in an attempt to determine the "best" actor to investigate (Zilli, Grippa, Gloor, & Laubacher, 2006; Tsvetovat & Carley, 2007). This is analogous to a heuristic search. Both studies selected a single actor to investigate at each iteration. In Tsvetovat's and Carley's study using randomly generated graphs, the actor to be explored was selected by their degree or betweenness centralities in the sampled network at each iteration. To test the performance of selecting actors by the specified SNA measure, the sampled network was compared via Hamming distance to the "true" network after a specified number of iterations. The results indicated that selection by degree or betweenness centralities performed better than random selection (Tsvetovat & Carley, 2007, pp. 69-70).

A real world data study conducted by Zilla, Grippa, Gloor, and Laubacher (2006) utilized betweenness and closeness centralities to select the next actor for each snowball iteration. The sampled graphs' betweenness and closeness centrality scores were compared to the "true" values. The experimental conclusions showed a quicker convergence of the sampled graphs' measures, either betweenness or closeness centrality, when snowballing was accomplished via selecting the node with the highest betweenness or closeness centrality, respectively. The stated reasoning lies within the global nature of betweenness and closeness centrality. The authors identified that "hubs that connect the different parts of the community, determine the global characteristic of the community" (Zilli, Grippa, Gloor, & Laubacher, 2006, p. 6). Through selection of actors to explore via betweenness and closeness centralities, there is an increased likelihood of including a hub into the sampled graph in an earlier iteration than by randomly selecting nodes.

2.5 Cognitive Social Structures

Krackhardt (1987) introduced cognitive social structures as an alternative perspective to avoid the informant accuracy problem. Noting that much of the social science theory of human behavior and interactions was based upon individuals' cognitive and psychological perceptions as opposed to objective behavior, Krackhardt partitions a network into each of the actor's own perspectives. Instead of the traditional sociomatrix representing the network, *n* sociomatrices, one for every actor in the network, are generated. Each sociomatrix represents an individual actor's perspective on the existence of relationships within the entire network. Individual actors assess the existence of relationships between other actors within the network (Krackhardt, 1987, pp. 113-114).

2.5.1 Consensus Structure Aggregation.

These numerous perspectives are aggregated into a single representation of the social network. Krackhardt (1987, pp. 115-118) presents three separate aggregation rules to construct the single sociomatrix to be used in the subsequent traditional social network analysis. However, additional value can be obtained from having multiple perspectives of the interactions within the social network. The first aggregation technique, referenced as slice, involves a single actor's perspective of the whole social network. It is simply one of the *n* sociomatrices that are provided by the network's actors (Krackhardt, 1987, pp. 115-116). Demonstrated in the empirical example accompanying the introduction of cognitive social structures, potential insights can be gleaned from comparing an individual actor's perspective of the network and their own position within it against the

aggregation of the other actors, compiled via either or both of the other two aggregation techniques (Krackhardt, 1987, pp. 119-125).

Another aggregation technique, locally aggregated structures, represents the traditional social network construction process. Relationships in the sociomatrix representing the entire network are included based upon single nomination or reciprocal nomination from the actors involved in the dyad. Extended to cognitive social structures, each element in the aggregated sociomatrix is constructed from either the intersection or union of the corresponding elements from the sociomatrices of the corresponding actors of the representative dyad. A relationship exists in the network if one or both, dependent upon the nomination rule, of the actors involved state the relationship exists in their respective sociomatrix. The intersection of the sociomatrices' data elements corresponds to a reciprocated nomination rule, while the union designates a single-nomination rule for inclusion in the aggregated social network (Krackhardt, 1987, pp. 116-117).

The final aggregation technique, consensus structures, constructs the aggregated sociomatrix as a function of all of the individual actor perceptions. In general application, the determination of each element in the aggregated sociomatrix results from a simple threshold function which tallies the number of sociomatrices, out of the *n* actor matrices, that identify the specific relationship. If the total number exceeds an analyst specified threshold, the relationship is included in the aggregated sociomatrix (Krackhardt, 1987, pp. 117-118).

2.5.2 Consensus Structure Aggregation Limitations.

There are several shortcomings of consensus structure aggregation. First, the requirement of possessing actor perspectives for all individuals within the social network implies each actor must indicate and characterize the relationships existing between all pairs of actors within the network. As this must be accomplished for each relation modeled, the data collection requirements may be quite extensive as the number of actors within the network expands or the number of relations grows. Within the literature, consensus structure aggregation has been performed on relatively small networks, approximately twenty-five actors (Krackhardt, 1987; Neal, 2008).

Neal (2008) applied consensus structure aggregation as a method to mitigate missing data. She investigated relationships among school children and as a condition of dealing with minors had to obtain parental consent forms. As a result, only fifteen of the twenty-three students participated in the data collection. This relaxation of obtaining perspectives for all actors in the social network due to experimentation necessity utilized a threshold function to construct the network. If a specified number of actors identify an existing relationship between two actors, a relationship is determined to exist, even if neither of the actors involved submitted data. This allows information to be acquired on the individuals not participating in the data collection.

The common usage of a threshold function in consensus structure aggregation highlights a significant gap in the literature. There are no guidelines on what is an appropriate threshold value for a given network. Krackhardt (1987) in his presentation of the empirical application of a threshold function arbitrarily chose one-half, i.e. at least half of the actors must nominate a relationship for it to be included in the aggregated

social network. Neal (2008) applied three different thresholds for comparison: a majority rule, an average rule, and a binomial rule. The majority rule is equivalent to a straightforward threshold of one-half, as applied by Krackhardt (1987). The average rule is not as stringent as the majority rule, and sets the threshold as the average number of respondents reporting a tie between actors. The binomial rule uses a value derived from the binomial distribution as the threshold to determine if a relationship is deemed to exist. For each pair of actors, the threshold is set at the value obtained from the binomial cumulative distribution function for a given α, with the number of trials defined as the number of respondents who reported on the pair and the probability of success equaling the total number of relationships identified across all pairs and sources divided by the total possible number of relationships across all pairs and sources (Neal, 2008, pp. 150-152).

The general application of the threshold function involves a common threshold for all relationships for entry into the aggregated social network, with the exception of Neal's (2008, pp. 151-152) binomial rule. This common standard is appealing due to its ease of implementation. However, as cognitive social structures have been generally applied to small networks, it can safely be assumed, particularly from the context of the empirical studies, that all actors involved are aware of the other actors and to some extent the inter-relationships among them. If larger networks are under investigation, this assumption may not be sound, especially if examining dark networks. Potentially, larger networks could be decomposed into smaller sub-networks to address this assumption; though the impact of network decomposition upon consensus structure aggregation is an open research question.

In a large or dark network context, as each actor provides their perspective on the network, they may be queried about relationships among actors of whose involvement in the social network they may not even be aware. As most sources provide information of the workings and organization and are not explicitly questioned on relationships and actors, this network perspective effect can even be more pronounced. Determining a threshold based upon the number of actors in the network may artificially skew the resulting aggregated social network representing the cognitive social structure.

If a source is unaware of the existence of one or both of the actors in the dyad under consideration, then they must be unaware of any potential relationship existing among them. Under the current application of thresholds found in the literature, if only a small proportion of a social network are aware of a few individuals, despite complete agreement on their inter-relations the relationships may fail to reach the specified threshold due to remaining sources non-confirmation of the dyad.

2.6 Bayesian Approach to Imperfect Social Network Data

Butts (2003) introduces a Bayesian approach to construct conditional probability distributions of a social network model utilizing several information sources of unknown reliability. The conditional probability distributions are used to simulate the joint probability distribution to ascertain quantities of interest. The joint probability distribution can also be used to construct point estimates of the joint posterior mode of the distribution and the modes of the individual edges of social network model resulting from an aggregation of the information sources (Butts, 2003, p. 129). Butts' provided framework enables an assessment of informant accuracy (Butts, 2003, p. 132).

Several assumptions and specification of prior probability distributions are necessary to implement the Bayesian approach. Butts assumes the set of vertices composing the social network graph is known (Butts, 2003, p. 107). The underlying social network graph's parameters are assumed to be known *a priori*, thus Social Network analysts must have, at the minimum, insight into characteristics of the network's structure (Butts, 2003, p. 110). Resulting from the Bayesian implementation, an individual arc inference is assumed to be independent of the other arcs. Thus, an inference or conclusion on the status of an arc is independent of inferences of other arcs (Butts, 2003, p. 113.). The methodology incorporates multiple sources providing social network data. An information source provides social network data with associated false positive and false negative probabilities, Type I and Type II error, which are assumed to be independent of each other (Butts, 2003, p. 115). The distribution of the Type I and Type II errors is assumed and specified by the Social Network analyst for each information source (Butts, 2003, p. 121).

The prior probability distributions of underlying network graph parameters and source Type I and Type II error assumptions are necessary to utilize the Bayesian framework to compute posterior conditional probability distributions of the final social network model constructed from this application. Quantities of interest are then estimated by simulated drawings from the joint probability distribution via a Gibbs sampler which utilizes the constructed posterior conditional probability distributions (Butts, 2003, p. 117). The conditional probability distributions are derived by estimating their associated parameters via "simple counts of successful and unsuccessful

identifications of the non-existence of ties, combined with the prior parameters (Butts, 2003, p. 119)."

The joint probability distribution can generate point estimates of the joint posterior mode of the distribution and the modes of the individual edges of the social network model resulting from an aggregation of the information sources. The complexity due to high dimensionality may necessitate heuristic search techniques to maximize the likelihood function. The aggregated social network model is constructed by examining the posterior joint probability distribution for each arc individually and including arcs whose posterior probability is greater than a specified threshold of 0.5 (Butts, 2003, p. 129).

Information source accuracy can be assessed by examining the posterior distribution Type I and Type II errors associated with each source. Butts noted that the Type I and Type II error posterior probability distributions are not independent (Butts, 2003, pp. 132-133). However, experimentation conducted on 15 node networks found the approach was robust to inaccuracy in the Type I and Type II errors' prior probability distribution parameters (Butts, 2003, p. 134).

2.6.1 Bayesian Approach Limitations.

When investigating dark networks, assumptions made on the presupposed underlying network structure could significantly depart from the actual social network. The Bayesian approach was introduced with a caution, "each researcher, however, should be careful to select network priors which are accurate depictions of his or her prior information, and should avoid blind reliance on pre-packaged choices (Butts, 2003, p.

111)." Considering it is unlikely for any single or even collection of sources to possess the ability to accurately characterize a social network's structure with enough specificity to accurately estimate network parameters, heeding this caution would prove to be difficult or impractical for the Social Network analyst. This difficulty of accurately estimating the network's structural parameters as an input to implementing this Bayesian technique imposes an additional challenge upon the Social Network analyst which may reduce likelihood of adoption or correct execution. Additionally, no method was provided to validate this *a priori* assumption. The potential exists for this assumption to substantially impact the subsequent SNA results and conclusions.

The Social Network analyst must specify the probability distribution types and their associated parameters characterizing Type I and Type II errors which must be accomplished for each information source. Initial testing by Butts, indicated a preliminary robustness to inaccuracy in the analyst specified probability distributions' parameters, further testing would be necessary to validate this claim on larger networks or graphs with varied structural characteristics (Butts, 2003, p. 134). Despite this presumed robustness, the methodology assesses informant sources' accuracy based on assumptions of the sources' accuracy. In some instances, social network analysts would probably balk at having to assess sources' accuracies based upon very limited data or no previous reporting history.

The Gibbs sampler allows simulation of draws from the joint posterior probability distribution, but is theoretically based upon a convergence in the limit. Thus, utilizing the Gibbs sampler introduces several standard issues associated with simulation approaches. Depending on the particular instantiation, the Gibbs sampler may have an

associated burn-in time, iterations occurring prior to convergence for which those results must be discarded, and a sampling frequency (Butts, 2003, p. 120). The complexity involves determining the number of iterations necessary for convergence. As these necessary specifications are critical to the methodology, proper execution would probably require automated decision criteria as they are probably beyond the scope and expertise of typical social network analysts.

2.7 Social Network Source Data

Sources provide information on actor inclusion in a social network via reporting on the existence of relationships among individuals. Sources identify, and sometimes quantify, relationships among actors, which in a SNA setting occurs by nominating dyads for inclusion into the social network model. Simultaneously and unwittingly, sources implicitly identify null relationships, dyads that are not present in the network. As sources provide listings of existing relationships between actors, information on the network structure can be gleaned from the relationships not explicitly mentioned.

Each source provides a social network model, either a limited or complete representation of the organization under investigation. Much of the focus in the traditional application of SNA is directed upon the relationships nominated by each source. Generally neglected are the null relationships, the nonexistence of interaction between two actors, identified by each source. Given that social networks are often sparse, the number of identified null relationships provided by sources, albeit not explicitly, will generally compose the majority of the information delivered.

Sources identify existing relationships within a social network, generally by explicitly identifying actors and the interrelationships among them. If a source identifies n actors and their associated dyads, they are in fact commenting on $n(n\text{-}1)$ directed relationships or $n(n\text{-}1)/2$ undirected relationships. As a source identifies m relationships among the identified n actors, it will be assumed that the remaining potential dyads are null relationships, totaling $n(n\text{-}1) - m$ or $[n(n\text{-}1)/2] - m$ for directed or undirected networks, respectively. Thus, dyads are identified as absent by either the source's explicit confirmation or by implicit confirmation by the source reporting on both actors of a dyad but failing to confirm an existing relationship between them. Empirically observed, social networks are generally considered sparse networks or low density networks; thus, the number of relationships present in the network is substantially less than the number of null relationships.

When constructing social network models, generally all information provided by sources is incorporated into the model. In some cases, social network analysts may withhold sources' information due to suspicions of inaccuracies or duplicitous behavior. This determination of source reliability is based upon a number of considerations including previous reporting accuracy, inherent source trustworthiness, and other factors. Ideally, assessing source reliability is based upon confirmation or discrediting of previous reporting.

Confirming or discrediting reported data can be accomplished when the true social network model is known. Due to difficulties in collecting social network information even with cooperating actors, let alone a dark network, the likelihood of possessing a correct model is often low. Additionally, possessing the true social network

model obviates the requirement for acquiring sources of information reporting on the network.

2.7.1 Source Evaluations.

The DoD's primary definition of a source is "a person, thing, or activity from which information is obtained (Department of Defense, 2001, p. 429)." The importance of a source's reliability and the credibility of their provided information is identified in DoD doctrine. However, despite providing a scale specifying levels of reliability and credibility as recreated in Table II-13, no scoring criteria are detailed (Department of Defense, 2004, pp. III-33 - III-35).

Table II-13 Evaluation of Source Reliablity and Information Credibility

Reliability of the Source		Credibility of the Information	
A	Completely Reliable	1	Confirmed by Other Sources
B	Usually Reliable	2	Probably True
C	Fairly Reliable	3	Possible True
D	Not Usually Reliable	4	Doubtful
E	Unreliable	5	Improbable
F	Reliability Cannot be Judged	6	Truth Cannot be Judged

(Figure III-17 in Department of Defense, 2004, pp. III-35)

A source's reliability is defined by the probability of correctly reporting an event occurring and the probability of correctly reporting the nonoccurrence when appropriate (Schum & Kelly, 1973, p. 406). Encapsulating this definition mathematically, a source's reliability is defined by their reports' false positive and false negative probabilities. In a SNA application, a source's reliability is determined by their correct reporting of the status of an existing or non-existing relationship between two actors.

An unreliable source is defined by the analyst, in which the source's reporting does not meet an acceptable level of reliability. Thus an unreliable source is defined by the situation faced by the SNA analyst. If the SNA conclusions are associated with high operational risk decision-making, the acceptable level of reliability may be greater than when compared against low operational risk decisions.

In dealing with dark networks, an additional consideration in data collection is deceptive sources. Deceptive sources are intentionally providing inaccurate information and, if unrecognized, are mathematically no different than unreliable sources. However, deceptive sources may collude to present multiple sources providing confirming reports, making detecting their inaccuracy more difficult, creating a "Jones' Dilemma" (Department of Defense, Joint Publication 3-13.4, 2006, pp. A-1). Multiple reports providing confirmation of the same data increase the likely acceptance of the imperfect information. Some of the provided information may be accurate to further improve the likelihood of acceptance and disguise the introduction of imperfect information.

2.7.2 Individual Source Assessment.

Individual sources will need to be characterized to determine their validity. Several attempts to predict accuracy are present in the literature; however, past efforts defined accuracy for a source relative to a known true social network. These studies investigated predicting the accuracy of a source by examining characteristics of the source and relating them to its overall performance, measured by comparing the source's reporting to the known social network. This is useful as a forensics methodology and a predictive model given that the analyst has accepted the inherent risk of applying

generalities to a specific problem, despite potential differences rendering the generalities non-applicable. Unfortunately, considering the risks and implications faced by government organizations employing SNA against dark networks, applying broad based generalities derived from open, bright networks may precipitate serious unintended consequences.

2.7.3 Informant Accuracy Assumptions.

Nonetheless, utilizing data collection sources which possess limitations necessitates some assumptions on the sources and the data derived from them. Romney and Weller (1984, p. 61) present four assumptions in their investigation of predicting informant accuracy. These assumptions will be expanded upon and extended to account for dealing with a collection of sources.

> Assumption 1. There exists an objective set of "facts" or reality pertaining to the pattern of interaction of the group under investigation.
>
> Assumption 2. Individuals vary in the extent to which they know all the facts or reality pertaining to the pattern of interaction of the group. We refer to this as knowledge.
>
> Assumption 3. The knowledge of each individual about the group is assumed to be independent of the knowledge of every other individual.
>
> Assumption 4. The correlation of knowledge between any two subjects is a function of the extent to which each has knowledge of the objective reality. Specifically, the correlation of knowledge between individual A and individual B is the product of the correlation of individual A with the "truth" and of individual B with the "truth". (Romney & Weller, 1984, p. 61)

2.7.3.1 Examining Informant Accuracy Assumption One.

Further clarification is required for examining each of these assumptions in the context of the methodological approach presented in Chapter III. Assumption one states

"there exists an objective set of 'facts' or reality pertaining to the pattern of interaction of the group under investigation (Romney & Weller, 1984, p. 61)." The collection of information generated from all sources produces a representation of the objective social network. Unfortunately, this representation is most likely not an exact model of reality; and, even with full participation from all actors involved in the social network, a completely accurate model is probably unobtainable. The representative social network that the analyst obtains from various information sources, will be missing data elements and probably contain spurious data, particularly in the case when dealing with dark network organizations.

The objective is to remove spurious information while minimizing the amount of missing data. Undoubtedly, any process detecting incorrect information also has a nonzero probability of misclassifying correct information, i.e. a false positive. Conversely, correct information can be misclassified as erroneous and selected for subsequent removal, i.e. a false negative. Extending this schema to the collection of sources, a source identified as generating substantial amounts of erroneous data will be classified as an unreliable source and all information obtained from that specific source is discarded in the analysis.

2.7.3.2 Examining Informant Accuracy Assumption Two.

Assumption two states "individuals vary in the extent to which they know all the facts or reality pertaining to the pattern of interaction of the group. We refer to this as knowledge (Romney & Weller, 1984, p. 61)." Romney and Weller were addressing social networks where the information was provided by individuals reporting their own

interactions. For the purposes here, individuals can be extended to individual sources. Each individual source possesses inherent capabilities and limitations generating their own perspective of the social network. Some sources are unable to collect specific types of relationships. Electronic surveillance or intercepting communications will provide a view of the social network but may be oblivious to face-to-face interactions. Human sources reporting on interactions within the social network are likely to be constrained to interactions they can directly observe or learn about from trusted confidants.

Each source may paint a different picture of the social network; hopefully, the collective canvas formed from their individual portrayals presents a complete picture of the social network. However, in terms of validating individual sources, the overlap between the sources will enable the decision criteria for source inclusion or exclusion. To assess an individual source, redundant information with other sources is necessary to validate the individual source. A source may contain some information that is redundant with the other sources and some information that is novel to the collection. Judging the accuracy and validity of the redundant information enables a decision to be made on the inclusion or exclusion of the novel information. Utilizing individual source information that can be compared, vetted, and validated against the aggregated sources will enable the accuracy of the source to be assessed.

2.7.3.3 Examining Informant Accuracy Assumption Three.

Assumption three states "the knowledge of each individual about the group is assumed to be independent of the knowledge of every other individual (Romney & Weller, 1984, p. 61)." Individual is extended to individual sources and the group refers

to the social network under study. Dark networks, when confronted with opposing government forces, may initiate a deception campaign. An individual or perhaps several sources could purposely provide false information on the interactions within the social network. It would be very difficult to ascertain if an individual source is providing erroneous data due to misconceptions or purposeful deception. However, if there are multiple sources, all of which are determined to be providing false data, especially if the false data is correlated, it might be indicative of an active deception campaign on behalf of the dark network organization being targeted.

2.7.3.4 Examining Informant Accuracy Assumption Four.

Assumption four states 'the correlation of knowledge between any two subjects is a function of the extent to which each has knowledge of the objective reality. Specifically, the correlation of knowledge between individual A and individual B is the product of the correlation of individual A with the 'truth' and of individual B with the 'truth' (Romney & Weller, 1984, p. 61)." This assumption is modified when applied to assess the individual source accuracy as part of a collection of sources. Some sources may be assumed to be trustworthy in the information they provide, for example, those utilizing technical means or a trusted agent. These trusted sources are used in a vetting role for the sources whose reliability is indeterminate. As such, individual sources are not in actuality being compared and correlated with the "truth", which is unknown, but are being compared against the collection of sources whereas trusted sources are held in higher regard.

Trusted sources are so named by the social network analysts' ability to discount the possibility of deceptive information being intentionally generated by these sources. Trusted sources will generally not provide a complete picture of the social network under investigation, but will only observe a subset of the interactions among the network's actors. Technical collection means, for example, may only observe a subset of actor interactions, as in the case of wiretapping telephones which are oblivious to face-to-face communications that may occur among individuals under suspicion. Additionally, trusted sources are not precluded from providing inaccurate information. A trusted source's observations may reflect an inaccurate assessment of an interaction in the social network. For example, a trusted agent may observe interactions among actors which are purely social, but may be misconstrued as being related to the organization of interest and its associated activities.

A social network analyst may also possess an *a priori* presumption of reliability and accuracy of data provided by trusted sources. For certain organizations under investigation, their detailed knowledge of technical collection means or emplaced agents within their organization may be limited. Operating under this assumption, social network analysts may incorporate information provided by trusted agents as reliable and accurate. This assumption may be completely unfounded, particularly if the organization under investigation has taken active measures to thwart and disrupt trusted sources' collection means. Activities, such as providing deceptive information over technical communication mediums with the expectation of government interception, could invalidate data provided by a trusted source. Failure to account for the possibility of

trusted sources providing unreliable or inaccurate data could corrupt the subsequent social network analysis results.

2.8 Measuring Source Agreement

Several statistical measures are available to measure the concordance, or agreement, of reporting among various social network information sources and informants. Any specific social network data stems from one or more sources of information and these sources may corroborate or disagree on the existence of relationships among specific actors in the network. Source comparisons may generate insight into the reliability of the provided information as sources in agreement may be more likely to provide reliable information and sources providing discordant data may indicate of potential erroneous reporting.

The complete collection of information sources characterize the status of dyads, whether existing or null, and ultimately actor inclusion, within the social network under investigation. If multiple sources are reporting the same status on a particular dyad, it is reasonable to assume that their agreement is an accurate reflection on the existence of the dyad. If there is disagreement among the reporting sources, it may be difficult to assess the presence of the dyad if one cannot trust some or all of the sources. "The degree of agreement among the raters [sources] provides no more than an upper bound on the degree of accuracy present in the ratings (Fleiss, Levin, & Paik, 2003, p. 598)." Thus, source agreement is a reflection of the accuracy of the data and an indication of individual sources' reliability.

2.8.1 Concordance Among All Sources' Reporting – Fleiss' Kappa.

By considering the entire collection of social network information sources, Fleiss' Kappa, $\hat{\kappa}_F$, can be used to assess the interrater reliability across all of the sources. Fleiss' Kappa ranges from [0, 1.0], with larger values indicating greater concordance among the information sources. Given a total of I dyads across all sources, it is necessary to calculate the number of sources commenting on each dyad i, denoted m_i, including sources identifying the dyad's existence or the null relationship. For each dyad i of the I total dyads, the number of sources confirming its existence, x_i, is also recorded. Fleiss' Kappa, Equation (2.36), is then computed based upon the average number of sources commenting on each dyad, $\bar{m}$ and the overall proportion of dyad confirmations, $\bar{p}$ (Fleiss, Levin, & Paik, 2003, pp. 611-612).

$$\bar{m} = \frac{\sum_{i=1}^{I} m_i}{I} \tag{2.33}$$

$$\bar{p} = \frac{\sum_{i=1}^{I} x_i}{I\bar{m}} \tag{2.34}$$

$$\bar{q} = 1 - \bar{p} \tag{2.35}$$

$$\hat{\kappa}_F = 1 - \frac{\sum_{i=1}^{I} \frac{x_i(m_i - x_i)}{m_i}}{I(\bar{m} - 1)\bar{p}\bar{q}} \tag{2.36}$$

A significance test evaluating the null hypothesis that Fleiss' Kappa is equal to zero compares the z statistic obtained from Equation (2.39) against a standard normal distribution. The z statistic divides the value obtained for Fleiss' Kappa against its standard error. Calculating the standard error involves computing the harmonic mean of the number of ratings per subject, $\bar{m}_H$ (Fleiss, Levin, & Paik, 2003, p. 613).

$$\bar{m}_H = \frac{I}{\sum_{i=1}^{I} 1/m_i} \tag{2.37}$$

$$\widehat{se}_0(\hat{\kappa}) = \frac{1}{(\bar{m}-1)\sqrt{I\bar{m}_H}} \sqrt{2(\bar{m}_H - 1) + \frac{(\bar{m}-\bar{m}_H)(1-4\bar{p}\bar{q})}{\bar{m}\bar{p}\bar{q}}} \tag{2.38}$$

$$z = \frac{\hat{\kappa}}{\widehat{se}_0(\hat{\kappa})} \tag{2.39}$$

2.8.2 Pairwise Comparisons.

To satisfy source to source pairwise comparisons, confusion matrices are utilized in this study. Though traditionally applied to compare a classifier's performance against a known set of objects, a minor adaptation makes confusion matrices suitable for the problem at hand. The confusion matrix, termed a comparison matrix here, can capture the amount of concurrence and disagreement between two sources. This enables simple construction of a comparison matrix by counting the number of agreements of dyad existence, cell *a* in Table II-14, the number of null relationship agreements, cell *d* in Table II-14, and the number of disagreements distinguished by every sources' selection of relationship agreement, cells *b* and *c* in Table II-14. Since the focus of concurrence is upon dyads, both nodes composing the dyad must be reported by both sources for inclusion into the comparison matrix. Thus binary vectors for each source can be constructed, restricted to only contain dyads in common. If consistently ordered to ensure dyad comparisons across sources, these binary vectors can be compared on an element by element basis to construct the comparison matrix.

Table II-14 Comparison Matrix for Source Comparison

		Source 2	
		Dyad Present	Null Relationship
Source 1	Dyad Present	*a*	*b*
	Null Relationship	*c*	*d*

2.8.3 Binary Similarity Measures.

Binary similarity measures have been continually introduced since 1884, when Sergeant Finley published a tornado forecasting ability with an accuracy greater than 95%. Gilbert (1884) responded noting that by always predicting no tornados an accuracy of 98.2% can be achieved. Gilbert introduced two measures, which were soon followed by Peirce (1884) and Doolittle (1885) presenting alternative measures (Murphy, 1996, pp. 3-7). Since these initial introductions, binary similarity measures have been proposed and utilized in a wide range of disciplines, such as weather forecasting, biology, and others (Murphy, 1996, p. 3; Choi, Cha, & Tappert, 2010).

As it is assumed that multiple sources are reporting on the social network under investigation, a confusion matrix can be generated for each source pairing. Evaluating every source pairing via cross-examination of each confusion matrix's elements quickly becomes unwieldy, particularly when considering the variance in the quantity of information provided by each source. Fortunately, a variety of binary similarity and distance measures have been introduced in the literature to reduce a confusion matrix to a single scalar value representing the similarity, or conversely the dissimilarity or distance, between the two objects.

Due to the widespread applicability of binary similarity and distance measures, numerous equations have been introduced. This plethora of measures has initiated debates on their utility, applicability, and interpretation. Several extensive survey studies of these measures exist: Warrens' dissertation (2008) presented 56 binary similarity measures, Choi *et al* (2010) recently surveyed 76 binary similarity and dissimilarity measures, and Eidenberger (2011) listed 75 measures. This author identified two additional measures present in the literature not identified in the survey studies: *Proportion of Specific Agreement* (*ignoring d*) and *Rogot & Goldberg* (1966) (Fleiss, Levin, & Paik, 2003, pp. 599-602). Duplication or alternative measure names exist on the survey studies, and a comprehensive listing of the measures, alternative names, and the associated equations is provided in Table A-1 and Table A-2 in Appendix A. Even synthesizing the lists requires evaluating 105 distinct binary similarity and dissimilarity measures to determine the most suitable for the proposed analysis. Selecting an appropriate measure from a large set of binary similarity and dissimilarity measures complicates the task of conducting pairwise source comparison to assess source reliability; clearly a methodology to select the appropriate measure or set of measures is required. A methodology to accomplish this has been developed and is presented in Chapter IV.

2.8.3.1 Characterizing Binary Similarity/Dissimilarity Measures.

Binary measures are first distinguished by whether they are attempting to capture the degree of similarity among objects. Choi's (2010) description of 76 binary measures included 17 binary dissimilarity measures. Of note, several of these dissimilarity

measures are directly derived from a corresponding similarity measure. For example, the *Yule-Q* measure was introduced as a similarity measure with a corresponding dissimilarity measure, also referenced as *Yule-Q*, computed as one minus the similarity measure. Another example, the *Gleason* similarity measure is one minus the *Lance & Williams* dissimilarity measure. Of greater complexity, the *Hamann* similarity measure is equivalent to the *Sokal & Michener* similarity measure minus the *Mean Manhattan* dissimilarity measure. The *Variance* dissimilarity measure is equal to one-fourth of *Mean Manhattan*. These measure pairings are equivalent to each other and will always be perfectly positively or negatively correlated. Therefore, it is only necessary to investigate one of the measures in each corresponding pairing for its suitability for a given application. Table II-15 includes identified relationships among the binary similarity and dissimilarity measures.

Additionally, several similarity measures are perfectly correlated due to scale changes in their associated equations as illustrated in Table II-15. The *Sokal & Sneath*-II similarity measure is twice the *Gower & Legendre* measure, ranging on [0, 1] and [0, 0.5], respectively. The *Driver & Kroeber* similarity measure is half of the *Johnson* measure, ranging on [0, 1] and [0, 2], respectively. The dissimilarity measures *Hellinger* and *Chord* are perfectly correlated with *Hellinger* ranging on [0, 2] and *Chord* [0, $\sqrt{2}$]. The *McConnaughey* and *Kulczyński*-II similarity measures range from [-1, 1] and [0, 1], respectively. Since these measures are perfectly correlated, only a single measure in each pairing is required in this analysis. The *Sokal & Sneath*-II, *Driver & Kroeber*, *Hellinger*, and *McConnaughey* measures' ranges are more suited for analyst interpretation due to resemblance with the conventional Pearson's correlation range of [-1, 1].

Some of the measures presented are known by several names (Choi, 2008). Several binary measures distinctly identified by Choi *et al* (2010) are derived from measures that reduce in the binary case to another measure. The *Squared-Euclidean*, *Canberra*, *City Block*, and *Minkowksi* dissimilarity measures all reduce to the *Hamming* distance in the binary case. Similarly, the *AMPLE* similarity measure reduces to the *Tarantula* similarity measure. Other measures are effectively the same, but introduced in the literature in different fields by various authors. The *Gleason*, *Dice*, *Sørenson*, *Czekanowski*, *Nei and Li* similarity measures are computed via the same expression. *City Block* and *Manhattan* are two common names for the same measure. The *Ochiai*-I and *Otsuka* similarity measures are binary versions of *Cosine* similarity (Choi, 2008, p. 44). The *Lance & Williams* dissimilarity measure is computed by the same equation as the *Bray & Curtis* dissimilarity measure. The *Tanimoto* similarity measure reduces to the much older *Jaccard* measure. *Loevinger's H* is algebraically equivalent to the *Forbes*-II measure. Alternative names for the measures are compiled in Table A-1 and Table A-2 in Appendix A.

Table II-15 Binary Measures Relationships

Similarity Measures [range] (alternative names)	Correlated Measures [range]	Obverse Dissimilarity Measure [range]	Negatively Correlated Measures [range]
Gleason Dice Sørenson (Coincidence Index) (Quotient Similarity) Czekanowski Nei & Li (Genetic Coefficient)		Lance & Williams = Bray & Curtis = 1 - Gleason	
		Hellinger [0, 2]	Chord [0, $\sqrt{2}$]
Sokal & Michener [0,1] (Simple Matching)	Hamann [-1,1] = Sokal & Michener [0,1] – Mean Manhattan [0,1]	Mean Manhattan [0,1]	Variance [0,0.25] = MeanManhattan/4 = (1- Hamann)/2
		Size Difference = (Mean Manhattan)^2	
Yule Q [-1, 1] (Coefficient of Association)		Yule Q dissimilarity [-1, 1]	
Driver & Kroeber [0, 1] Kulczynski-II [0, 1]	Johnson [0,2] = 2 x Kulczynski-II McConnaughey [-1, 1]		
Baroni-Urbani & Buser-I [0,1]	Baroni-Urbani & Buser-II [-1,1]		
Sokal & Sneath-II [0, 1]	Gower & Legendre [0, 0.5]		
Sorgenfrei (Correlation Ratio)	Sorgenfrei = (Ochiai-I)^2		

Several of the binary measures are not scaled and do not possess a theoretical limit. Though useful if the length of the binary vectors is constant across the generated confusion matrices, for the source reliability assessment application, conducting comparison with varying total number of binary vector lengths, scaleless measures may hinder interpretability among sources. If direct comparisons among measures are important for a specific application, several measures can be eliminated from consideration. The lack of scale in several measures can easily be seen from direct observation of the equations. The *Intersection*, *Innerproduct*, *Koppen* (1884), *Browsing Pattern*, and *Tversky* binary similarity measures and the *Euclidean* dissimilarity measure are clearly not scaled due to the lack of a denominator. The equivalent *Hamming*, *Squared-Euclidean*, *Canberra*, *Manhattan*, *City Block*, and *Minkowksi* dissimilarity measures are not scaled as well since they also do not possess a denominator. The range bounds of the collected binary similarity and dissimilarity measures are available in Table A-1 and Table A-2 in Appendix A.

Additional measures are unbounded in range, though this is not readily apparent by their equations. Table II-16 displays binary similarity and dissimilarity measures that are unbounded in either their minimum possible value, maximum possible value, or both minimum and maximum values.

Up to this point, binary similarity and dissimilarity measures have been eliminated from consideration for this study due to their inherent perfect correlation with other identified measures, thus no information is lost by reducing the set of measures from 105 to the 96 distinct uncorrelated measures listed in Table II-17. We now turn our attention to what the binary similarity and dissimilarity measures attempt to describe in

Table II-16 Unbounded Range Binary Measures

Unbounded Similarity Measures		
Minimum Value	**Maximum Value**	**Both**
Fleiss	Batagelj & Bren	Browsing
Goodman & Kruskal Tau	Clement	Eyraud
	Cole-I	Köppen (1870)
	Cole-III	Maron & Kuhns
	d Specific Agreement	Stiles
	Dennis	Stuart's τ_c
	Forbes-I	Tversky
	Fossum	
	Gilbert & Wells	
	Harris & Lahey	
	Inner Product	
	Intersection	
	Köppen (1884)	
	Kulczyński-II	
	Pearson-I	
	Sokal & Sneath-III	
	Tarantula	
	Warrens-V	
Unbounded Dissimilarity Measures		
Minimum Value	**Maximum Value**	**Both**
	Euclidean	
	Hamming	

comparing binary vectors. The confusion matrix in Table II-14 displays commonalities and disagreements in reporting of the presence or absence of dyads. The weighting of these two states may not be equal. "[A]symmetric binary variables represent cases where the two states are not equally important (Choi, 2008, p. 18)." As social networks are generally described as sparse networks, equally weighting confirmation of present dyads with confirmation of null relationships may not be appropriate due to the preponderance

of null relationships. Returning to the confusion matrix, when investigating assumed sparse social network source comparison, we naturally expect cell *d* to be significantly larger in number than cell *a*. As a result, a source reporting large numbers of null relationships probably scores well under measures that consider negative matches, i.e. null relationships, in their calculations. The weighting of cells *a* and *d* varies substantially across the measures dependent upon the specifics of their associated equations. The question remains, which measures, considering the variability in the weighting mechanisms, are suitable for the application at hand--measuring social network source agreement. A methodology is demonstrated in Chapter IV that aids analysts attempting to answer that question.

Table II-17 Reduced Set of Binary Similarity and Dissimilarity Measures

Binary Similarity Measures		
Anderberg	Goodman & Kruskal Lambda	Pearson-I
Anderberg's D	Goodman & Kruskal Tau	Pearson-II
Baroni-Urbani & Buser-II	Goodman & Kruskal Max	Pearson-III
Batagelj & Bren	Goodman & Kruskal Min	Phi Coefficient
Benini	Goodman & Kruskal Prob	Peirce-I
Braun-Blanquet	Gower	Peirce-II
Browsing	Hamann	Peirce-III
Clement	Harris & Lahey	Relative Decrease of Error Probability
Cohen's Kappa	Hawkins & Dotson	Rogers & Tanimoto
Cole-I	Inner Product	Rogot & Goldberg
Cole-II	Intersection	Russell & Rao
Cole-III	Jaccard	Scott
Cosine	Jaccard-3W	Simpson
Dennis	Kent & Foster-I	SokalSneath-I
Dice-I	Kent & Foster-II	SokalSneath-II
Dice-II	Koppen 1870	SokalSneath-III
Digby	Koppen 1884	SokalSneath-IV
Dispersion	Kuder& Richardson	SokalSneath-V
Doolittle	Kuhn	Sorgefrei
d Specific Agreement	Kuhn Proportion	Stiles
Eyraud	Kulczyński-I	Tarantula
Fager & McGowan	Kulczyński-II	Tarwid
Faith	Loevinger's H	Tversky
Fleiss	Maron & Kuhns	Warrens-I
Forbes-I	Maxwell & Pilliner	Warrens-II
Fossum	McConnaughey	Warrens-III
Gilbert	Michael	Warrens-IV
Gilbert & Wells	Mountford	Warrens-V
Gini	Pearson & Heron-I	Yule Q
Modified Gini	Pearson & Heron-II	Yule W
Binary Dissimilarity Measures		
Euclid	Hellinger	ShapeDifference
Hamming	PatternDifference	SizeDifference

2.9 Classifier Performance

The objective of this research is to provide a methodology to distinguish between reliable and unreliable information sources when constructing a social network model. The test the methodology's effectiveness a method of evaluating classification performance is required.

2.9.1 Response Operating Characteristics (ROC) Graph.

A Response Operating Characteristics (ROC) Graph provides a method to examine a classifier's performance and conduct comparisons among classifiers. A classifier when examining an object that can be one of two classes produces four possible outcomes: a true positive, a false positive, a true negative, or a false negative (Fawcett, 2006, p. 862). In this experimentation, a true positive denotes the methodology correctly identifying a reliable source for inclusion into the social network model. A true negative indicates the methodology correctly discarding an unreliable source from the social network model. False positives and false negatives occur when the methodology incorporates an unreliable source into the social network model and discards a reliable source, respectively. Arranging the total number of true/false positives and true/false negatives into a matrix gives the standard confusion matrix, also referred to as a contingency table, as displayed in Table II-18 (Fawcett, 2006, p. 862).

Table II-18 Confusion Matrix

		True Class: Class 1	True Class: Class 2
Classifier Assessment: Class 1		True Positives	False Positives
Classifier Assessment: Class 2		False Negatives	True Negatives

2.9.1.1 Confusion Matrix Metrics.

There are several commonly used metrics that can be obtained from the confusion matrix. The true positive rate, also referenced as the hit rate, recall or sensitivity, is defined by Equation (2.40). The analogous false positive rate, also referenced as the false alarm rate, is detailed in Equation (2.41) (Fawcett, 2006, p. 862).

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2.40}$$

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \tag{2.41}$$

The ROC graph can be constructed by plotting the true positive rate on the y-axis and the false positive rate on the x-axis. "An [sic] ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives) (Fawcett, 2006, p. 862)." The ROC graph can be constructed from plotting points of the classifier's performance as the classifier's threshold is varied. As the threshold is monotonically increased, a curve will be displayed on the ROC graph as shown in Figure II-10 (Fawcett, 2006, p. 863).

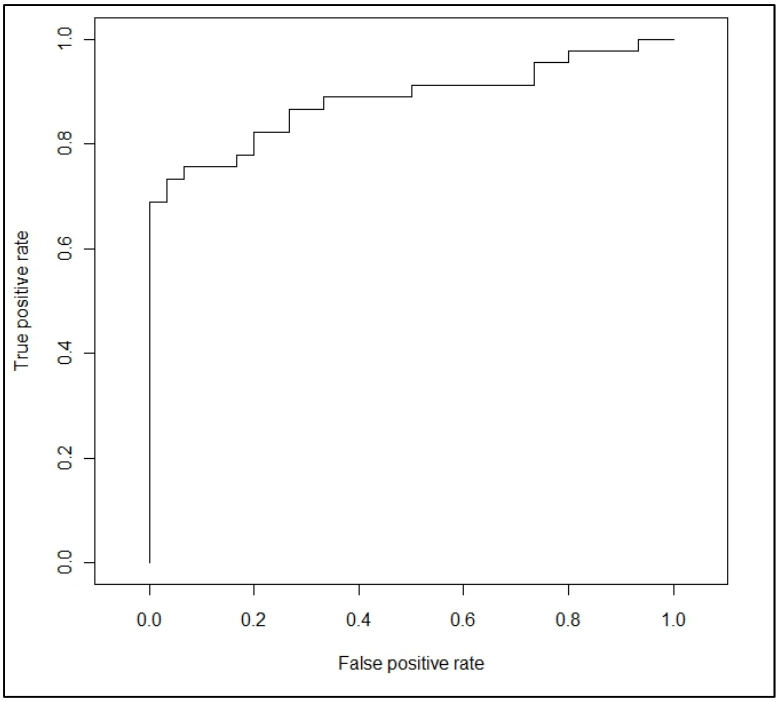

Figure II-10 ROC Graph Example

2.9.1.2 Area Under the Curve (AUC).

The area underneath the ROC curve is a commonly used summary statistic to compare classification accuracy. The AUC ranges from [0, 1.0] with a larger value indicating better classification performance. However, random guessing would generate an AUC equal to 0.5, so in practice, classifiers should at a minimum exceed this threshold. This appears as a diagonal line from the origin to (1,1) on the ROC curve. The AUC "is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chose negative instance (Fawcett, 2006, p. 868)." For comparison, the AUC for the ROC curve in Figure II-10 is 0.883.

2.9.1.3 ROC Graph Properties.

ROC graphs are insensitive to imbalances among the probability of occurrence between classes. If a great disparity in the frequency of occurrence is present, the ROC graph is unaffected as it is based on the true positive and false positive rates. These rates are indifferent to number of objects in each class (Fawcett, 2006, p. 864). Real world social network information sources are presumed to present disparity in the number of reliable and unreliable sources. Ideal situations would have greater numbers of reliable sources with only a few unreliable sources reporting.

2.9.1.4 Additional Confusion Matrix Metrics.

Additional commonly referenced confusion matrix metrics include: precision, accuracy, specificity, and F-measure. Precision, also referenced as positive predictive value, is defined in Equation (2.42). Accuracy is presented in Equation (2.43). Specificity is one minus the false positive rate. The F-measure, Equation (2.44), incorporates the precision and recall metrics (Fawcett, 2006, p. 862).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2.42}$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{False Positives} + \text{True Negatives}} \tag{2.43}$$

$$\text{F-measure} = \frac{2}{\left({}^{1}/_{\text{Precision}}\right) + \left({}^{1}/_{\text{Recall}}\right)} \tag{2.44}$$

2.10 Statistical Analysis Techniques

Several statistical analysis techniques are utilized in this research to examine the proposed methodology's performance. The design of experiment is discussed in Section

3.3. Traditional ordinary least squares (OLS) regression is applied in the analysis, but as is discussed in Section 5.7, additional statistical techniques are required due to violations of OLS assumptions. This section overviews some statistical techniques that are employed to address issues present with applying OLS on these data sets.

2.10.1 Analysis of Covariance (ANCOVA).

Analysis of Covariance (ANCOVA) is a statistical technique that has the potential for reducing large error variances (Neter, Wasserman, & Kutner, 1985, p. 845.). The high variability structural characteristics of the randomly generated graphs may account for some of the error variances in the model. These structural characteristics cannot be controlled in the random graph generation, but they can be observed using traditional SNA network measures. These covariate or concomitant variables may inflate the mean square error and mask treatment effects. By accounting for these structural characteristics, the resultant model will address the uncontrollable nuisance variables associated with individual graph characteristics (Montgomery, 2005, p. 575).

ANCOVA models the experimental factors, the covariates, and, if desired, the interaction terms among the experimental factors, the covariates, and interactions between the experimental factors and covariates. In this experimentation, the full factorial design, coupled with the center point runs, the runs from the space filling design and ten replications of each design point, will ensure there are sufficient degrees of freedom are available to conduct an ANCOVA with all interaction terms present in the model. There is justification for investigating such a complicated model:

> Leaving interaction effects involving covariates out of the model may result in biased estimates of the factor effect. Therefore, whenever the

> number of experimental runs is sufficiently large, we recommend verifying whether these interaction effects are significantly different from zero, even when they are not of primary interest. (Goos & Jones, 2011, p. 207)

The ANCOVA model with up to two variable interactions, Equation (2.45), with response vector Y, m experimental factors denoted x_1 through x_m, c covariate variables denoted z_1 through z_c can be solved for the intercept, β_0, experimental factor main effects, β_0 through β_m and factor interaction coefficients, β_{ij}, covariate main effects, γ_1 through γ_c, covariate interaction coefficients, γ_{ij}, and factor covariate interaction coefficients, β_{ij}^{EC} (Goos & Jones, 2011, p. 207).

$$Y = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \beta_{ij} x_i x_j + \sum_{i=1}^{c} \gamma_i z_i + \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} \gamma_{ij} z_i z_j + \sum_{i=1}^{m} \sum_{j=i+1}^{c} \beta_{ij}^{EC} x_i z_j + \varepsilon \tag{2.45}$$

2.10.2 Ordinary Least Squares Regression Error Term Assumptions.

Ordinary Least Squares (OLS) regression assumes that the error terms of a linear regression model are independently normally distributed with a mean of zero and the error distribution is homoskedastic (Kmenta, 1971, p. 348). However, since this experimentation data set is based on graphs these assumptions may prove to be invalid. Several statistical techniques are available to test the validity of these assumptions on an OLS model and attempt to rectify deficiencies.

Of particular interest in this experimentation is the normally distributed error terms assumption. Empirical observations of some social network measures' distributions do not follow normal distributions. Barabási and Bonabeau (2003, pp. 63-

64) note that the degree distribution of networks appear to follow a power-law distribution. As such, it can be anticipated that other network characteristics or social network algorithms possess non-normal distributions as well. Quantile regression is conceptually similar to OLS, but possesses different assumptions regarding the error terms distribution and is an alternative statistical technique when OLS' assumptions are not met.

Empirical SNA models have been found to possess certain characteristics, such as clustering, degree correlation, and power-law degree distributions (Newman & Park, 2003, pp. 036122:2-4; Barabási & Bonabeau, 2003, pp. 63-64). Even the classic random network model of Erdös and Rényi (1959) produced degree distributions following a binomial distribution, a Poisson distribution in the limit as the number of nodes approaches infinity. Noticeably absent within the networks' literature is the regular appearance of the normal distribution.

The recent ascendance of network science has brought forth discoveries of the appearance of scale-free distributions and inequalities among network significance of nodes and edges. As such, one must consider the assumptions of traditional statistical techniques when applying them to network models. Notably, the normality assumption of the error terms is prominent in ANOVA, ANCOVA, and linear regression, which are traditional analysis techniques used in Design of Experiments (DOE). Another assumption on the error term's distribution in OLS is homoskedasticity. However, in the presence of heteroskedasticity, confidence intervals cannot be placed on the regressor coefficients and testing coefficient significance is inappropriate (Kmenta, 1971, p. 255).

2.10.3 Quantile Regression Overview.

Least squares linear regression models the conditional mean response of the dependent variable for given values of the independent variables. However, if one wants to investigate the effects of variables on the response's distribution of values, techniques based on the generalized linear model are not sufficient. Understanding how network parameters and characteristics affect a response variable necessitates utilizing techniques that evaluate the full distribution of the dependent variable as opposed to only investigating mean responses.

Quantile regression is a semi-parametric linear regression model extension that models the dependent variable's quantile response, as opposed to the mean response, conditioned on independent variables (Cade & Noon, 2003, p. 414). The parametric portion of the model involves the independent variables and their associated regression coefficients, as in least squares linear regression, with the regression coefficients being interpreted as rates of change in the dependent variable. One advantage that lends this technique to network analysis includes that the model's non-parametric error terms are not assumed to follow a specified parametric distributional form, such as the normal distribution in linear regression (Cade & Noon, 2003, p. 414).

Quantile regression enables a linear model to be calculated for any desired quantile, allowing a characterization of the dependent variable's response at the tails of its distribution conditioned on modeled independent variables (Cade & Noon, 2003, p. 414). Thus, by computing quantile regression hyperplanes for a series of quantiles, an estimate of the response variable's conditional distribution can be constructed (Koenker, 2005, p. 16). Additionally, due to the ordering nature of quantiles, the regression quantile

model maintains its statistical properties under any linear or nonlinear monotonic transformation of the dependent variable (Cade & Noon, 2003, pp. 414-415).

Figure II-11 displays a comparison of ordinary least squares (OLS) and quantile regression of food expenditure by household income. The data is "based on 235 budget surveys of the 19th century working-class households (Koenker, 2005, p. 78)." As can be observed in Figure II-11, the OLS of the mean generates a distinctly different regression line than the quantile regression of the median. Additionally, Figure II-11 displays how conducting quantile regression for multiple quantiles assists in understanding the impact of hetereoscedasticity.

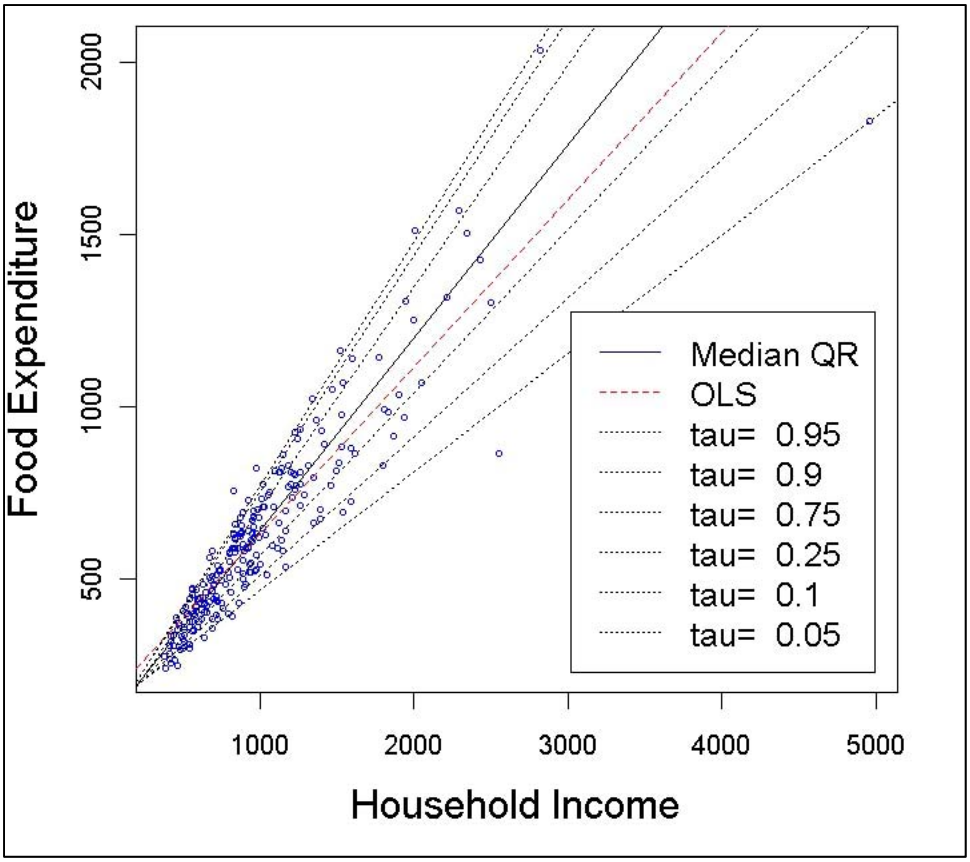

Figure II-11 OLS and Quantile Regression Comparison (Koenker, 2011)

The following description of the mechanics of quantile regression is based on two primary sources: Koenker's Quantile Regression book (Koenker, 2005), and a series publication on quantile regression (Hao & Naiman, 2007). The following sections generally follow the description, mathematical properties, and assumptions discussion of quantile regression as presented by Koenker (2005) and Hao and Naiman (2007).

2.10.4 Quantile Regression Mathematical Underpinnings.

Quantile regression is based upon an optimization problem. The objective is to minimize an expected piecewise linear loss function, given by Equation (2.46), for a specified quantile, $\tau \in (0,1)$, with indicator function *I* (Koenker, 2005, p. 5).

$$\rho_\tau(u) = u\big(\tau - I(u < 0)\big) \tag{2.46}$$

Analogous to linear regression, the objective is to minimize an aggregation of a difference function between the observed responses and the predicted responses. Linear regression minimizes the sum of squared differences between these quantities. Quantile regression uses the loss function of Equation (2.46) to minimize the weighted differences between observed and predicted response, Equation (2.47) (Koenker, 2005, p. 10). Equation (2.47) is piecewise linear and continuous. It is differentiable everywhere except the points when $y_i - x_i^T\beta = 0$ (Koenker, 2005, p. 32).

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) \tag{2.47}$$

Equation (2.47) can be reformulated as a linear program, Equation (2.48), by decomposing the residuals, into positive and negative slack variables, *u* and *v*. *X* is the standard design matrix as in traditional linear regression (Koenker, 2005, p. 10).

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}}\{\tau 1_n^T u + (1-\tau)1_n^T v \,|X\beta + u - v = y\} \tag{2.48}$$

2.10.4.1 Coefficients' Standard Errors and Confidence Intervals.

Confidence intervals and standard errors for estimated coefficients in a quantile regression model can be computed via asymptotic calculations which assume errors are identically and independently distributed. Alternatively, it is common practice to estimate the coefficients' standard deviations and confidence intervals via a bootstrap method. The bootstrap method draws samples of size *n* with replacement from the data set. The quantile regression coefficients are computed for the selected sample and a new sample is generated. Estimates of each coefficient's distribution are then calculated. The coefficients' distributions derived from the bootstrap method can be assessed by assuming a normal distribution and evaluating the means and standard deviations, or by analyzing the distributions' quantiles (Hao & Naiman, 2007, pp. 47-49; Koenker, 2005, pp. 105-107).

An example of the visualization of the quantile regression coefficients is provided in Figure II-12 for seven covariates and the intercept. The data involves a response variable of the birth weight of 198,377 babies, with fifteen covariates recorded, such as mother demographic data and health factors. The OLS regression coefficients are presented as dotted lines along with their 95% confidence intervals. As can be seen in Figure II-12, the coefficients vary by quantile, which indicates the effects of these covariates are not constant across the conditional distribution of the response variable. Confidence intervals on the quantile regression coefficients are computed via a bootstrap method. Quantile regression coefficients which appear constant across the range of

quantiles, indicate a pure location shift in the response variable due to the covariate. Non-constant coefficients indicate scale or shape changes in the response variable's conditional distribution due to the covariate (Koenker & Hallock, 2001, pp. 148-151).

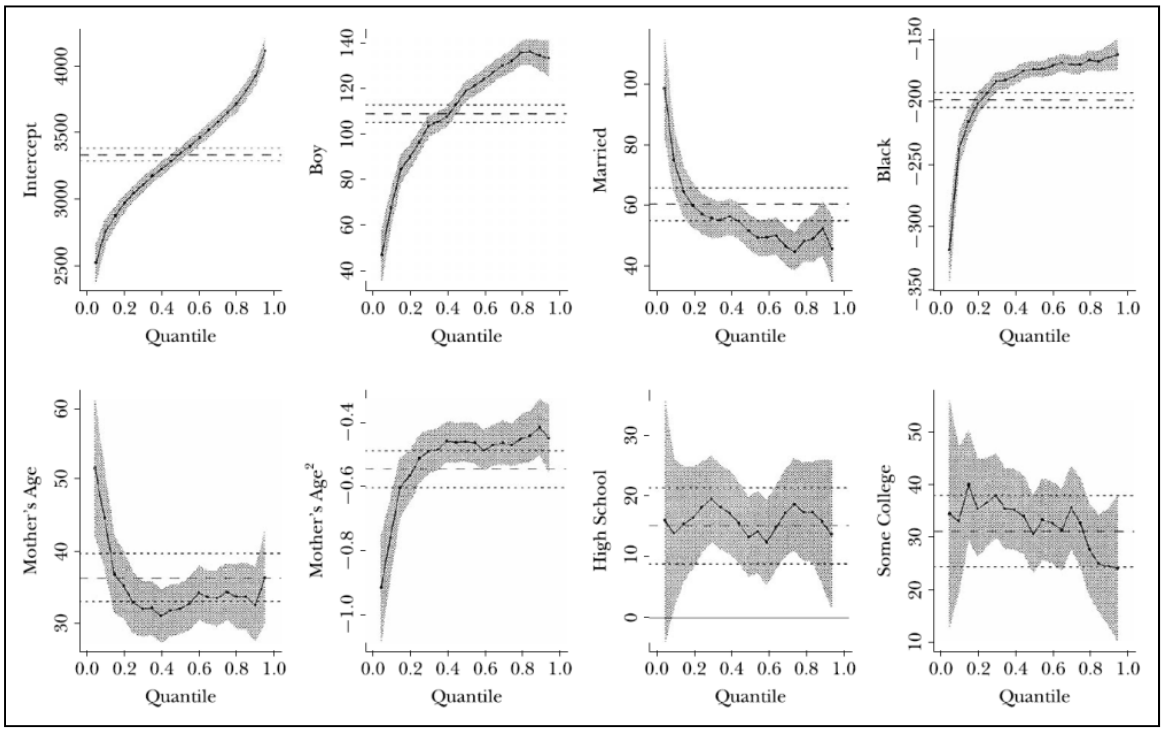

Figure II-12 Covariate Coefficient Comparison (Koenker & Hallock, 2001, p. 150)

2.10.5 Quantile Regression Properties.

Quantile regression possesses several properties that distinguish it from OLS regression. This subsection describes several properties that are relevant to the analysis conducted in Section 5.8.

2.10.5.1 Quantile and Mean Robustness.

The mean of a data set is susceptible to large outliers. The median or any quantile, on the other hand, is relatively robust to the presence of outliers. The quantile's value is only affected if a data point is perturbed enough to cross the quantile. In linear regression, which computes the conditional mean, an extreme outlier can alter the entire mean regression hyperplane. In quantile regression, the quantile hyperplane is only altered if the extreme outlier crosses the quantile hyperplane. Thus, if a median regression hyperplane is computed, changes to a data point only affects the regression equation if it changes the data point's position from above to below, or vice versa (Koenker, 2005, p. 11; Hao & Naiman, 2007, pp. 41-42).

2.10.5.2 Equivariance to Monotone Transformations.

Quantiles possesses a property referred to as equivariance to monotone transformations, which if given a variable Y, $h(\cdot)$ is a nondecreasing function on $\mathbb{R}$, the quantiles, $Q(\cdot)$, of the transformed variable are equal to the transform of the variable's quantiles, $Q\big(h(Y)\big) = h(Q(Y))$. The mean does not possess this property, but in linear regression a commonly applied technique if analysis of the residuals indicates hetereoscedasticity or non-normality is to transform the dependent variable (Montgomery, Peck, & Vining, 2006, pp. 160-161). The transformation is intended to correct the hetereoscedasticity and non-normality, while maintaining the response variable as linearly dependent on the covariates. Thus, transformations in quantile regression are subject to greater interpretability (Koenker, 2005, p. 39; Hao & Naiman, 2007, pp. 38-41).

2.10.5.3 Quantile Crossing.

Quantile regression allows one to independently construct quantile hyperplanes for any set of given quantiles. As each one is derived from an independent optimization to calculate the regression coefficients, it is possible to produce a set of quantile hyperplanes that cross. As this is inconsistent with the definition of quantiles and their monotonicity, methods are available to ensure generation of monotonic quantile hyperplanes, but in practice, it is recommended to perform quantile regression and use quantile hyperplane crossing as a potential indicator of model misspecification (Koenker, 2005, pp. 55-59). Figure II-13 shows an example of quantile crossing on data consisting of glacier lily seedlings with the number of flowers observed as an independent variable (Koenker, 2005, p. 54).

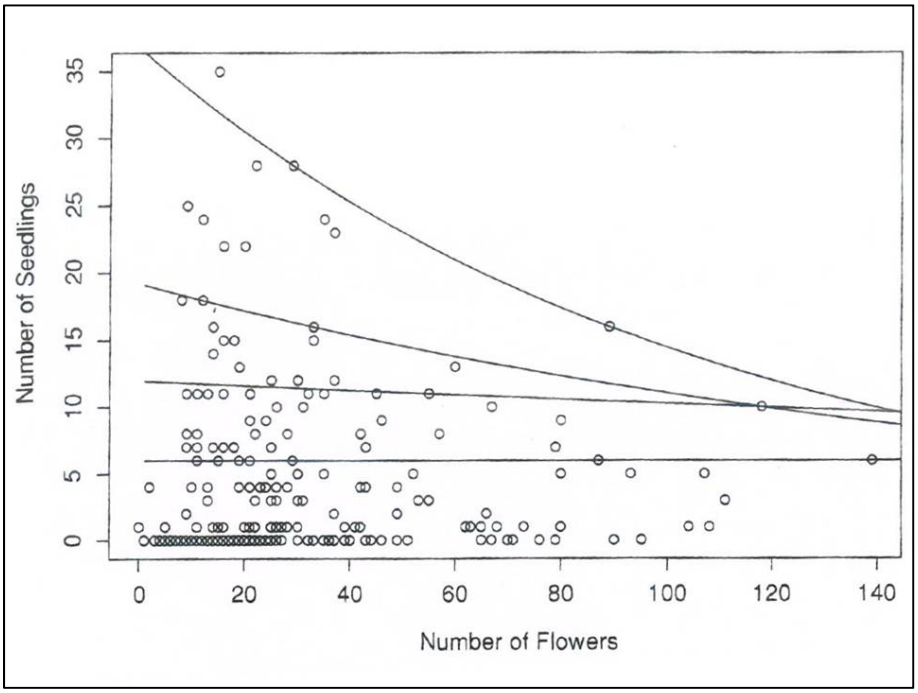

Figure II-13 Example of Quantile Crossing (Koenker, 2005, p. 55)

2.10.6 Location and Scale Shifts.

Traditional linear regression can detect location shifts of a response variable's distribution conditioned upon the independent variables. However, it assumes that the covariates affect the response variable only via a location shift in the response conditional distribution and that there are no effects impacting the distribution's scale or shape (Hao & Naiman, 2007, p. 57). Linear regression cannot detect scale changes in the response variable if the scale change does not affect the conditioned distribution's mean, unless replications in the design matrix, *X*, allow for variance computations. Additionally, linear regression may or may not detect location and scale shifts in the conditioned distribution.

Quantile regression can detect location shifts by examining the independent variables' impact upon the response variable's median, if central tendencies are of interest, or other quantiles if investigating the response variable's distribution's tail behavior. Scale changes in the response variable's conditioned distribution can be detected by quantile regression by generating several quantile regressions and estimating an interquantile range conditioned on *X*. By estimating the conditioned distribution's central location changes, coupled with scale change detection, location and scale changes of the response variable's distribution due to effects from the independent variables can be measured (Hao & Naiman, 2007, pp. 7-8).

To measure location shift, the median is analogous to the mean used in traditional linear regression (Hao & Naiman, 2007, pp. 12-13). However, if one is interested in location shifts in the tails of the distribution, any quantile could be used. For scale changes, linear regression uses the standard deviation. For symmetric distributions, such as the normal distribution, the standard deviation possesses a straightforward

interpretation. For asymmetric distributions, such as heavy-tailed distributions, the interpretation of the standard deviation is more difficult. To measure scale changes in quantile regression, an interquantile range, termed quantile-based scale measure (QSC), for a selected quantile τ, with value $Q^{(\tau)}$, as given in Equation (2.49) can be used (Hao & Naiman, 2007, pp. 12-13).

$$QSC^{(\tau)} = Q^{(1-\tau)} - Q^{(\tau)} \text{ for } p < 0.5 \tag{2.49}$$

Another aspect of a distribution's shape is skewness. With symmetric distributions, the skewness is zero. "A negative value indicates left skewness and a positive value indicates right skewness. Skewness can be interpreted as saying that there is an imbalance between the spread below and above the median (Hao & Naiman, 2007, p. 13)." Within quantile regression, a quantile-based skewness measure (QSK), can be defined as the ratio of the upper spread against the lower spread, as depicted in Equation (2.50). The QSK for a symmetric distribution is zero, is negative for left-skewed distributions and positive for right-skewed distributions (Hao & Naiman, 2007, pp. 13-14).

$$QSK^{(\tau)} = \frac{Q^{(1-\tau)} - Q^{(0.5)}}{Q^{(0.5)} - Q^{(\tau)}} - 1 \text{ for } p < 0.5 \tag{2.50}$$

2.10.6.1 Wald Statistic.

If the variance and covariances of a quantile regression model's coefficients are available, such as being estimated via the bootstrap method, the Wald statistic can be used to test whether multiple coefficients are equal across various quantiles. They can be applied to the interquantile ranges of multiple samples. "Thus, they may be considered to be tests of homogeneity of scale or tests for heteroscedasticity (Koenker, 2005, p. 75)."

The simpler comparison of a pair of coefficients, testing the null hypothesis $\beta_1^{(p)} = \beta_1^{(q)}$ for quantiles, p and q respectively, can be accomplished with the Wald statistic, which has an approximate χ^2 distribution with one degree of freedom. The test can be performed on quantile regression coefficients according to Equation (2.51), with the variance estimated by Equation (2.52). Multiple comparisons are accomplished via a covariance matrix (Hao & Naiman, 2007, pp. 49-50).

$$\text{Wald Statistic} = \frac{\left(\hat{\beta}_i^{(p)} - \hat{\beta}_i^{(q)}\right)^2}{\hat{\sigma}^2_{\hat{\beta}_i^{(p)} - \hat{\beta}_i^{(q)}}} \tag{2.51}$$

$$\begin{aligned} \widehat{\beta}_i^{(q)} &= Var\left(\hat{\beta}_i^{(p)} - \hat{\beta}_i^{(q)}\right) \\ &= Var\left(\hat{\beta}_i^{(p)}\right) + Var\left(\hat{\beta}_i^{(q)}\right) - 2Cov\left(\hat{\beta}_i^{(p)}, \hat{\beta}_i^{(q)}\right) \end{aligned} \tag{2.52}$$

2.10.7 Quantile Regression Model Goodness of Fit.

R^2 is a typical measure of goodness of fit for linear regression models. For quantile regression, an analog to R^2, a pseudo-R^2 is available. Psuedo-R^2 is restricted to [0, 1], similarly as R^2. However, psuedo-R^2 is a comparison between a model with covariates and a model only containing an intercept and does not reflect the proportion of variance explained. Quantile regression minimizes a weighted sum of the difference between the data points and the quantile regression hyperplane. Quantile regression's pseudo-R^2, $R(\tau)$, is the sum of the weighted differences of the quantile regression model, $V^1(\tau)$, compared against the sum of the weighted differences of a quantile regression model containing only an intercept, $V^0(\tau)$. The sum of the weighted differences for a τ^{th} quantile regression model with dependent variable y and independent variables X in

standard design matrix form is described in Equation (2.53) (Hao & Naiman, 2007, p. 51; Koenker & Machado, 1999, p. 1297).

$$V^1(\tau) = \sum_{i|y_i \geq \beta^{(\tau)} X_i} \tau \left| y_i - \beta^{(\tau)} X_i \right| + \sum_{i|y_i < \beta^{(\tau)} X_i} (1-\tau) \left| y_i - \beta^{(\tau)} X_i \right| \tag{2.53}$$

The sum of the weighted differences for a τ^{th} quantile regression model containing only an intercept term, with quantile estimates $\hat{Q}^{(\tau)}$, is defined in Equation (2.54) (Hao & Naiman, 2007, pp. 51-52).

$$V^0(\tau) = \sum_{i|y_i \geq \hat{Q}^{(\tau)}} \tau \left| y_i - \hat{Q}^{(\tau)} \right| + \sum_{i|y_i < \hat{Q}^{(\tau)}} (1-\tau) \left| y_i - \hat{Q}^{(\tau)} \right| \tag{2.54}$$

Thus, a τ^{th} quantile regression's pseudo-R^2, R(τ), is defined in Equation (2.55) (Hao & Naiman, 2007, p. 52).

$$R(\tau) = 1 - \frac{V^1(\tau)}{V^0(\tau)} \tag{2.55}$$

R(τ) can be used to compare two nested quantile regression models, coined relative R(τ) in the literature. If $V^2(\tau)$ is more restricted than quantile regression model, $V^1(\tau)$, which possess all of the covariates denoted in $V^2(\tau)$, then relative R(τ) is given by Equation (2.56) (Hao & Naiman, 2007, p. 52).

$$R(\tau) = 1 - \frac{V^2(\tau)}{V^1(\tau)} \tag{2.56}$$

As R(τ) and relative R(τ) are functions of quantile τ, comparisons between models can be accomplished across the entire conditional distribution by sampling various quantiles and examining which model possess better performance in regions of interest (Hao & Naiman, 2007, pp. 52-54).

2.11 Chapter Summary

This chapter provided a summary overview of Social Network Analysis, definitions of measures used in relevant research, germane modeling aspects and considerations, and an overview of experimental studies investigating imperfect data in SNA. Evidenced in the contradictory nature of the findings derived from the experimental studies, the impact of imperfect data upon SNA results is not completely understood by the academic community or practitioners. Furthermore, no guidelines are available indicating when SNA is an appropriate technique for a given data set. Additionally, imperfect data mitigation techniques, such as imputation, exist but are presented without guidance on the applicability or necessity for a given problem and its associated data set.

Limitations with consensus structure aggregation and Butt's Bayesian approach, the discussed SNA methods dealing with social network model construction in the face of unreliable sources, were identified. The next chapter will present an overview of a methodology that alleviates these shortcomings. This methodology utilizes the statistical methods reviewed here to measure source agreement in reporting. Additionally, an experimental design examining factors affecting classifier performance is presented to test the methodology. Chapter V employs the statistical analysis techniques overviewed in this chapter to analyze the experimentation results.

III. Methodology Overview and Experimental Design

SNA analysts are attempting to construct a model of a true underlying social network. The social sciences have various data collection means which were overviewed in Section 2.3.2. When constructing social network models of dark or clandestine networks, information is acquired through intelligence collection means and reported. Regardless of how the data is obtained, it generally is derived from multiple information sources of indeterminate reliability. These information sources' reporting are aggregated into a social network model. This resultant aggregate social network model in all likelihood will contain correct and incorrect information, and will additionally be missing information as overviewed in Figure III-1. The problem at hand is to assess sources' reliability when constructing the social network model to discount information being provided by unreliable sources.

This chapter provides an overview of the methodology used to compare sources' reports and to subsequently assess source reliability. To examine the effectiveness of this proposed methodology, an experimental design is detailed to test the methodology's performance across a variety of conditions faced by SNA analysts in practice. Due to a lack of available real world data, generation of artificial social network information sources is described for use in the experimental testing.

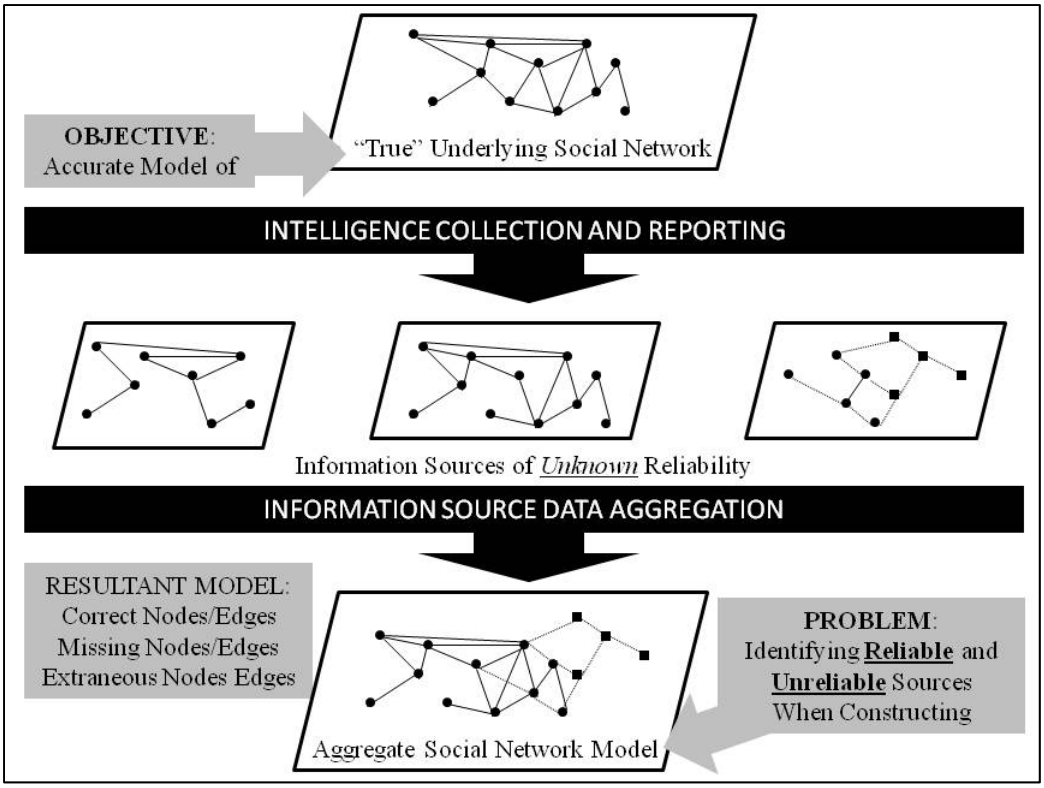

Figure III-1 Graphical Representation of Problem

3.1 Methodology Overview

This section provides an overview of the methodology developed in this research. The components of the methodology are described in general aspects, while details of the components are provided in Chapters IV and V.

3.1.1 Comparing Sources' Reporting.

Social network information sources proved reports on the existence and status of relationships among actors within the social network under investigation. These reports can prove to be conflicting or confirming with other source reporting. Lacking in the SNA literature is a systematic method of examining source reporting to ascertain whether

it is confirmed by other sources or there is dissention. Information sources may be in agreement, concordance; or they may counter other sources' reporting, discordance.

The basic data element provided by information sources is dyad reporting, the reporting of a relationship between two actors in the social network. Information sources provide statements of existing relationships between actors in the social network. Simultaneously, directly or indirectly, information sources are reporting null relationships between actors. Comparisons can be conducted among sources that are reporting on identical dyads with methods such as consensus structure aggregation as described in Section 2.5. As it is likely that sources are reporting on dyads of which only a subset are common with another information source, methods that can account for partially overlapping reporting are required.

Utilizing the reported dyads for each source and noting confirmation and dissentions, sources can be assessed for the amount of similarity. This similarity assessment can consider all of the sources as a collection, using a measure such as Fleiss' Kappa, as presented in Section 2.8.1. Conversely, the sources can be examined on a pairwise basis, with each source assessed for its similarity with every other individual source.

The approach developed in this research quantitatively assesses the sources based on the similarity of their reporting. However, it requires a measure of similarity between sources. As sources are either explicitly or implicitly reporting the presence or non-presence of a dyad, a source can be represented as a binary vector with each element representing a specific dyad as presented in Section 2.8.2. Two information sources can then be compared by examining the dyads they have in common and using a binary

similarity measure as described in Section 2.8.3. This will produce a set of pairwise source similarity scores from the social network information sources by examining confirmations and dissentions among the reported dyads and null relationships. This is represented in Figure III-2.

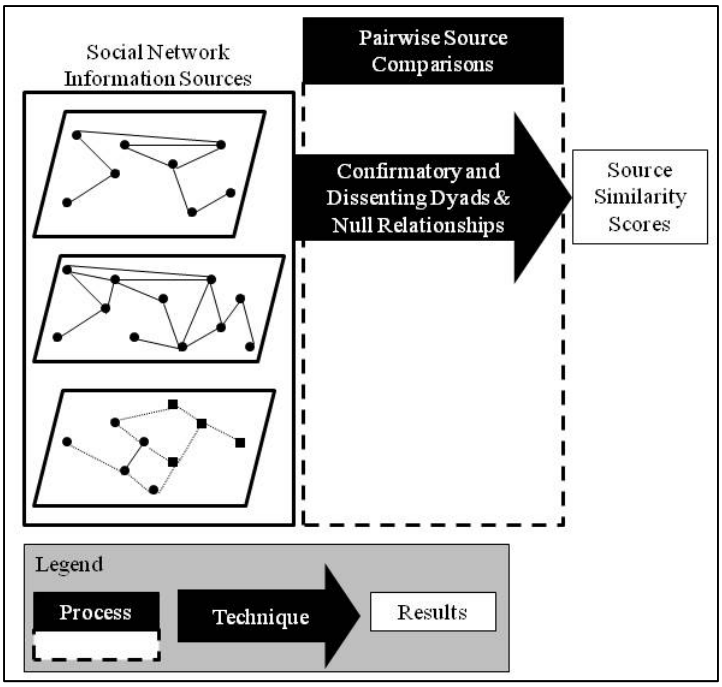

Figure III-2 Source Similarity Scores Generation

However, the problem remains of selecting an appropriate binary similarity measure from the 105 documented available measures as listed in Table A-1 in Appendix A. Chapter IV details the methodology component for selecting an appropriate binary similarity measure to use for pairwise source comparisons to produce source similarity scores.

3.1.2 Sources' Network Perspective.

Social network information sources provide reporting which reflects their knowledge of the underlying social network. Each information source, in effect, provides their representation of the social network model. These reported social network models may be artifacts of the sources' perspective of the network. If a source is a member of a social network, they may only be aware of the relationships of which they are a participant. It is also possible they may be aware of other members' relationships due to the prominence of the actors involved.

If information sources are reporting on different aspects of the underlying social network, the sources will not possess many dyads in common and thus it will be difficult to confirm or contradict dyads. Conversely, if information sources are reporting on similar aspects of the social network, the expectation is that there will be many dyads in common on which confirmation and dissentions can be determined. The number of dyads in common that sources are reporting on reflects the importance of the similarity between the sources. If two sources are commenting on the same aspect of the network, and possess high levels of dissention, that is a probable indication that at least one of the sources is unreliable. If two sources are reporting on the same aspect of the network and possess high levels of agreement, the reporting confirmation increases the likelihood that the sources are reliable.

If information sources are reporting on different aspects of the network, i.e. they do not possess many actors in common, there is no expectation that the sources' should be concordant. If two sources are reporting on different aspects of the network, they should not be penalized or rewarded in terms of their reliability assessment based on the

information obtained from that source pairing. Thus, the source pairings should be weighted according to their perspective of the underlying social network. Sources reporting on the same aspects of the social network, share a similar perspective and their similarity score should accordingly be weighted higher than sources that are providing information on different parts of the network. The methodology component of weighting source pairings by their network perspective will be discussed in detail in Section 5.1.2.

At this point, the reporting obtained from all information sources has been distilled by pairwise source comparisons into: a set of source similarity scores which measures the similarity, in terms of confirming and dissenting dyads and null relationships between each pair of sources; and a set of source pair weightings, which reflect the varying network perspectives of each of the information sources. Figure III-3 displays the methodology as described through this section.

3.1.3 Assessing Sources.

Once source similarity scores and their associated weightings have been obtained, the objective is to group information sources to begin ascertaining whether they are reliable or unreliable. Sources that report dyads and null relationships that are being confirmed are more likely to be reliable, while sources whose information is discordant with other sources' reporting are more likely to be unreliable all other factors being equal. With large, or possibly even moderate, number of sources, the majority of sources may have reported relationships that have been discredited by at least one other source. This will be reflected in the similarity scores, with sources potentially exhibiting high

similarity with some sources while simultaneously possessing low similarity scores with other sources.

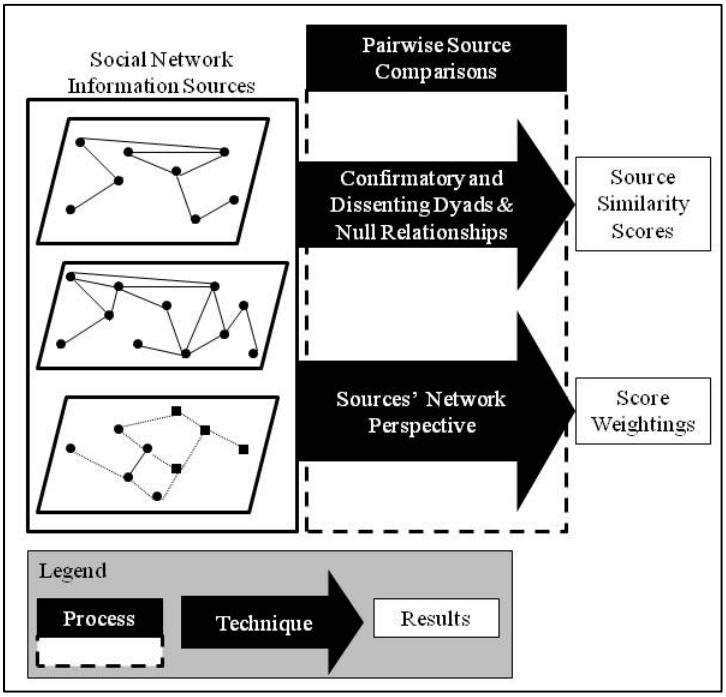

Figure III-3 Pairwise Source Comparisons Methodology

Incorporating the score weightings representing every sources' network perspective, adds further complexity. This multidimensional data composed of pairwise similarity scores and associated pairwise score weightings represents sources' concordance and the sources' individual perspective of the underlying social network. The objective at this phase of the methodology is to group sources by clustering together those who are concordant and share similar network perspectives, while separating the sources that are discordant. Information sources, providing unique information as a function of the network perspective, are unable to be assessed for reliability as there is no

relevant information to confirm or discredit their reporting. A detailed explanation of the methodology component of grouping sources into clusters is provided in Chapter V, along with analysis of the methodology's performance in the experimentation that will be discussed in Section 3.3.

Sources are clustered according to their concordance via weighted multidimensional scaling and fuzzy clustering techniques presented in described in Chapter V. The resultant cluster of sources will group sources providing confirmed information together. Sources that are providing confirmed information are presumed to have an increased likelihood of reliability than those sources whose reports have been discredited. Figure III-4 displays the methodology of conducting pairwise source comparisons to determine source concordance while accounting for the sources' network perspective. This leads to grouping the sources to determine their likelihood of reliability.

3.1.4 Methodology Practical Application.

The methodology presented in this research is an analyst aid and not an authoritative discriminator or confirmer of source reliability. The methodology utilizes only the sources' reports and does not consider *a priori* information concerning the sources or any known information regarding the underlying social network's structure. Additionally, the methodology groups sources based on confirmations and dissentions of their reporting. It is conceivable that several sources are providing information that each confirms and there is a single lone source whose information is contradictory, yet the

single source is providing the best representation of the underlying social network while the several sources are unreliable.

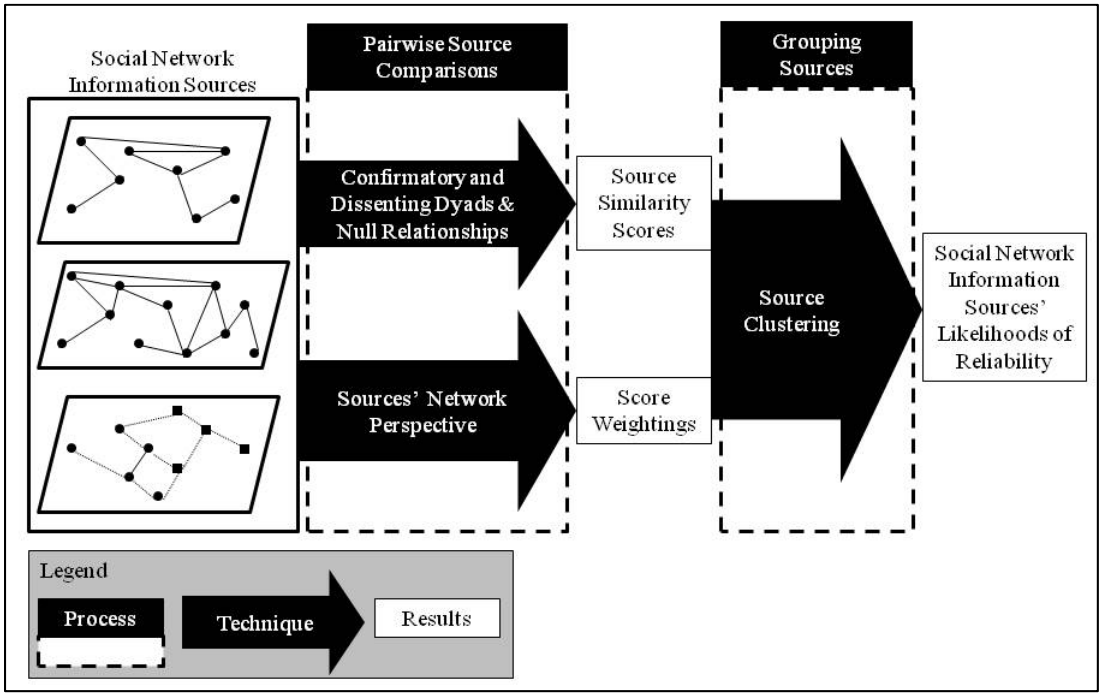

Figure III-4 Overall Methodology Framework

3.2 Experimentation Data

In order to examine the methodology's performance, experimentation needs to be conducted. Ideally, the experimentation would be conducted on a collection of real world dark network data sets, each composed of the independent information sources' reporting used to generate the model and the discarded information deemed not suitable for model inclusion. These real world dark network data sets would need to vary enough to account for differences in network structures and graph characteristics to represent the spectrum

faced by SNA analysts. Unfortunately, such a collection of real world dark network data sets does not exist in the open literature.

Searching the literature, there are no existing publically available social network data sets which are suitable for testing the effectiveness of the methodology presented here. The literature is replete with social network examples, including several dark network cases; however, the raw data used to construct these networks is not published whether due to privacy concerns, contractual agreements or the assumption of lack of relevance to a particular study. An additional aspect of this research vein is that it would also be necessary to possess the raw data that was intentionally omitted from the social network models. This inclusion of unreliable information sources precludes using currently existing social network models of real world dark networks available in the literature.

Due to the lack of suitable real world data, the testing of the methodology will be conducted on simulated networks. There are several random graph generators present in the literature with each differing in its capabilities to control various desired graph structural characteristics. The first requirement is to generate a graph to serve as the true underlying social network. Next, sources reporting true and false information are created.

3.2.1 Random Social Network Generation.

Random graph generation dates to the first algorithm by Erdös and Rényi (Erdös & Renyi, 1959). This earliest algorithm suffers from the fact that its degree distribution is not scale-free, unlike many common social networks. There are scale-free graph

generation algorithms, such as the preferential attachment model (Barabási & Albert, 1999). Although quite capable of creating scale-free random networks, the Barabási-Albert preferential attachment algorithm's utility for testing social network algorithms can be questioned for their failure to adequately mimic real world social networks in terms of clustering coefficients and degree correlation (Newman, 2002, p. 208701-2; Newman & Park, 2003, p. 036122-1).

Information regarding real world dark networks' characteristics is limited with only a few data sets' clustering coefficients and degree correlations published. Xu and Chen (2008) reported the clustering coefficients and degree correlations for three dark networks—Sageman's Global Salafi Jihad dataset, a methamphetamines trafficking network, and a network of criminals involved in gang-related crimes. The average clustering coefficients for these three networks are 0.55, 0.60, and 0.68, respectively, notably outside of the range of graphs produced by the Erdös-Rényi or Barabási-Albert algorithms. Erdös-Rényi and Barabási-Albert generated graphs possess analytic results of degree correlations of 0 in the limit as the number of nodes becomes large (Newman & Park, 2003). However, the degree correlations reported for the three dark networks are 0.41, -0.14, and 0.17, respectively (Xu & Chen, 2008), while a drug importation network researched by Morselli and Petit (2007) possessed a degree correlation of -0.47 (Keegan, Ahmed, Williams, Srivastava, & Contractor, 2010). The variability of degree correlations among the data sets reviewed here demonstrates the potential unrealistic nature of traditionally used random graph generation models to adequately represent real world dark networks.

To account for the deficiencies in creating realistic random social networks, random graph generation was accomplished via the Prescribed Node Degree, Connected Graph (PNDCG) generation algorithm (Morris, O'Neal, & Deckro, Forthcoming). The algorithm allows for varying parameter settings in graph construction to create random networks with certain desired properties. The algorithm generates connected graphs according to a desired degree distribution, while accounting for clustering and degree correlation. An advantage of the PNDCG generation algorithm is its capability to produce random graphs that possess clustering coefficients and degree correlations commonly associated with social networks. Other graph generations algorithms, such as preferential attachment method, do not consistently construct networks whose characteristics match empirical social networks.

The user employing the PNDCG generation algorithm specifies the desired graph size in terms of number of nodes, the degree distribution, and a clustering parameter. One parameter, the network size, was adjusted according to a design of experiments construct detailed in Section 3.3.2.1, while the other graph generation parameters were held static to isolate effects in the binary similarity and dissimilarity measures. All generated networks were restricted to undirected graphs. Additionally, the generated degree distribution followed a power-law (scale-free), as described in Section 2.10.1, with alpha set constant at 2.4, a value typically in the range of empirical networks. The clustering extension of the PNDCG algorithm, which encourages larger clustering coefficients, was not utilized.

3.2.2 Source Generation.

For each generated "true" network, a collection of reliable sources and unreliable sources reporting social network information were created. The reliable sources' generation procedure assumes the source reports true information, although subject to random errors. The unreliable sources are intended to report false information. The reliable and unreliable source generation techniques are designed to create sources that are reporting on the same actors in the social network, but are providing different accounts of existing and null relationships among them. Each source, both reliable and unreliable, is providing a report detailing a subset of the actors in the generated "true" social network and the existing relationships among them, with the difference between source types reflected in the accuracy of the information regarding the relationships. However, since each source is only reporting information on a subset of actors in the social network, it is possible that either a reliable or unreliable source may be the sole information provider on some of the actors in the network.

The information sources may or may not be members of the social network under investigation. The sources are reporting relationships among actors in the network. For this study and experimentation, it is immaterial whether the source is a member of the social network reporting on relationships that they directly or indirectly participate, or if the information source is external to the network, providing information they collect by observing members of the network. As discussed in Section 2.3, studies investigating social network information collection have found that participants within the social network can be unreliable when reporting information on the social network's structure.

Dark network members may take active measures to prevent accurate collection by external organizations, causing the information collected and reported to be unreliable.

The reliable and unreliable sources provide reporting in the form of an edge list. The edge list details a relationship between two nodes. Examining and aggregating the nodes composing each dyad, a list of nodes composing the network as reported by the source can be constructed. Additionally, by examining the reported nodes and noting the missing relationships on the edge list, a list of reported null relationships is obtained.

3.2.2.1 Reliable Sources Generation.

Reliable sources are generated by selecting edges uniformly at random in the true social network, with the probability specified by the design of experiments described in Section 3.3. Reliable sources then report the edges and their accompanying nodes. As a result of this construction method, a reliable source will never report a relationship occurring if it is not present in the true network, i.e. a false positive identification of a relationship. Error is introduced intentionally to reflect inaccuracies in reporting by reliable sources omitting relationships between reported nodes, i.e. a false negative report of a relationship. As they were constructed via edge selection, it is conceivable that a reliable source may report on nodes and omit an existing relationship between them.

3.2.2.2 Unreliable Sources Generation.

Unreliable sources are generated by selecting edges uniformly at random with the same probability as the reliable sources. The difference lies in that unreliable sources are sampling from a "red herring" or diversionary network, a network generated under the same parameter specifications as the true network. Since the diversionary network

possesses the same number of nodes as the true network, unreliable sources are reporting on the same actors but are drawing from a different set of relationships among them. This method enables unreliable sources to report nonexistent relationships between nodes as occurring and to omit existing relationships, false positive reports and false negative reports, respectively. It is conceivable for unreliable sources to report dyads accurately if by chance the specific dyad's status is the same in the "true" network as it is in the diversionary network.

Each unreliable source is constructed from a diversionary network exclusive to that source, though the diversionary networks are constructed according to the same parameters as the true network. Unreliable sources are independently constructed from other unreliable source as each utilizes its own diversionary network for construction as depicted in Figure III-5. Reliable and unreliable sources both utilize the same technique to sample from their respective networks. Thus, the only generation difference between a reliable and unreliable source is the underlying network from which they sample, with reliable sources sampling from the true underlying network and unreliable sources sampling from diversionary networks that are structural similar to the true network. Since social networks are sparse networks, it is expected that there will be substantial agreement on null relationships between sources, whether they are reliable or unreliable.

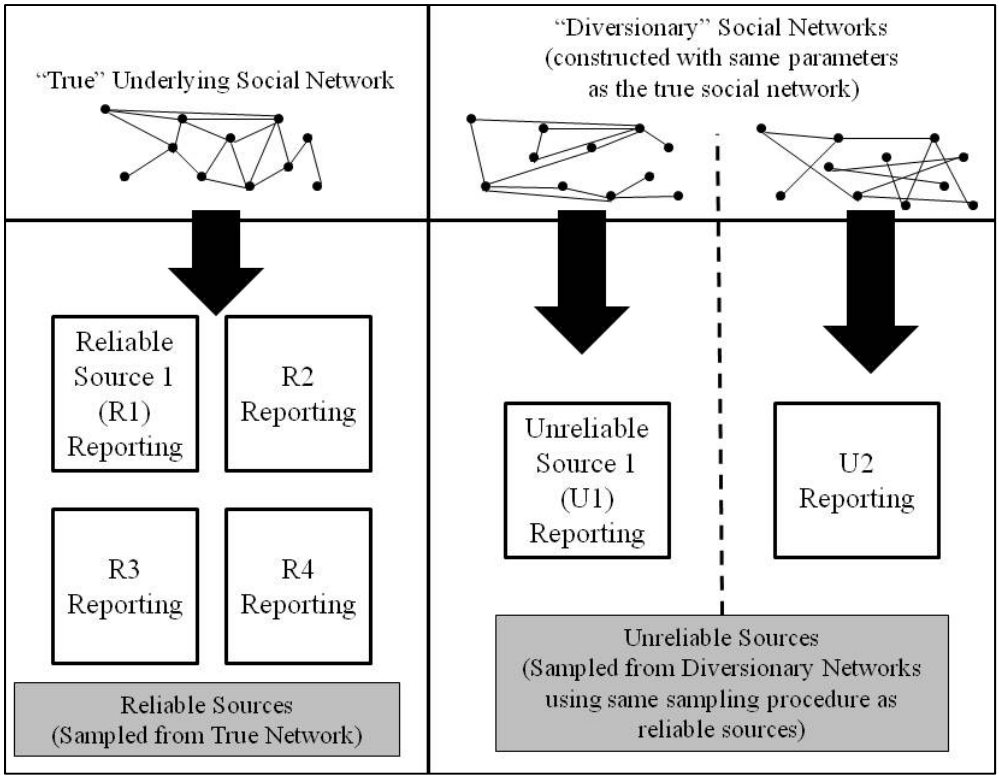

Figure III-5 Source Generation Overview

3.3 Design of Experiment (DOE)

The objective of the experimental testing is to characterize the methodology's performance in correctly discriminating between reliable and unreliable sources providing social network information. As social network analysts investigate varying types of social networks, the methodology's performance response in the face of different conditions is of interest. The majority of these conditions is outside of the analysts' realm of control, but reflects the operating environment in which the SNA is being conducted. To accurately characterize the methodology's performance, it is necessary to determine and estimate the effects of various factors on the methodology's performance.

3.3.1 Response Variable.

The objective of the methodology is to classify information sources providing social network data for incorporation into a social network model. The methodology's performance is measured on its ability to correctly distinguish between the generated reliable and unreliable sources. The fuzzy clustering in the methodology possesses a cutoff parameter for which the sources are separated into the groups. In practice, the SNA analyst specifies an appropriate cutoff parameter, perhaps as a reflection of operational risk, or whether the analysis is exploratory or being used for decision making. Accounting for these potential varying cutoff levels, the performance of the methodology can be examined at each possible cutoff value. To summarize across all cutoff values the Area Under the Curve (AUC) provides a single value detailing the Receiver Operating Characteristics (ROC) curve performance as detailed in Section 2.9.

The response variable for the experimentation is the Area Under the Curve (AUC) of the ROC graph. This summary statistic captures the overall methodology's performance over a range of parameter settings in the algorithm. In the experimentation, the AUC was computed for each selected binary similarity measures to examine their overall performance when applied in the methodology.

3.3.2 Factors.

The Design of Experiment (DOE) was developed with four explanatory factors. In practical application, these factors are unknown quantities, yet SNA analysts conduct their analysis without knowing the factors' values or effects. Ideally, the methodology is robust in its performance, regardless of the factors' values. This experimentation is

designed to determine the factors impact upon the methodology. Constructing a DOE with these factors attempts to estimate the methodology's robustness when applied against dark networks of unknown size with unknown graph structural characteristics. The four factors selected to test the methodology and the rationale for their selection follows.

3.3.2.1 Network Size.

The underlying network size was selected as a design factor to test the algorithms robustness when dealing with networks of differing sizes. The network size is measured by the number of nodes present in the true network and the corresponding diversionary networks. Real world social networks vary in size from relative small networks to in some cases, hundreds of thousands or millions of arcs and nodes if dealing with online social networks. In the case of dark networks, the size of the network is unknown and in some cases may be quite difficult to accurately estimate. One might also wish to model only a subset of the actors of the dark network dependent upon the analytical objectives.

As the focus of this methodology is to assess sources reporting on dark networks, the network size was restricted to 200 node graphs in the small case and 1,500 node graphs in the largest case. Due to the dearth of data on real world dark networks, the network sizes were selected to represent typical sized graphs faced by investigating social network analysts. Two-hundred nodes were selected as a lower bound. Although dark networks of smaller sizes certainly exist, it is unlikely that there are several sources of information on those networks whose reliability requires examination. If a small network is being investigated and there are a few information sources, graph visualization

techniques, or consensus structure aggregation as discussed in Section 2.5.1, may be sufficient to assess the sources' reporting concordance. The upper bound of 1,500 nodes was selected arbitrarily to represent a large dark network. There is nothing that restricts employing the methodology to this restricted range of networks. However, the expected performance of the methodology would be an extrapolation from the experimental design investigated here.

3.3.2.2 Number of Sources.

The next exploratory factor is the number of sources reporting information on the social network. The number of sources includes both reliable and unreliable sources. In this experimentation, the number of sources was selected to be a percentage of the network size. This percentage represents the number of sources reporting on the dark network, both reliable and unreliable. In a real world application of a dark network, these sources could be informants, undercover collectors, electronic means, walk-ins, and so forth. A lower bound of 5% and an upper bound of 10% were chosen. As there is no reported open source data detailing informant statistics and other collection means of dark networks, the percentages selected here are arbitrary. The lower bound was set at 5% because, in practice, if there are only a few reporting sources, it is possible to intensively examine each source to assess its reliability. In this experimentation, the minimum number of sources reporting on a social network is ten. Ten sources provide enough complexity that may overwhelm a social network analyst's ability to accurately assess the sources' reporting reliability. The upper bound of 10% was selected to

consider that dark networks undertake OPSEC activities which limit data collection efforts.

An advantage of expressing the number of sources as a percentage of network size is there is a large difference between ten sources reporting on a 100 node network compared against ten sources informing on a 1,000 node network. Utilizing a percentage mitigates this effect. This inherently assumes defection rates are relatively constant and investigative resources are applied against dark networks proportionally to their size. While this latter assumption may be incorrect for an especially violent or effect small group, it has been adopted for the purpose of the experimentation.

3.3.2.3 Number of Reliable Sources.

The number of reliable sources is expressed as a percentage of the number of sources. Again no real world data is available detailing the frequency of reliable sources to unreliable sources reporting social network information. In an attempt to stress the methodology, the lower bound on the percentage of reliable sources was set at 60%. This ideally aligns with real world worst case scenarios. For the upper bound, the percentage was set to 80%, which is still a substantial number of false reports on a social network. The upper bound was selected at 80% because higher percentages lose analytical interest. If all sources are reliable, or if there are only a few unreliable sources against a backdrop of many reliable sources, then the necessity of applying the methodology is diminished.

3.3.2.4 Sampling Percentage.

The sampling percentage is a proxy for the amount of data the sources are reporting. The sampling percentage is the probability of an individual edge being

reported by a source. The more information individual sources report, the more aggregate information is available. Increased amounts of information should ease assessment of reporting sources. Of note, both reliable and unreliable sources possess the same sampling percentages. Thus, the only distinguishing difference between a reliable and unreliable source is the network they are reporting on, the true underlying network for reliable sources and diversionary networks for the unreliable sources.

3.3.2.5 DOE Factors' Values.

The DOE varies the number of nodes in the "true" social network graph and by construction the diversionary network graphs, the number of reliable and unreliable sources, and the sampling probability used to generate sources by random selection of edges from either true or false networks. The total number of sources, both reliable and unreliable, was expressed as a percentage of the true social network graph's size. The number of reliable sources was expressed as a percentage of the total number of sources. By using percentages of the network's size and number of sources, two levels for each factor could be examined while producing a variety of reliable and unreliable source combinations as presented in Table III-2.

Table III-1 DOE Factors

Factors		-1	1
Network Size	A	200	1,500
# of Sources (% network size)	B	5%	10%
% Reliable Sources	C	60%	80%
Sampling %	D	10%	20%

3.3.3 DOE Justification.

The experimental design space presented in Table III-1 precludes an exhaustive search and requires a DOE approach. The network size, factor A, is restricted to integers, but complete enumeration of only this factor would require 1,301 design points to cover the specified range. The remaining factors, expressed as percentages, are continuous variables, theoretically capable of taking on an infinite number of values. In this experimentation, the number of sources and the number of reliable sources, both computed via the percentage specification, are restricted to integer solutions and could be enumerated. Although, for every specified value of factor A, all possible integer solutions for factors B and C would have to be examined. Factor D, the sampling percentage, is continuous and must be sampled at some set of levels. Complete enumeration of the design space would be prohibitive as it would take 916,661 design points to cover all possible integer combinations of the first three factors, multiplied by the number of levels examined for the continuous factor D.

3.3.4 Experimental Hypothesis.

Ideally, the methodology would demonstrate consistent performance regardless of the factors' values, i.e. the methodology is robust. However, this is unrealistic as these factors are expected to have at least a minimal impact upon the methodology's performance. What must be discovered is the effect these factors have on the methodology's performance to characterize under what conditions is it appropriate to apply the approach. It is difficult to anticipate the impact of the underlying social network's size and the number of information sources, yet these factors are present in

every practical application of SNA. The DOE will assist in statistically determining the impact of these factors on the methodology's performance.

The percentage of reliable sources and the sampling percentage are contributing factors present in every application of SNA. In this experimentation, the sources are not all assumed to be reliable sources, as in most SNA applications. The expectation is that the greater the percentage of sources that are reliable, the methodology's performance would correspondingly improve. The greater number of reliable sources providing correct reporting in comparison to unreliable sources should clearly distinguish the incorrect reporting as anomalies. The sampling percentage should also possess a similar effect as it reflects the amount of data provided by each reporting source.

Analysis of the experimentation results should estimate the effect of each factor and their interactions upon the methodology's performance. The statistical hypothesis tests whether for each of the four factors, denoted A through D, and their interaction terms, denoted AB through ABCD, their treatment effects, τ, are equal to zero as shown in Equation (3.1). Statistically characterizing the factors effects should identify the utility of the methodology under varying parameter conditions. This will allow SNA analysts to properly employ the methodology under conditions for which it is appropriate.

$$H_0: \tau_A = \tau_B = \cdots = \tau_{ABCD} = 0$$
$$H_A: \text{at least one } \tau_i \neq 0 \text{ for any factor or factor interaction } i \tag{3.1}$$

3.3.5 2^4 Full Factorial Design.

With the four independent variables identified in Section 3.3.2, it was decided to conduct a 2^4 full factorial design. A 2^4 full factorial design only encompasses 16 design points, but allows all main effects and interaction effects, including up to four factor

interactions, to be investigated. The 2^4 full factorial design points, expressed in standard order, are listed in Table III-2 (Montgomery, 2005, pp. 224-226).

Table III-2 2^4 Full Factorial Design Points

	Factors (Percentages)				Factors (Absolute Numbers)		
	Size	# Sources	# Reliable	Sample %	# Sources	# Reliable	Sample %
Run	A	B	C	D	B	C	D
(1)	200	5%	60%	10%	10	6	10%
a	1500	5%	60%	10%	75	45	10%
b	200	10%	60%	10%	20	12	10%
c	200	5%	80%	10%	10	8	10%
d	200	5%	60%	20%	10	6	20%
ab	1500	10%	60%	10%	150	90	10%
ac	1500	5%	80%	10%	75	60	10%
ad	1500	5%	60%	20%	75	45	20%
bc	200	10%	80%	10%	20	16	10%
bd	200	10%	60%	20%	20	12	20%
cd	200	5%	80%	20%	10	8	20%
abc	1500	10%	80%	10%	150	120	10%
abd	1500	10%	60%	20%	150	90	20%
acd	1500	5%	80%	20%	75	60	20%
bcd	200	10%	80%	20%	20	16	20%
abcd	1500	10%	80%	20%	150	120	20%

3.3.6 Center Point Runs.

Center point runs were added to the full factorial design to test for the existence of quadratic curvature within the design space. The center point runs do not affect the factors' effect estimates (Montgomery, 2005, p. 247). For this experiment, conducting runs at the true center points is not feasible due to the center points of some of the percentage factors generates non-integers for number of sources and number of reliable

sources. The center point design points were rounded to the nearest integer to provide integer solutions to the number of sources and number of reliable sources. This means the center point runs will affect the factors' effect estimates, albeit in a small manner, and that the full factorial design matrix augmented with the center point runs will not be truly orthogonal. Section 3.3.7 will describe another augmentation which has a similar impact in regards to the design matrix's orthogonality. Table III-3 compares the center point runs in original space where the rounding occurred and also in design space to highlight the deviation from the ideal center point located at zero for all factors.

Table III-3 Center Point Runs

Original Space				Design Space			
A	**B**	**C**	**D**	**A**	**B**	**C**	**D**
850	64	45	15	0	0.012	0.031	0

3.3.7 NOLH Space-Filling Design.

Due to the expected complexity and nonlinearities of the response surface, a space-filling design was employed to augment the original full factorial design. Specifically, conducting experiments on graphs with varying structural characteristics, does not lend one to justify assumptions of the response variable's behavior. Assumptions such as error terms' distribution properties, i.e. the traditional normally distributed errors, may not apply to graphs and the associated response variable. As such, the potential complex response surface may not be adequately investigated with a full factorial design of experiments.

A potential solution to this problem is a space-filling design. "A good space-filling design is one in which the design points are scattered throughout the experimental region with minimal unsampled regions (Cioppa & Lucas, 2007, p. 45)." One approach, Latin hypercube sampling, involves dividing an input variable's range into several strata and draw a value from each strata. The process is repeated for each of the input variables and the values for each input variable are then assigned to each run, with each variables' values appearing only once in the design matrix (Cioppa & Lucas, 2007, p. 47). Cioppa and Lucas (2007) provide an algorithm that produces nearly orthogonal Latin hypercube (NOLH) designs. While their algorithm is computational intensive, an Excel spreadsheet implementation that provides previously generated NOLH designs within specified orthogonality and space-filling parameters is available (Sanchez, 2005).

Using the Excel spreadsheet (Sanchez, 2005), the four design factors were inputted to create a NOLH design comprising 17 runs. The Excel spreadsheet assumes integer factors, which were obtained by multiplying the percentage factors by 100. The maximum absolute correlation found between the factors in the NOLH design is 0.125. The NOLH design points are displayed in Table III-4 and graphically, along with the full factorial design points, in Figure III-6. As illustrated by Figure III-6, the NOLH design points fill in the space in the settings for each factor pairing, in comparison to the full factorial design points, which bounds the region by the extreme points: $(1,1)$, $(1,-1)$, $(-1,1)$, and $(-1,-1)$. Due to rounding for some of the factors, the center point runs generated by the full factorial design do not perfectly align with the center point runs generated by the NOLH design. Both sets of center points were included in the experimentation.

Table III-4 NOLH Design Points

Original Space				Design Space			
A	**B**	**C**	**D**	**A**	**B**	**C**	**D**
606	61	46	14	-0.375	1.026	0.541	-0.20
281	17	13	16	-0.875	-0.580	0.647	0.20
363	25	15	13	-0.749	-0.245	-1.0	-0.40
444	36	24	20	-0.625	0.243	-0.333	1.0
1175	118	81	11	0.50	1.017	-0.136	-0.80
1500	105	71	18	1.0	-0.20	-0.238	0.60
1013	61	49	13	0.251	-0.591	1.033	-0.40
931	84	63	19	0.125	0.609	0.5	0.80
850	68	48	15	0.0	0.20	0.059	0.0
1094	55	35	16	0.375	-0.989	-0.636	0.20
1419	128	81	14	0.875	0.608	-0.672	-0.20
1338	107	85	18	0.751	0.199	0.944	0.60
1256	88	65	10	0.625	-0.197	0.386	-1.0
525	26	18	19	-0.50	-1.019	-0.077	0.80
200	16	12	12	-1.0	0.20	0.50	-0.60
688	62	37	17	-0.249	0.605	-1.032	0.40
769	46	30	11	-0.125	-0.607	-0.478	-0.80

3.3.1 Replications.

Replications were conducted at each design point in order to obtain an estimate of experimental error (Montgomery, 2005, p. 13). "Replication reflects sources of variability between runs and (potentially) within runs (Montgomery, 2005, p. 14)." For this experimentation, there is no available data in the literature indicating the magnitude of the experimental error to be expected.

Operating characteristic curves can be used to compute the number of replications, n, necessary to specify a level of acceptable type II error (Montgomery, 2005, pp. 177-178). The experiment was designed to detect a 0.05 difference in the performance means between treatment main effects, D. The standard deviation, σ, was

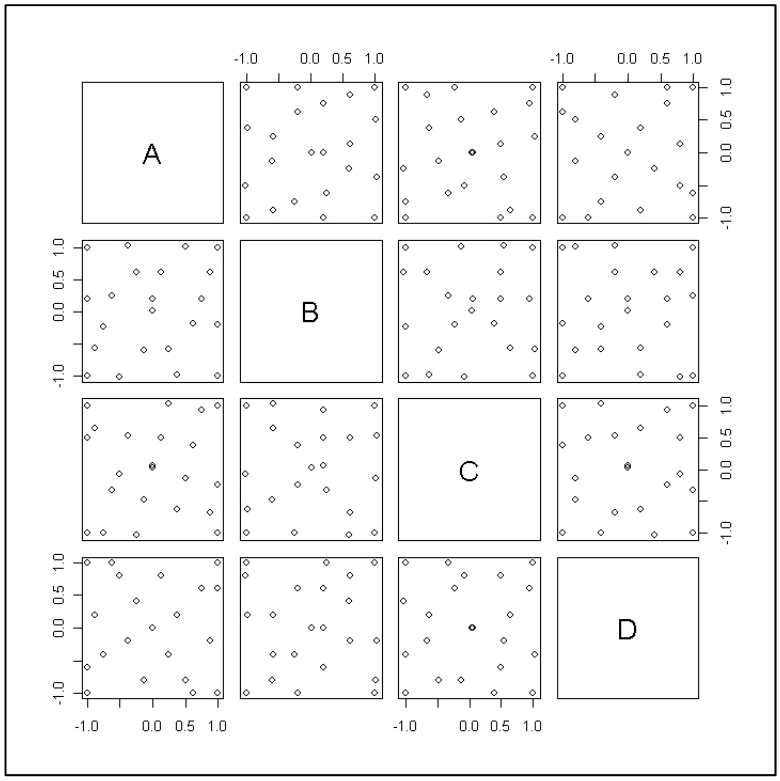

Figure III-6 Experimental Design Points (in Design Space)

estimated to be 0.025. The degrees of freedom of any main effect in the model is the number of levels, a, minus one. The other three factors can be combined into a single factor with 8 levels, b, and its associated degrees of freedom is $b - 1$. Using the Φ^2 parameter computed as in Equation (3.2), Table III-5 displays the associated β and the power of the test, $(1 - \beta)$, for several values of n (Montgomery, 2005, pp. 177-178). As displayed in Table III-5, five replications are necessary to have a power greater than 0.80, a common statistical threshold. Ten replications were utilized in the experimentation to compensate for inaccuracies in the standard deviation estimate.

$$\Phi^2 = \frac{naD^2}{2b\sigma^2} \tag{3.2}$$

Table III-5 Replications Calculations

n	Φ^2	Φ	Numerator DOF	Error DOF	β	(1-β)
2	2	1.41	1	16	0.60	0.40
3	3	1.73	1	32	0.32	0.68
4	4	2.00	1	48	0.21	0.79
5	5	2.24	1	64	0.17	0.83
6	6	2.45	1	80	0.09	0.91
7	7	2.65	1	96	0.055	0.945
8	8	2.83	1	112	0.022	0.978
9	9	3.00	1	128	0.014	0.986
10	10	3.16	1	144	0.011	0.989

As each replication involves a new set of "true" and "false" networks to generate sources, it is expected the replications will reflect the graph to graph variation at each combination of factor settings. Additional statistical approaches, such as ANCOVA, to account for the expected graph to graph variation are discussed in Section 2.10.1.

3.4 Statistical Analysis of the DOE

Conducting the experiment under a DOE framework allows for efficient statistical analysis of the factors and their impacts upon the response variables. The methodology's performance was measure by calculating the Area Under the Curve (AUC) as a single measurement of the algorithm's classification accuracy. Collapsing all source classifications into a single dimension allows each set of sources, both reliable and unreliable, describing a true social network to be considered a single case or run within

the experimental design. However, mathematical properties of the AUC require consideration before the analysis can begin.

3.4.1 Transforming the Response Variable.

The AUC ranges from its minimum value of 0.5, representing poor classification accuracy, to its maximum value of 1.0, indicating perfect identification of sources as either reliable or unreliable. However, traditional DOE utilizes linear regression for its analysis. In this instance, the dependent variable, AUC, is continuous but bounded on [0.5, 1.0]. Examining Figure III-6 shows that many observed values lie close to a bound. This may lead to deriving a regression equation whose predictions regularly lie outside the bound (Montgomery, Peck, & Vining, 2006, p. 429). To correct for this, response variable transformations are required.

Several response variable transformations are in common practice to account for bounded variables. Conducting regression with a proportion as the response variable has been addressed in the statistics literature (Bottai, Cai, & McKeown, 2010, p. 310). As a proportion is restricted to range between zero and one, transformations are applied to enable regression analysis. A commonly used transformation, the logit transformation, takes variables mapped on (0, 1) and using logarithms, transforms them to the $(-\infty, \infty)$ domain. Thus to take advantage of this commonly applied transform, the AUC must be mapped from its range of [0.5, 1.0] to (0, 1).

The initial step is to transform a response variable, y, from $[a, b]$ to [0.0, 1.0]. This can be accomplished by using Equation (3.3) (Smithson & Verkuilen, 2006, p. 54).

$$\frac{(y-a)}{(b-a)} \tag{3.3}$$

Transforming a response variable, *y*, from [0.0, 1.0] to (0.0, 1.0), so that the extreme values are no longer possible, can be accomplished by Equation (3.4), where *n* is the sample size (Smithson & Verkuilen, 2006, p. 55). In this application, the sample size is adjusted to account only the cases in which the AUC could be computed. This necessity will become evident when the logit transform is presented as the natural logarithm of zero is negative infinity.

$$\frac{y(n-1)+0.5}{n} \tag{3.4}$$

Finally, the dependent variable, *y*, whose domain is (0, 1) can be mapped to the continuous real line, (-∞, ∞), by using the logit transform, commonly used in logistic regression, as described in Equation (3.5) (Hosmer & Lemeshow, 2000, p. 6).

$$\ln\left[\frac{y}{1-y}\right] \tag{3.5}$$

Utilizing these three transforms, the AUC is transformed from its initial [0.5, 1.0] range to (-∞, ∞) on which regression can be applied.

3.5 Experimentation Implementation

The underlying social networks are generated using C code implementing the PNDCG algorithm. The sources are generated using Java code that extends the Java Universal Network/Graph Framework (JUNG) library, version 2.0.1 (O'Madadhain, Fisher, Nelson, White, & Boey, 2010). The JUNG library was also the basis for computing the SNA network measures used in Section 5.7.2 in conjunction with additional Java code. The source comparisons utilizing the binary similarity/dissimilarity

measures are conducted through Java code. Some of the main components of the Java code are available in Appendix F and Appendix G.

The methodology components of weighted multidimensional scaling, fuzzy clustering, and ROC curves were implemented in R, x64 version 2.13.0 (R Development Core Team, 2011). The weighted MDS was accomplished via the smacof package (de Leeuw & Mair, 2009), the fuzzy clustering was conducted using the cluster package (Maechler, Rousseeuw, & Struy, 2005), and the ROC curves were computed using the ROCR package (Sing, Sander, Beerenwinkel, & Lengauer, 2009). The statistical analysis of Chapter V was conducted in R and the quantile regression of Section 5.8 was performed via the quantreg package (Koenker, 2011). The R script for executing the methodology is available in Appendix E.

The experimentation was conducted on a desktop computer containing a 2.70 GHz AMD Athlon™ II X2 215 processor. The desktop was equipped with 4.00 GB of RAM. The operating system was Windows 7 x86 Enterprise edition, 64 bit version. The Java code was composed and compiled in Netbeans IDE 6.9.1 (Oracle Corporation, 2010).

3.6 Chapter Summary

This chapter outlined the overall methodology and detailed the experimental testing to examine the methodology's performance. The experimentation is conducted under a DOE framework, with the factors, design, replications, and necessary data transformations for analysis provided in this chapter. Additionally, the need for artificially social network data was justified and a method for generating reliable and

unreliable social network information sources was described. Chapter IV discusses and conducts the methodology for selecting binary similarity measures to conduct source reporting comparisons. Utilizing the binary similarity measures identified in Chapter IV, Chapter V will describe the methodology for source assessment, present a detailed example of employing the methodology, and then analyze the results from the experimentation detailed in this chapter.

IV. Pairwise Source Concordance Measure Selection

This chapter details the methodology for measuring concordance among the social network information sources. It begins with a statistical analysis utilizing Fleiss' Kappa, a standard nonparametric statistic for inter-rater reliability. Following the Fleiss' Kappa analysis is a description of the methodology to select a measure to conduct pairwise source comparisons. Finally, a detailed example of selecting appropriate binary similarity measures to compare sources is presented.

4.1 Inter-Rater Reliability with Fleiss' Kappa

Given a collection of information sources reporting social network data, one has a listing of dyads with various sources classifying them as present or absent. The objective is to determine how consistently information sources classify the dyads. This is analogous to assessing inter-rater reliability, for which a stastistical technique, Fleiss' Kappa, is commonly applied. Fleiss' Kappa measures inter-rater reliability across all information sources as discussed in Section 2.8.1. Fleiss' Kappa does not identify reliable or unreliable sources, but merely assesses the collection as a whole. One would assume the greater the percentage of reliable sources in a collection of sources, the greater the Fleiss' Kappa score. If true, this would imply that Fleiss' Kappa could be used as an indicator of whether unreliable sources are present in a collection of sources. Unfortunately, the results obtained from the experimentation conducted here do not support such a conclusion.

Fleiss' Kappa was computed for each experimental run and generally showed weak to moderate concordance among the reporting sources, which is expected due to the

purposeful inclusion of unreliable information sources. These are displayed in Table IV-1 which averages the ten replications for each design run. The overall average for Fleiss' Kappa is 0.232.

The Pearson product moment correlation between Fleiss' Kappa and the percentage of reliable sources in the sample was calculated and found to be 0.375. This exhibits a weak correlation to the expectation that with an increasing percentage of reliable sources in the sample, the sources' concordance, measured by Fleiss' Kappa, would improve. Of interest, Fleiss' Kappa's correlation with the total number of sources, both reliable and unreliable, is 0.705. This may explain Fleiss' Kappa's weak correlation with the percentage of reliable sources as the number of sources is a more important contributor to the Fleiss' Kappa scoring. Unfortunately, this experimentation demonstrates that Fleiss' Kappa is unsuitable as a measure of concordance of social network information sources. The alternative was to investigate pairwise similarity between sources.

Table IV-1 Fleiss' Kappa Averaged by Experimental Run

Run	Average Fleiss' Kappa	% Reliable Sources	Number of Sources
(1)	0.028	60%	10
a	0.188	60%	75
b	0.071	60%	20
c	0.031	80%	10
d	0.028	60%	10
ab	0.248	60%	150
ac	0.307	80%	75
ad	0.232	60%	75
bc	0.153	80%	20
bd	0.135	60%	20
cd	0.075	80%	10
abc	0.391	60%	150
abd	0.282	60%	150
acd	0.405	80%	75
bcd	0.249	80%	20
abcd	0.451	80%	150
centerpt	0.300	70%	64
SF1	0.279	75.4%	61
SF2	0.171	76.5%	17
SF3	0.108	60%	25
SF4	0.207	66.7%	36
SF5	0.313	68.6%	118
SF6	0.302	67.6%	105
SF7	0.337	80.3%	61
SF8	0.356	75%	84
SF9	0.288	70.6%	68
SF10	0.238	63.6%	55
SF11	0.262	63.3%	128
SF12	0.409	79.4%	107
SF13	0.324	73.9%	88
SF14	0.202	69.2%	26
SF15	0.116	75%	16
SF16	0.220	59.7%	62
SF17	0.184	65.2%	46

4.2 Source Comparison Methodology

The methodology presented here proceeds with conducting comparisons among the sources in a pairwise fashion. These pairwise source comparisons are accomplished by examining the concordance of the sources reporting whether dyads are present or absent, for the dyads they have in common. On a dyad by dyad basis, it is noted if the sources confirm the existence, concur on the absence or if disagreement exists. This allows construction of the confusion matrix described in Section 2.8 and the subsequent application of a selected binary similarity measures. This section begins with a discussion of a methodology to select one or a set of suitable binary similarity measures. The objective of this phase of the methodology is to generate source similarity scores, as shown in Figure IV-1, which are derived from the selected binary similarity measure.

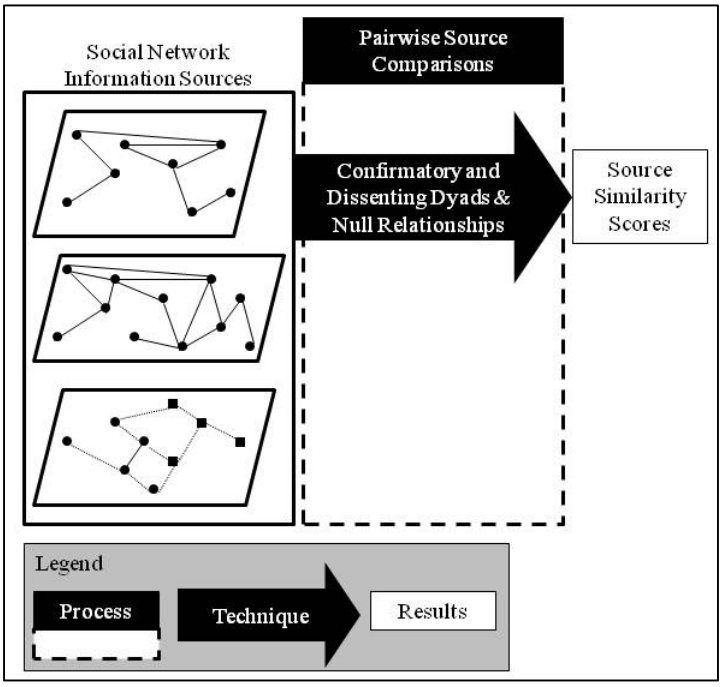

Figure IV-1 Source Similarity Scores Generation

4.2.1 Empirical Justification of Source Reporting Comparisons.

Romney and Weller (1984, p. 63) defined an informant's reliability as "the correlation between the recall data of an individual and the total aggregated recall data of the group (minus the individual's own data)." Using the four social network data sets explored in Bernard's, Killworth's, and Sailer's (1979/1980) study, described in Section 2.3.2.2, and measuring accuracy by examining information reported by informant participants in the social network as compared against data derived from independent observations, they determined "the more reliable an individual is the more accurate he or she is (Romney & Weller, 1984, p. 66)" and further hypothesized that "individuals or informants can be weighted by their reliability, i.e. the answers of 'better' informants would be taken more seriously or weighted more than the answers of the less reliable informants (Romney & Weller, 1984, p. 76)." As they defined reliability in terms of a source's agreement with the consensus obtained from the complete collection of information, their results lend credence to the approach of assessing individual sources via comparisons to the collection of sources.

4.2.2 Confusion Matrices Creation.

Sources reporting on the same social network were compared pairwise and the results tabulated in a series of confusion matrices. The procedure for source comparison was as follows. All dyads reported by both sources were examined. If both sources agreed upon the presence of the dyad, a positive match was recorded in cell *a* of the confusion matrix. If both sources agreed upon the absence of the dyad a negative match

was recorded in cell *d* in the confusion matrix. If one source confirmed a relationship and the other denied its presence, the dispute was recorded in cells *b* or *c* as appropriate.

4.2.3 Selection of Source Comparison Measures.

Section 2.8.3 introduced 105 binary similarity and dissimilarity measures found in the literature, which are listed in Table A-1 and Table A-2 in Appendix A. These 105 binary similarity and dissimilarity measures can be reduced to 96, as several measures were algebraically shown to be perfectly correlated in Table II-15. As even this reduced number would prove to be unwieldy, a methodology, presented here, was developed to select a reduced set of measures. Selecting appropriate binary similarity or dissimilarity measures can be accomplished by examining several desirable characteristics. The potential measures' characteristics are derived empirically on data sets relevant to the application. One characteristic is the computability of the various binary measures. Another is the correlation among the measures. Measures that are highly correlated indicate a redundancy and allow for reductions in the total number of binary similarity and dissimilarity measures required to adequately describe the data set.

Selection of appropriate binary similarity and/or dissimilarity measures depends upon the particular application under investigation. Since the measures were introduced as theoretically measuring different aspects of similarity, it is likely that results can be quite divergent when applied to specific data sets. For example, the discussion on the importance of negative matches highlights that a data set possessing a large number of negative matches will present different values across the collected binary similarity and dissimilarity measures as compared against a data set with minimal negative matches.

Ideally, several data sets would be available to generate a variety of confusion matrices that are possible for the specific application under consideration.

4.2.3.1 Testing Measures' Computability.

Dependent upon characteristics of the confusion matrix, some binary similarity and dissimilarity measures cannot be computed. This results from issues such as square roots of negative numbers or zero appearing in the denominators of equations. These circumstances can occur when two cells of the confusion matrix possess zeros, particularly if one of them is cell *a*. If some of the cells of the confusion matrix are significantly larger in terms of magnitude than the remaining cells, negative numbers can occur, which cause some measures utilizing square roots to become incomputable. For example, two of the measures, *Gilbert & Wells* and *Stiles*, utilize logarithms, but occasionally proved incomputable as they attempted to take the logarithm of zero. Note, that computability of measures is highly dependent upon the data set being tested as certain characteristics of confusion matrices drive computability. Generalizing to other data sets should therefore be considered cautiously.

4.2.3.2 Group the Measures.

Binary similarity and dissimilarity measures that are strongly positively or negatively correlated are providing redundant information. As the confusion matrices are composed of only four numbers, the magnitude of these numbers in relation to each other substantially impacts the binary similarity measures' values. Dependent upon the specific application and the associated confusion matrices' characteristics, binary similarity measures' correlation among themselves may vary.

As a result, the correlation profiles of the measures may be dependent on the application. However, if confusion matrices exist from previously obtained application data, or can be simulated with proper confusion matrix characteristics, the correlation profile of the measures can be empirically constructed. The correlations among the binary similarity measures can be summarized in a correlation matrix.

4.2.3.3 Identify Similarity Measures Clusters.

Clusters of measures can be observed in the correlation matrix, by identifying groups of measures which possess strong positive or negative correlations. There are numerous clustering techniques to identify these groupings. In this research, Multi-Dimensional Scaling (MDS) is performed on the correlation matrix. Groupings of binary similarity measures are then identified via visual inspection. In some instances, measures can be identified as isolates, i.e. not associated with any cluster.

4.2.3.4 Determine Representative Measures.

For each cluster obtained from the MDS, a representative binary similarity measure may be selected. As each cluster involves strongly correlated measures, a representative measure ideally exhibits the same dimensions of similarity expressed by the other measures composing the cluster. In this research, for every identified cluster, MDS was executed for only the measures comprising that cluster. A binary similarity measure was selected based on its positioning in the MDS visualization, with the intention of identifying a measure in the center of the visualization as the best representative of that cluster.

4.2.3.5 Select Desired Measures.

At this point, the complete listing of candidate binary similarity measures has been reduced to representative and isolate measures. Working with a reduced set, the analyst can further reduce the number of measures by examining specific characteristics of each measure, such as the measures' ranges. Representative measures that are deemed unsuitable can be replaced by other measures from their respective clusters if advantageous. Criteria specific to the application can be used to reduce the number of selected measures to a reasonable size, or potentially increase the number of measures if greater diversity is required. The binary measure selection methodology overview is provided in Figure IV-2 and the experimental implementation is detailed in Section 4.3.

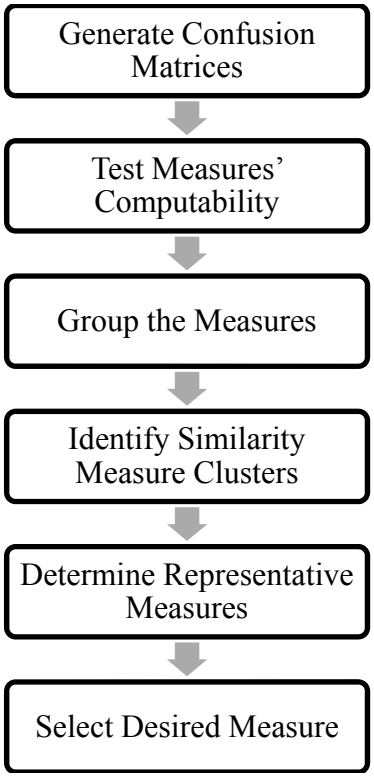

Figure IV-2 Source Comparison Binary Measure Selection Process

4.3 Pairwise Similarity Measure Selection

To generate confusion matrices to select an appropriate similarity measure following the procedures outlined in Section 4.2.3, experimental data was utilized. Reliable and unreliable sources were generated according to the full factorial design runs as described in Section 3.3.3 and depicted in Table III-2. Confusion matrices were then generated for every pairing of sources from each design point and the ten replications corresponding to the design points. For each factor combination run, pairwise comparisons generated a grand total of 56,690 confusion matrices. The 96 distinct binary similarity measures identified in Section 4.2.3 were then calculated for each of the test confusion matrices.

4.3.1 Testing the Computability.

For every one of the 56,690 generated confusion matrices, the proportion of matrices that each measure could be computed was found. Figure IV-3 displays a histogram of the number of measures for each fraction bin, which shows that the majority of measures are computable on at least 80% of the confusion matrices, with the exception of *Batagelj & Bren*, *Gilbert & Wells*, and *Pearson*-III. The reduction in binary similarity and dissimilarity measures under consideration by removing those which could be computed in less than 80% of the cases left 93 measures. These 93 measures are contenders for selection due to their appropriateness to the data set representative of this specific application. A different data set could potentially lead to substantial differences in the measures' computability distribution. Of note for this particular data set, cell *d*, representing agreement between information sources on null relationships, is significantly

larger than the elements in the other cells of the confusion matrices. This results from social networks being sparse networks, consisting of relatively few edges in the graph. This peculiarity of the source reliability application is the driving factor for some measures possessing surprisingly low computability percentages.

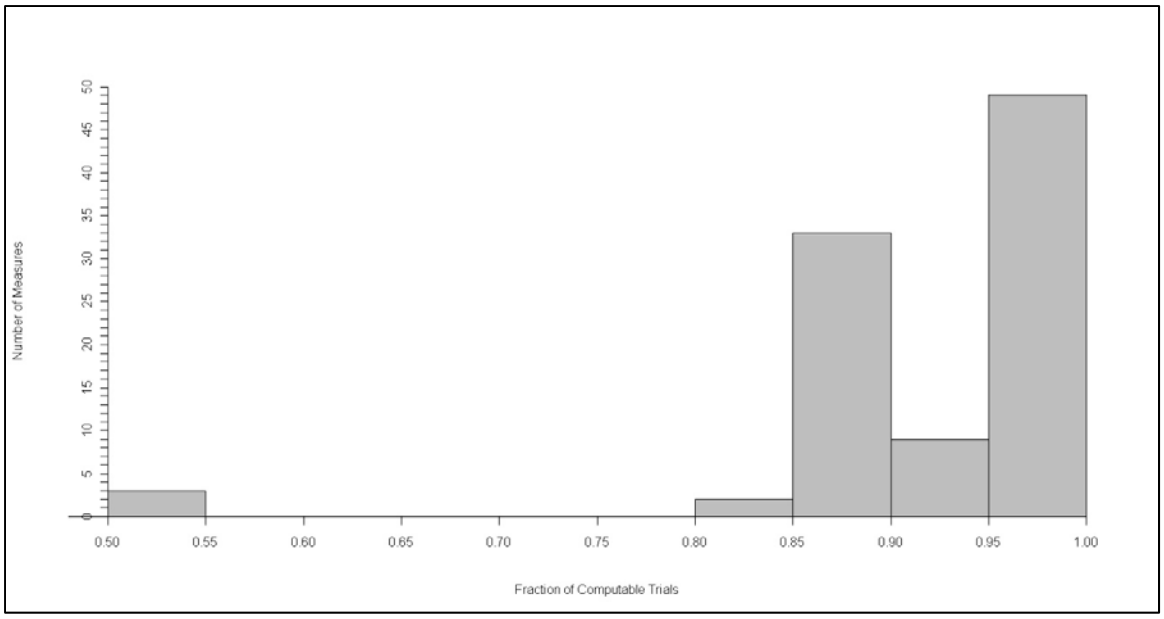

Figure IV-3 Measures' Computability Percentages

4.3.1.1 Measures' Computation Times.

The binary similarity measures' equations involve at most four variables: elements *a*, *b*, *c*, and *d* from the confusion matrix. Despite the varying complexity of the equations in Table A-1 and Table A-2 in Appendix A, substantial differences in measures' computation times are not exhibited. In this experimentation, generating the confusion matrices proved to be more computationally intensive than calculating the complete collection of binary similarity measures.

4.3.2 Discovering Similar Measures Groupings.

The correlation among the remaining measures was investigated to discern whether further reductions were possible. Figure IV-4, produced in the software package R (2011) via the corrplot package (Wei, 2011), displays a graphic version of the pairwise correlation matrix. The pairwise correlation is computed via Pearson's product-moment between the two measures values for every confusion matrix where both measures could be calculated. Strong positive or negative correlations among measures indicate a redundancy, as strongly correlated measures are capturing the same similarity/dissimilarity dimensions within the data set. As Figure IV-4 illustrates ordering the measures via hierarchical clustering, there is high correlation among several groupings of measures.

4.3.2.1 Identify Groupings Among Binary Measures.

Various algorithms exist to cluster the correlation matrix to identify groupings of measures. An alternative approach taken here utilizes Multidimensional Scaling (MDS). The pairwise correlation matrix was converted to a distance matrix by applying the transformation one minus the absolute value of each element's correlation, as suggested by Choi (2008, p. 70). This places measures with strong positive or negative correlations close to each other in the MDS visualization, while increasing the distance of measures with low correlations. The MDS visualization, generated via R's corrmds function (2011) depicted in Figure IV-5 with an aspect ratio of one, leads to visually aggregating the measures into five main groupings and identification of three isolated measures.

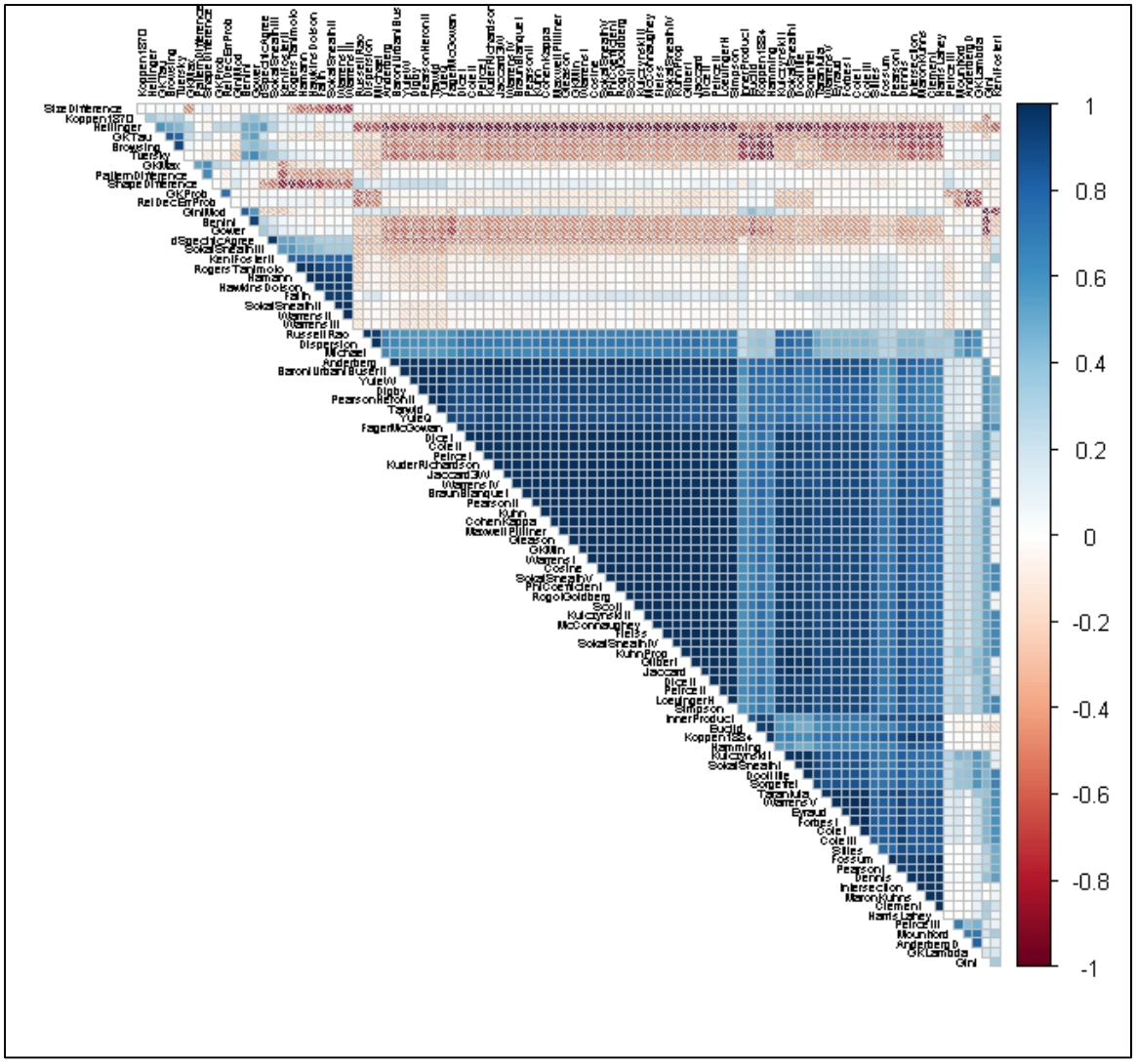

Figure IV-4 Correlations Among the Reduced Set of Measures

To find a single representative binary measure within each grouping, MDS was again applied to only the members of the designated group. Reapplying MDS sequentially with only the group members' measures is necessary to avoid MDS positioning due to influence from measures in other groups. MDS utilizes all distance pairings to place points. Conducting MDS in a sequential manner using only a subset of the data, i.e. the group's members, identifies a central representative for each group.

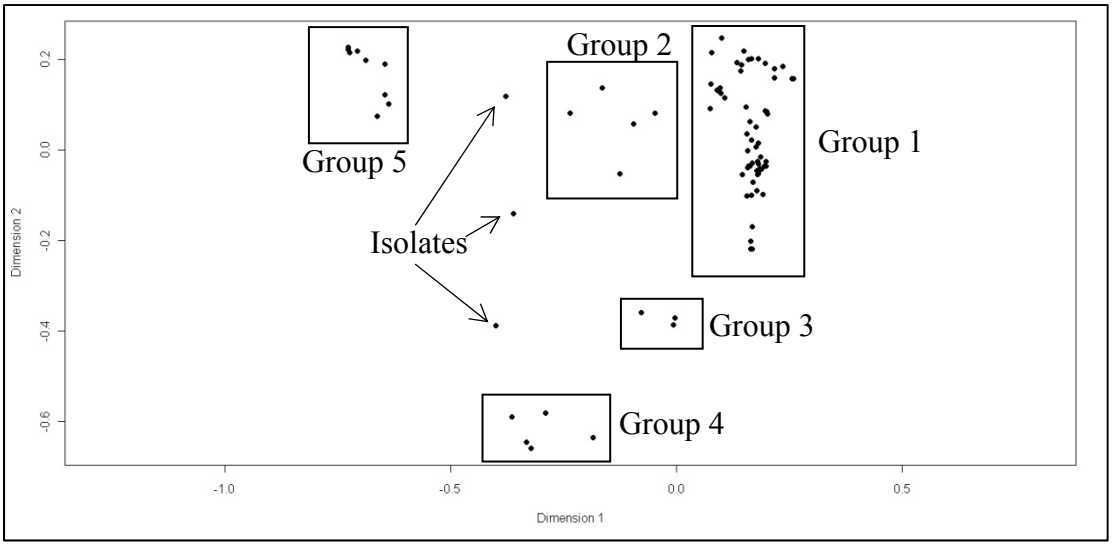

Figure IV-5 MDS of Reduced Set Measures with Groupings

Specific measures centrally located by the subsequent MDS were selected as representatives. Figure IV-6, Figure IV-7, Figure IV-8, and Figure IV-9 display the MDS visualizations for groups 1, 2, 4 and 5, respectively; each with an aspect ratio of 1. As Group Three is only composed of three measures, the subsequent MDS visualization is uninteresting, merely placing *Dispersion* in between *Michael* and *Russell & Rao*. Examining the MDS of the 66 measures composing Group One listed in Table IV-2 and shown in Figure IV-6, *Cohen's Kappa*'s central location, identified visually as being close to the origin, and its greater prevalent usage, led to its selection to represent the measures clustered in Group One listed in Table IV-2.

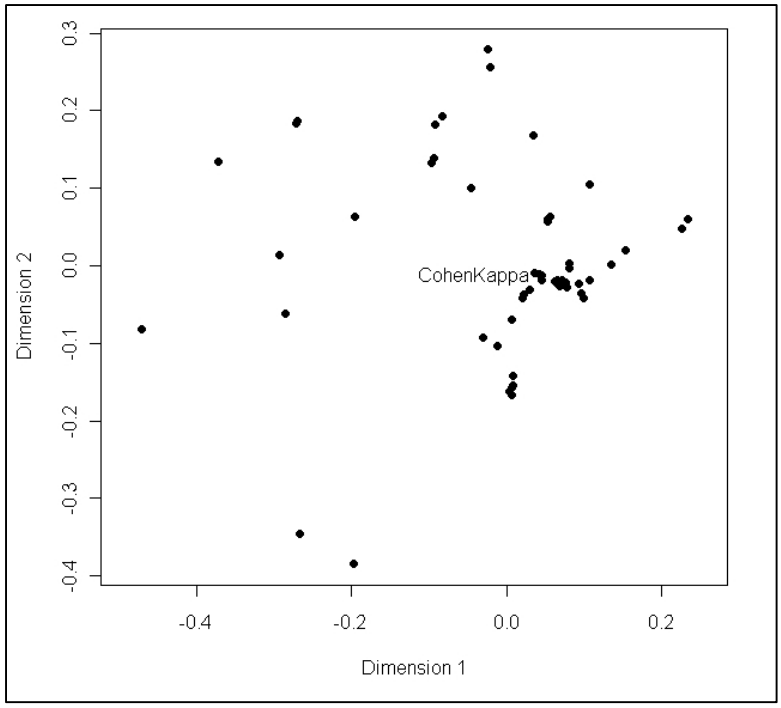

Figure IV-6 MDS of Group 1

Table IV-2 Group 1 Measures

Anderberg	Gilbert	Pearson-I
Baroni-Urbani & Buser-II	Goodman & Kruskal Min	Pearson-II
Benini	Goodman & Kruskal Tau	Peirce-I
Braun-Blanquet	Gleason	Peirce-II
Browsing	Gower	Phi Coefficient
Clement	Hamming	Scott
Cohen's Kappa	Harris & Lahey	Simpson
Cole-I	Hellinger	Sokal & Sneath-I
Cole-II	Inner Product	Sokal & Sneath-IV
Cole-III	Intersection	Sokal & Sneath-V
Cosine	Jaccard	Sorgefrei
Dennis	Jaccard-3W	Stiles
Dice-I	Koppen 1884	Tarantula
Dice-II	Kuder & Richardson	Tarwid
Digby	Kuhn	Tversky
Doolittle	Kuhn Proportion	Warrens-I
Euclidean	Kulczyński-I	Warrens-IV
Eyraud	Kulczyński-II	Warrens-V
Fager & McGowan	Loevinger's H	Yule Q
Fleiss	Maxwell & Pilliner	Yule W
Forbes-I	McConnaughey	
Fossum	Pearson & Heron-II	

Gini was selected to represent the five measures of Group Two, listed and displayed in Figure IV-7, due to its central location in the MDS. *Gini* is the measure closest to the origin, which is the centroid of the group, computed by taking the average of all of the datapoints.

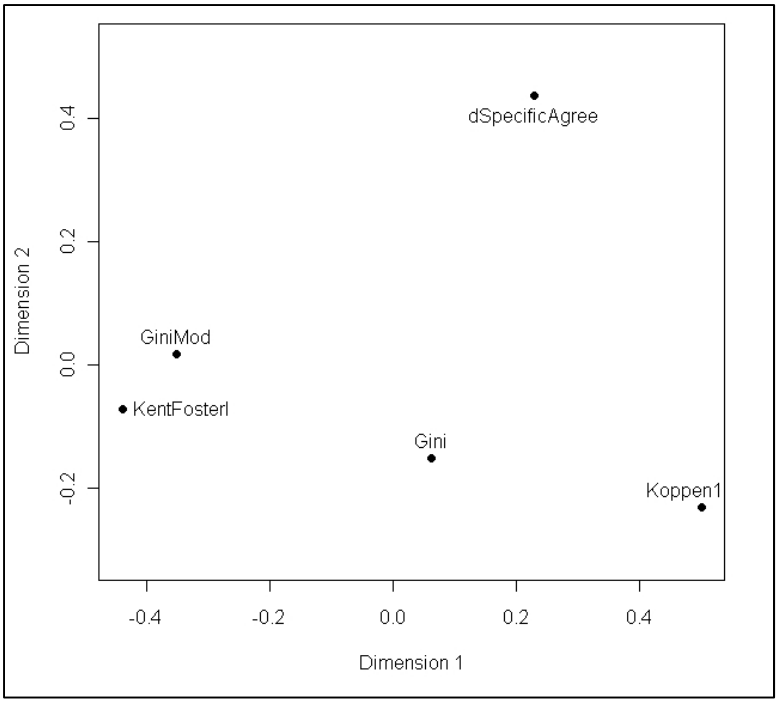

Figure IV-7 MDS of Group 2

Anderberg's D similarity measure was selected to represent the five measures of Group Four displayed in Figure IV-8, due to its central placement in the MDS.

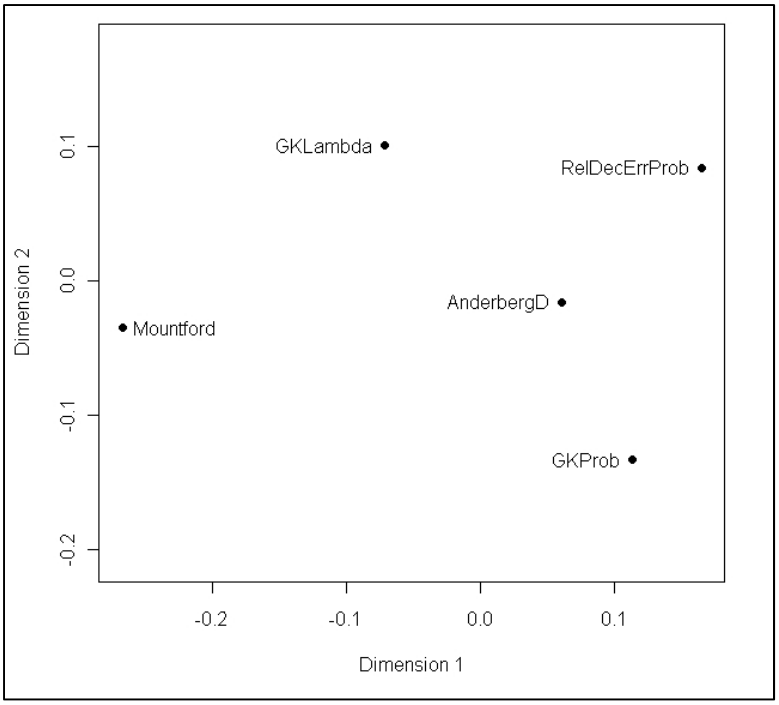

Figure IV-8 MDS of Group 4

Group Five's MDS, depicted in Figure IV-9, contains the placement of eleven measures, of which, *Hamann* was chosen to represent Group Five, due to its proximity to the centroid of the group and grerater familiarity than *Hawkins & Dotson*.

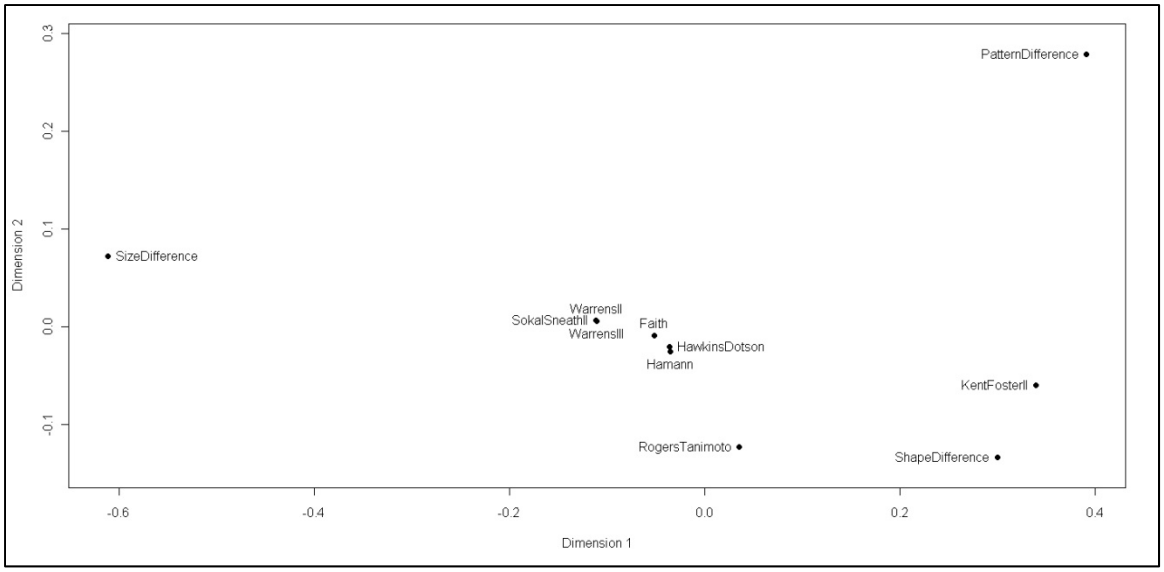

Figure IV-9 MDS of Group 5

Excluded from any of the five groupings, three isolated measures remain: *Goodman & Kruskal's Maximum Formula*, *Peirce*-III, and *Sokal & Sneath*-III. These measures appear to capture an aspect of similarity not represented by any of the other binary measures investigated here and so they remained as viable candidate measures for testing.

4.3.3 Selected Measures.

Initially facing 105 binary and dissimilarity measures, the methodology presented here reduced the set to eight measures: *Cohen's Kappa*, *Gini*, *Dispersion*, *Anderberg's D*, *Hamann*, *Goodman & Kruskal's Maximum Formula*, *Peirce*-III, and *Sokal & Sneath*-III. These measures were selected due to an empirical investigation of relevant data sets for this particular application. First, the computability of measures was investigated, garnering a set of measures that are suitable for the application. Correlation comparisons allowed for further reductions in the measures under consideration by removing

extraneous highly correlated measures. In this experiment, the correlation comparisons led to *Cohen's Kappa*, *Gini*, *Dispersion*, *Anderberg's D*, and *Hamann* similarity measures representing 90 similarity measures by examining the groupings resulting from multidimensional scaling. Most importantly is the lack of observed correlation among the eight selected similarity measures. Apparently, these eight measures address different dimensions of similarity for this specific data set and thus all should be utilized if possible to adequately describe this application's data set.

Additional considerations can be incorporated to account for application specificities. For the purpose of social network source assessment, the binary similarity measure between two sources is converted into a dissimilarity measure. This is accomplished by simply subtracting the measures' scores from the measures' theoretical maximum scores. For the set of eight measures selected in this experimentation, *Sokal & Sneath*-III is not bounded and therefore cannot be converted into a dissimilarity measure. As *Sokal & Sneath*-III was an isolated measure as opposed to representing a grouping of measures, it can be eliminated from consideration. If it had been a representative measure, another measure from the grouping could have been selected as a replacement measure.

4.4 Chapter Summary

This chapter provided details of the developed methodology with an example of its application to determine suitable binary similarity measures. The example was conducted on simulated data obtained in accordance with the data generation techniques and experimental design as described in Sections 3.2 and 3.3, respectively. The

methodology presented in this chapter can derive different sets of binary similarity measures dependent upon the specific application. Based on the experimentation conducted here, the methodology was able to reduce the initial 105 binary similarity measures candidates to seven suitable measures for the information source comparison application. The methodology also demonstrated its flexibility in selection of representative measures to account for application specific considerations. As such, it can be applied for other applications outside the thrust of this dissertation.

With a set of source comparison measures selected, Chapter V proceeds to examine that set to determine a single comparison measure to conduct the remaining components of the methodology. Armed with a manageable number of binary similarity measures with which to generate source similarity scores, source weightings can be determined and the sources can be grouped to assess their likely reliability.

V. Source Comparison

This chapter details and conducts the remaining components of the methodology. This chapter begins by assuming a binary similarity measure has been selected and describes in detail the remaining components of the methodology. It then steps through a small example employing the methodology for demonstration purposes. Then, utilizing the experimental design detailed in Section 3.3 and the set of binary similarity measures identified using the methodology developed in Chapter IV, the chapter presents the selection of a single measure to generate source similarity scores. It will then demonstrate the methodology components of calculating source weightings and the clustering of sources to assess their likelihood of reliability. Next, it examines the performance of the methodology by utilizing the DOE described in Section 3.3. The experimentation investigates factors that affect the methodology's performance in distinguishing between reliable and unreliable sources. The chapter concludes with a discussion of the analytical results.

5.1 Examining the Collection of Sources

Binary similarity measures enable direct pairwise comparisons among sources, but do not allow simultaneously consideration of all S sources. The pairwise comparisons of sources and the selected binary similarity measure, φ_{ij}, can be assembled into a binary similarity measure matrix. From this a dissimilarity matrix can be constructed as displayed in Figure V-1. The dissimilarity between sources i and j, δ_{ij}, is obtained from the obverse of the binary similarity measure, with δ_{ij} equaling the theoretical maximum possible value minus the theoretical minimum value for φ minus φ_{ij},

which will be nonnegative by construction. Source comparisons for which the binary similarity measure could not be computed are recorded at the minimum value of φ_{ij} which is the maximum value of δ_{ij}. This indicates the sources are not confirming the information presented by each other. This procedure transforms the pairwise binary similarity scores into nonnegative dissimilarities. It also ensures that every pairwise source combination has a value by placing the theoretical maximum dissimilarity for the source pairings for which computation of the binary similarity score was impossible.

		Sources				
		1	2	3	…	*S*
Sources	1	-	δ_{12}	δ_{13}	…	δ_{1S}
	2	δ_{21}	-	δ_{23}	…	δ_{2S}
	3	δ_{31}	δ_{32}	-	…	δ_{3S}
	⋮	⋮	⋮	⋮	-	⋮
	S	δ_{S1}	δ_{S2}	δ_{S3}	…	-

Figure V-1 Dissimilarity Matrix

5.1.1 Multidimensional Scaling (MDS).

To analyze the amount of consistency in reporting across all sources, multidimensional scaling (MDS) can be applied to the dissimilarity matrix to visualize the conformity and disagreement of the complete collection of information. It should be noted, however, that this technique does not validate or identify reliable sources, but merely displays groupings of commonalities in reporting.

5.1.1.1 MDS of Source Comparisons.

Multidimensional scaling visually groups objects based upon their proximities to other objects. In this instance, objects' proximities are derived from the binary similarity measure as described by the correlation matrix. Since the proximities are restricted to be interval scaled, metric multidimensional scaling can be used, and an exact equation can be specified to convert the proximity values in the correlation matrix to the distances displayed on the MDS mapping (Dillon & Goldstein, 1984, pp. 108, 126). The visual groupings of the various sources could be informative, once quantitative assessment of individual source reliability has been performed. A notional MDS visualization consisting of eight hypothetical social network information sources, denoted S1 through S8, is presented in Figure V-2.

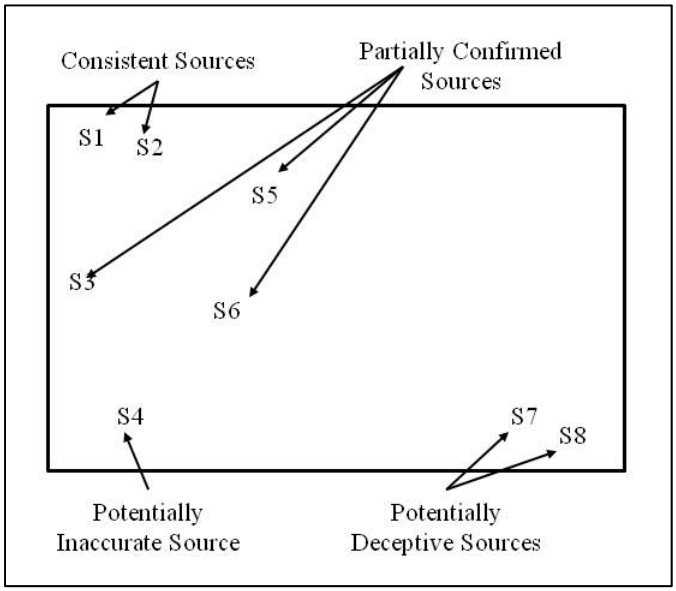

Figure V-2 Notional MDS Visualization of Social Network Information Sources

It is expected that the visual dispersion of sources as depicted through a MDS representation will generate clusters of the sources. Sources reporting similar social network data will group together, while sources providing incongruous information will be spatially distant in the MDS mapping. Furthermore, the expectation of significant distance between sources in the MDS mapping can be classified as stemming from two causal mechanisms.

5.1.1.2 MDS Distance Interpretation.

First, it is conceivable, and probably likely, that two network information sources are reporting data representing different aspects of the social network, as these sources may report non-overlapping network information that may reflect their observations of mutually exclusive sets of actors in the network. This non-overlapping characterization of the network, reporting on different components within the social network, may be visually represented in the MDS mapping as a significant distance between the reporting sources.

Second, large distances between sources in the MDS visualization may stem from sources reporting discordant information. Sources could be providing data representing the relationships among a specific subset of actors in the social network; however, the individual source reports may be in direct conflict on the status, present or null, of the relationships. This discordance would be illustrated in the MDS visualization as significant distance between incongruent sources. In terms of source reliability assessment, this second case illustrates the utility of applying multidimensional scaling against the sources' reporting patterns.

Potentially coupled with statistical clustering techniques, veracity assessments of source reliability could be derived from the MDS visualization. Sources that are concordant will group together in the MDS visualization, while discordant sources will possess distance between them. SNA analysts visually identifying sources that are tightly clustered can assess that those comprising sources are concordant which may indicate source veracity. Conversely, visually identified isolates sources are not concordant. Non-concordance may be a reflection of dissention in reporting or a function of the sources' network perspectives.

5.1.2 Weighting Source Comparisons.

High binary similarity measures' scores reflect the confirmation of dyads or null relationships among information sources. Low binary similarity measures' scores reflect disagreement among the sources. However, the amount of confirmations or disagreements can vary from a single confirmation/disagreement to many. Thus, as displayed in Figure V-3, there are two dimensions to consider when comparing sources: the level of concordance/disagreement which is reflected in binary similarity measure scores, and the amount of information being compared between the sources.

Information sources may be reporting on different aspects of the social network, with each source presenting its own perspective. If information sources are reporting on the same subgraph structures of the underlying social network, then confirmation and dissenting reports between these sources should be considered extensively, and should be reflected in a weighting of the sources similarity score. Conversely, if two information sources are reporting on different structural aspects of the social network, then their

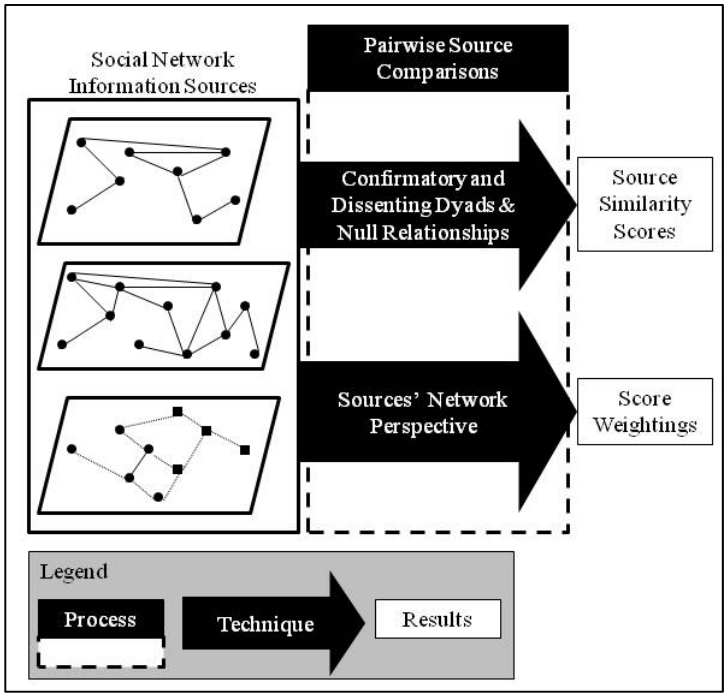

Figure V-3 Similarity Score Weightings

associated similarity score should be relatively discounted via a low weighting. This process will result in pairwise score weightings characterizing every sources' network perspective.

The pairwise weightings of sources reflect the social network perspective of the two sources. A high weighting value indicates that the sources are reporting on similar substructures of the network. A low weighting value implies that the sources are reporting on different portions of the social network. The weightings do not provide any information regarding the concordance among the sources, but only characterize the number of dyads in common between each pair of information sources in the combined network. Thus a low weighting of a pair of sources implies that little inference can be drawn between the sources' reports. With a high weighting, a pair of sources are

reporting on the same aspect of the social network and one would expect a high similarity score; if a low similarity score is present the sources are providing conflicting reporting indicating one or both are unreliable.

As each source is likely to provide differing amounts of social network information, applying weights to the MDS may better reflect the overall structure present in the complete data collection. In MDS, weights may be incorporated on each pair of sources, in this case each δ_{ij}, (Borg & Groenen, 2005, p. 254). The nonnegative dissimilarity between sources i and j, δ_{ij}, in this case, is obtained from the obverse of the binary similarity measure, $\delta_{ij} = \varphi - \varphi_{ij}$.

In this research, weights are derived by the amount of data that can be used for confirmation between two sources as a proportion of the total data presented by the two sources. Two sources, S_i and S_j, provide social network information involving the set of actors, N_i and N_j, respectively. The weighting, w_{ij}, of the relevancy of the corresponding δ_{ij} is then dependent upon the information that is reporting the same aspects of the network divided by the combined total amount of data provided by the two sources. This is reflected in Equation (5.1), which is the number of dyads reported by both sources divided by the total number of dyads reported by both sources. This enables sources that are reporting information regarding the same relationships to have their agreement or disagreement weighted heavier than sources reporting information on different aspects of the social network. It is possible for two sources reporting on completely different aspects of the social network to possess a weighting of zero.

$$w_{ij} = \frac{\left|N_i \cap N_j\right|}{\left|N_i \cup N_j\right|} \tag{5.1}$$

5.1.3 Weighted MDS.

MDS maps the δ_{ij} onto a m-dimensional MDS configuration X according to the mapping that minimizes a badness-of-fit measure, normalized stress, denoted $\sigma_n(X)$. The distance between each of the points in the mapping, $d_{ij}(X)$, are measured in terms of Euclidean distance. Euclidean distance is used to aid analyst interpretation of the visualization, where greater spatial distance between objects in the visualization represents greater difference between the objects in the original space. Normalized stress, $\sigma_n(X)$, is the proportion of the sum-of-squares of the original δ_{ij} that is not accounted for in the new mapping of the data points and their associated distances taking into account the source weightings, w_{ij}, as shown in Equation (5.2). There are several available methods and techniques developed for MDS to attempt to minimize normalized stress while also achieving dimensionality reduction (Borg & Groenen, 2005, pp. 42, 122, 248).

$$\sigma_n(X) = \frac{\sum_{i<j} w_{ij}\left(\delta_{ij} - d_{ij}(X)\right)^2}{\sum_{i<j} w_{ij}\delta_{ij}^2} \tag{5.2}$$

Due to incorporating weights, source pairings that are reporting on different aspects of the network have no impact on the MDS placement of each other. Thus, the weighted MDS layout is constructed only by sources that possess a positive weighting indicating at least some level of complimentary reporting. Source pairings possessing substantial overlap in their reports are more influential in the MDS layout.

The resulting visualization gives the social network analysts a visual aid depicting the sources' concurrence. In some cases, the visualization may be sufficient to provide clear interpretation of which sources should be included in the social network model. In

more complex cases, the inherent analyst subjectivity of interpreting a visualization may create difficulties in assessing the appropriateness of source inclusion.

5.2 Grouping Sources

Addressing the subjectivity and the associated induced variabilities in assessing of visualization interpretation, a more quantified approach is presented here. The weighted MDS visualization of the sources arranges the sources in accordance with their level of concurrence and the amount of information they are reporting. Cluster analysis is a collection of statistical techniques to group objects based on similarity or distance measures. The objective is to group objects into clusters "that display small within-cluster variation relative to the between-cluster variation (Dillon & Goldstein, 1984, p. 158)."

Grouping sources aids the SNA analysts in absorbing the large amount of information captured by the similarity scores while taking into accounting weightings reflecting the sources' network perspectives. Statistical clustering techniques can be applied to group sources, so that the SNA analyst can visually inspect source concordance. The clustering is based on the source similarity sources and the score weightings, as shown in Figure V-4. Clustering can be accomplished by visual inspection by the SNA analyst, or via statistical clustering techniques to provide a quantified approach.

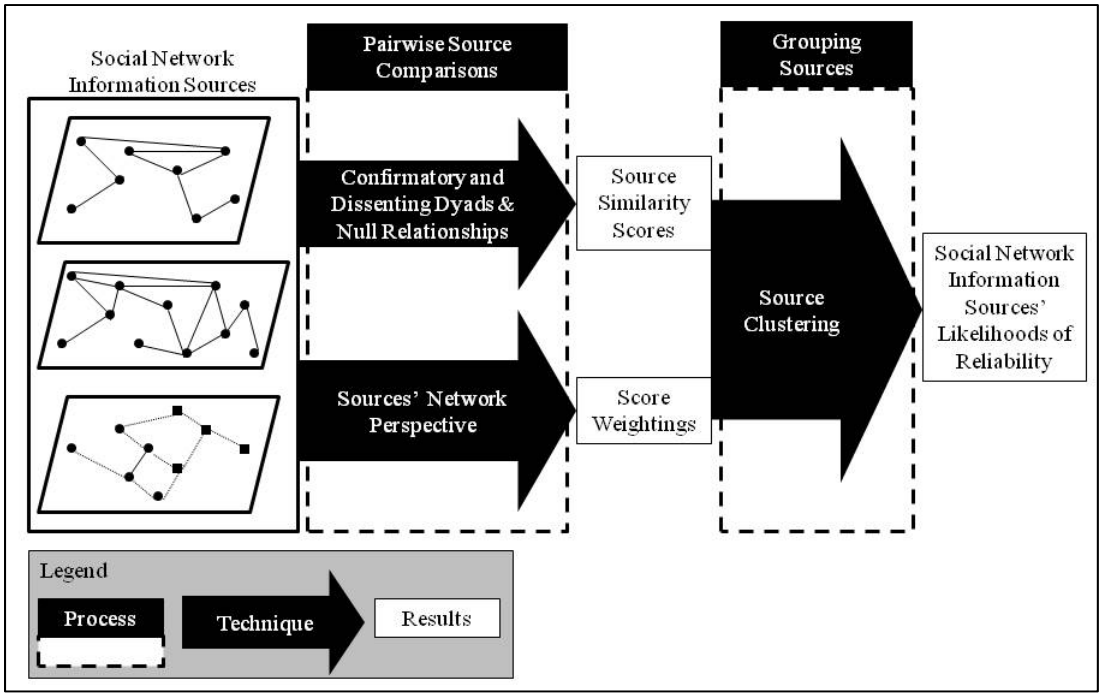

Figure V-4 Grouping Sources

5.2.1 Fuzzy Clustering.

Clustering analysis positions each object into a cluster. Alternatively, fuzzy clustering methods scores each object with membership coefficients. Membership coefficients range from 0 to 1, with greater values indicating a stronger preference for a cluster. The membership coefficients for an object will sum to one across all clusters. This approach gives more detailed information than traditional "hard" clustering techniques. One method for conducting fuzzy cluster is the FANNY procedure. FANNY's heuristic attempts to minimize the objective function presented in Equation (5.3) for objects 1 through n with associated dissimilarities $d(i,j)$, clusters $v = 1 \dots k$, and u_{iv} is the membership coefficient of object i for cluster v. The minimization is

subject to two constraints: the membership coefficients are nonnegative and sum to one for each object (Kaufman & Rousseeuw, 1990, pp. 164-166, 182).

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^2} \tag{5.3}$$

Fuzzy clustering's output is the membership coefficients of each object. For this methodology, setting the number of clusters to two partitions the sources into two groups: social network model inclusion and exclusion. Two clusters are arbitrary, as the weighted MDS visualization may suggest more clusters. But in the interest of providing a quantified consistent methodology, conducting fuzzy clustering for two clusters on the sources is conducted in the experimentation.

5.2.2 Fuzzy Clustering of Weighted MDS.

Fuzzy clustering could be performed directly on the dissimilarities matrix composed of the δ_{ij} dissimilarities between sources *i* and *j*. The weightings in weighted MDS account for sources reporting on different aspects of the social network. Performing fuzzy clustering on the dissimilarities matrix would only address the lack of confirmatory reporting between sources, but could not ascribe whether this is a result of disagreement or a lack of common reporting.

5.2.3 Membership Coefficient Interpretation.

The membership coefficients indicate the likelihood that an object belongs to a cluster or group. For source assessment in the case of restricting the fuzzy clustering to two clusters, the sources are given likelihoods of belonging to one of two groupings. The clusters have several potential interpretations. One interpretation involves the grouping

reflecting sources that are confirming other group members' reporting. If that occurs, the remaining cluster could reflect sources whose reporting is unconfirmed or discredited. It could also reflect a grouping of sources who are reporting on different aspects of the social network.

The fuzzy clustering technique produces membership coefficients based on the weighted MDS of the source similarity scores and their associated score weightings. These membership coefficients can be interpreted as source likelihoods of reliability. In the case of non-statistical based source clustering, the grouping of the sources indicates their likelihood of reliability in a qualitative manner. Sources grouped in the same cluster are interpreted as reporting concordant information on the same social network aspects. Sources grouped in difference clusters are interpreted as either reporting discordant information or on different aspects of the social network. An examination of the pairwise source similarity scores and weightings between sources located in different clusters can quickly identify if their separation is a function of their network perspective or disagreement in information reporting.

The fuzzy clustering membership coefficients can also be considered a threshold. If a SNA analyst determines one of the clusters to contain reliable sources, the membership coefficients for that group are a threshold for inclusion into being considered reliable. Dependent upon analytical considerations, such as operational risk that result from the SNA conclusions, the SNA analyst can specify an appropriate threshold. In the experimentation conducted in the research, utilizing the AUC as the response variable in effect examines every potential threshold that could be selected by an SNA analyst. This

allows the methodology's performance to be tested with every possible scenario of SNA analyst specified thresholds.

5.3 Methodology Overview

The methodology begins with social network information sources being compared in a pairwise manner. These pairwise source comparisons are conducted by examining the dyads they have in common and noting agreement and disagreement on the existence of relationships and null relationships. This is distilled into source similarity scores which are based on binary similarity measures, following the procedures in Section 5.1. A methodology for selecting binary similarity measures was developed and described in Chapter IV.

The sources' network perspective was examined via a source reporting weighting function as described in Section 5.1.2. The source similarity scores and score weightings are transformed into a small dimension space via weighted MDS as described in Section 5.1.3. The weighted MDS also serves as a visual depiction of the source reporting and concordance for SNA analysts. The information sources are then grouped based upon their concordance and perspective. The grouping can be accomplished by visual inspection of the weighted MDS, but a quantified approach was developed here based on fuzzy clustering of the weighted MDS and described in Section 5.2. The grouping of the sources enables identification of concordant sources, discordant sources, and sources reporting unique information on aspects of the social network. These groupings allow the SNA analysts to draw conclusions on individual information sources' likelihood of reliability. Utilizing the fuzzy clustering's membership coefficients, this likelihood is

quantifiable. Figure V-5 provides an overview of the methodology, noting the techniques used as each step.

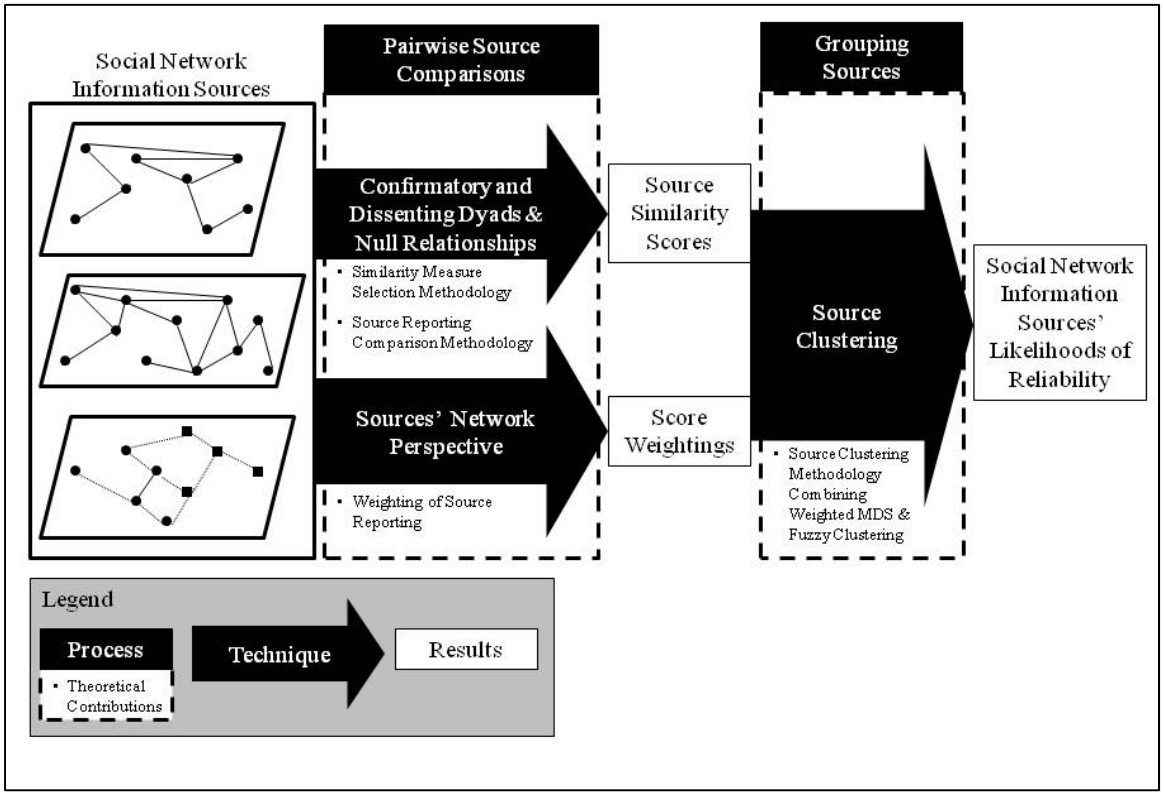

Figure V-5 Overall Methodology Framework

5.4 Trusted Sources

In some instances, a source is known to be reliable and accurate in the information it reports. These trusted sources can come in several varieties in regards to intelligence collection on dark networks. Technical means can passively observe a dark network organization's activities and communications. Human agents can insert themselves into the organization or cultivate informants; both provide an active means of determining the organization's structure, participating actors, and relationships.

Regardless of how the information is collected, its associated reliability is assumed to be accurate and free of deception; therefore, trusted sources are used to assist in assessing other sources of information via confirmation or contradiction.

Some sources' nature eliminates, or at a minimum diminishes, the need of assessing the source's accuracy. Sources such as signals or communications intelligence provide accurate reporting, although still only a specific perspective on the networks. Signals intelligence may confirm communication occurring between members of the network, but is limited to electronic communication and probably will not accurately capture face-to-face interactions. Trusted sources, such as undercover agents, can be assumed to provide accurate, reliable information. These *bona fide* sources will not need to be assessed for accuracy, but instead can provide a basis for the other sources to be compared against.

5.4.1 COMINT.

Communications intelligence (COMINT), utilizing monitoring of a target's electronic communications, generates data that can be construed as reliable. Assuming the adversary is unaware which actors, conversations, and communication devices are being monitored, inhibiting collection can only be achieved by broad measures, such as having all members avoid usage of specific types of electronic communication devices. The adversary can complicate COMINT collection by using multiple means of communication creating difficulties in accurately recreating the entire communication network and associated patterns. However, information collection via COMINT in either of these scenarios is still unlikely to be inaccurate, unless opposed by an extensive denial

and deception campaign. COMINT derived social network data has the advantage that the communications are known to occur as Actor A talked with Actor B and the relationship existence is not in question. Difficulties can arise in actor attribution, and identifying all actors in intercepted communications.

COMINT is unlikely to generate random observations of the social network, but more likely will be similar to the snowball data collection technique delivering observations on local structures within the social network. As an actor is targeted by COMINT collection means, the number of communication devices associated with individuals in the network grows. Monitoring a known dark network actor's phone identifies the phone numbers of those they call, creating an ever growing network of communication devices that can be surveyed. However, a potential drawback is the incorrect assessment of actor participation in the dark network based on irrelevant communications. For example, a dark network actor may regularly communicate with family members who have no affiliation with the organization. COMINT may not be able to discern the nature of the relationship, but merely that communication occurs and with which frequency and associated pattern.

5.4.2 HUMINT.

Human intelligence (HUMINT) sources can also be considered trusted under certain circumstances. The HUMINT asset employed may possess a substantial history of reporting or be an agent of government forces purposely emplaced in the dark network. Other than these special cases, informants providing social network data on a

dark network organization may report accurate, erroneous, or deceptive information. The difficulty lies in discerning an untested HUMINT source's reliability.

5.4.3 Modeling Trusted Sources.

Trusted sources can be modeled using any of the reliable source generation techniques. The only difference between a trusted source and a reliable source is that a trusted source is known and assumed to provide accurate information, while a reliable source's veracity must be determined. Thus, sources that are being confirmed as concordant with trusted sources are assessed as reliable, and conversely sources whose information is being discredited by trusted sources are assessed as unreliable. Applying the methodology, information sources that possess high membership coefficients into groupings that contain trusted sources can be inferred to be reliable sources.

5.5 Example

This section provides an example of the methodology applied to a problem to demonstrate the individual steps of the methodology. The problem is pulled from the experimentation (details are provided in Section 3.3) and is obtained from the second replication of the full factorial design point with all factors set at -1, i.e. run "1". For this data, the true underlying social network is composed of 200 actors. Ten sources reported information on the network with 6 sources providing reliable information and 4 sources providing unreliable information.

5.5.1 Source Reporting Data.

The reliable sources are denoted R1 through R6 and the unreliable sources are denoted U1 through U4. The sources' reporting is presented in Table V-1 and Table V-2

as edge lists. Actors in the network are uniquely identified by a number ranging from 1 to 200 in this case. As can be seen in Table V-1 and Table V-2, each source is reporting on different relationships in the network as well as reporting varying amounts of information.

Table V-1 Reliable Sources' Edge Lists

R1		R2		R3		R4		R5		R6	
127	190	55	183	77	135	148	180	148	180	71	155
71	139	55	105	149	151	155	190	39	63	38	71
20	45	17	190	28	151	28	151	71	116	133	140
37	71	15	190	78	182	129	190	53	112	23	145
44	142	90	103	65	158	55	139	23	115	60	82
19	172	52	55	158	190	23	145	6	190	53	112
155	166	107	162	55	134	155	166	4	49	14	112
60	82	51	70	44	107	53	112	58	190	165	170
158	194	48	75	71	190	5	11	107	164	112	167
7	156	64	164	7	73	68	190	51	190	155	184
135	190	47	155	107	190	15	190	43	79	91	155
7	170	46	59	112	173	84	105	77	135	112	190
9	107	65	158	1	107	10	158	64	190	4	74
107	198	133	140	107	164	110	190	29	190	7	41
190	191	67	150	19	112	51	70	67	150	109	151
24	129	112	167	51	190	190	195	60	82	51	71
23	74	7	74	12	143	69	190	126	133	51	70
48	116	135	190	185	190	23	195	176	196	43	117
		142	159	21	180	44	62	107	129	48	118
		98	155	44	82	21	161	107	121	71	95
		110	132	48	113			142	159	35	90
		32	93	43	118			10	158		
		190	195					110	190		
		34	155					112	155		
		74	112					23	74		
		44	82								

Table V-2 Unreliable Sources Edge Lists

U1		U2		U3		U4	
181	197	63	152	140	158	72	194
13	109	79	80	59	165	63	93
24	169	148	185	32	154	132	158
16	29	8	171	46	171	30	102
13	16	61	80	45	67	75	179
170	189	114	116	6	142	102	132
13	18	104	124	82	89	6	156
38	57	21	150	97	106	101	132
2	119	60	107	67	165	85	132
2	61	97	121	67	137	100	173
2	48	34	179	41	125	21	145
72	149	35	79	68	176	56	150
45	74	79	198	36	145	33	136
44	52	32	63	29	165	133	145
62	84	152	167	59	143	71	166
63	165	37	152	140	171	45	101
64	166	66	68	119	140	69	156
13	163	50	108	33	186	92	145
2	191	84	123	84	183	12	145
2	159	28	185	122	153	13	145
13	124	84	125	85	94	105	135
13	75	65	139	31	140	105	145
		94	144	124	145	145	187
		94	157	82	173	2	147
		13	180	33	125	34	105
		14	107	10	149		
		107	115	67	198		
		93	158	2	129		
		107	121	2	76		
		4	170	4	53		
		107	144	166	192		
		64	82	3	72		
		170	185	2	185		
		79	131	74	189		
		181	189	32	105		
		79	120				

5.5.2 Source Comparisons.

Next, the comparison matrices were computed as described in Section 2.8 and then summarized by a binary similarity measure for each pairwise combination. Cohen's Kappa was the binary similarity measure employed. Details of Cohen's Kappa are provided in Section 5.6.1. The pairwise Cohen's Kappa scores are provided in Table V-3.

Table V-3 Cohen's Kappa Scores for the Example

	R1	R2	R3	R4	R5	R6	U1	U2	U3	U4
R1	0	0.87	0.46	0.97	1.05	0.87	0.50	0.46	0.00	0.45
R2	0.87	0	0.87	1.14	0.80	1.15	0.45	0.45	0.48	0.50
R3	0.46	0.87	0	0.88	1.15	0.43	0.00	0.45	0.50	0.00
R4	0.97	1.14	0.88	0	1.13	1.22	0.30	0.50	0.50	0.50
R5	1.05	0.80	1.15	1.13	0	0.82	0.00	0.76	0.48	0.00
R6	0.87	1.15	0.43	1.22	0.82	0	0.50	0.45	0.46	0.50
U1	0.50	0.45	0.00	0.30	0.00	0.50	0	0.46	0.46	0.50
U2	0.46	0.45	0.45	0.50	0.76	0.45	0.46	0	0.50	0.44
U3	0.00	0.48	0.50	0.50	0.48	0.46	0.46	0.50	0	0.50
U4	0.45	0.50	0.00	0.50	0.00	0.50	0.50	0.44	0.50	0

Cohen's Kappa's maximum possible value is 1.5, and the dissimilarity matrix was constructed by subtracting the Cohen's Kappa score from its maximum possible value (Fleiss, Levin, & Paik, 2003, p. 603). The dissimilarity matrix is displayed in Table V-4. A dissimilarity score of 1.5 indicates either maximum disagreement on provided information or that the sources reported data on completely different aspects of the networks. Excluding the dual potential interpretations of maximum possible scores of 1.5, high dissimilarity scores indicates disagreement between sources.

Table V-4 Dissimilarity Scores for the Example

	R1	R2	R3	R4	R5	R6	U1	U2	U3	U4
R1	0	0.63	1.04	0.53	0.45	0.63	1.00	1.04	1.50	1.05
R2	0.63	0	0.63	0.36	0.70	0.35	1.05	1.05	1.02	1.00
R3	1.04	0.63	0	0.62	0.35	1.07	1.50	1.05	1.00	1.50
R4	0.53	0.36	0.62	0	0.37	0.28	1.20	1.00	1.00	1.00
R5	0.45	0.70	0.35	0.37	0	0.68	1.50	0.74	1.02	1.50
R6	0.63	0.35	1.07	0.28	0.68	0	1.00	1.05	1.04	1.00
U1	1.00	1.05	1.50	1.20	1.50	1.00	0	1.04	1.04	1.00
U2	1.04	1.05	1.05	1.00	0.74	1.05	1.04	0	1.00	1.06
U3	1.50	1.02	1.00	1.00	1.02	1.04	1.04	1.00	0	1.00
U4	1.05	1.00	1.50	1.00	1.50	1.00	1.00	1.06	1.00	0

The source comparison weightings were computed using Equation (5.1) as described in Section 5.1.2 and are provided in Table V-5. The weightings range from zero to one with large scores indicating reporting on the same aspects of the network. As can be seen by examining the R1-U3 comparison, the weight is zero, indicating that the two sources' reporting did not contain any dyads in common. This, in effect, nullifies the 1.5 maximum dissimilarity score in Table V-4 when the weighted MDS is applied. Following the same logic, sources R1 through R6 possess relatively strong source weightings among themselves, indicating they are reporting on the same substructures in the social network. Conversely, U1 through U4 possess relatively low source weightings which is indicative or reporting unique information.

$$w_{ij} = \frac{\left|N_i \cap N_j\right|}{\left|N_i \cup N_j\right|} \tag{5.1}$$

Table V-5 Source Weightings Matrix for the Example

	R1	R2	R3	R4	R5	R6	U1	U2	U3	U4
R1	0	0.17	0.19	0.15	0.22	0.19	0.14	0.12	0.00	0.12
R2	0.17	0	0.21	0.19	0.25	0.21	0.10	0.12	0.13	0.13
R3	0.19	0.21	0	0.19	0.19	0.18	0.00	0.10	0.08	0.00
R4	0.15	0.19	0.19	0	0.20	0.16	0.07	0.11	0.12	0.10
R5	0.22	0.25	0.19	0.20	0	0.21	0.00	0.18	0.15	0.00
R6	0.19	0.21	0.18	0.16	0.21	0	0.10	0.09	0.10	0.05
U1	0.14	0.10	0.00	0.07	0.00	0.10	0	0.12	0.16	0.11
U2	0.12	0.12	0.10	0.11	0.18	0.09	0.12	0	0.15	0.10
U3	0.00	0.13	0.08	0.12	0.15	0.10	0.16	0.15	0	0.14
U4	0.12	0.13	0.00	0.10	0.00	0.05	0.11	0.10	0.14	0

Source comparisons with large dissimilarity scores and large weightings, such as U1's and U3's values in Table V-4 and Table V-5, indicate substantial disagreement on the social network. Conversely, small dissimilarity scores and large weightings, exemplified by the R2 and R3 source comparison, indicate substantial agreement on the social network's structure. Comparisons possessing small weightings imply that little inference can be drawn from the pairwise comparison as the sources are reporting on different aspects of the social network. Visual inspection of the dissimilarity and source weightings matrices by SNA analysts can prove difficult if facing numerous information sources. Utilizing weighted MDS to summarize the information contained within these matrices visually is presented next.

5.5.3 Weighted MDS.

The weighted MDS visualization is displayed in Figure V-6 with an aspect ratio of one, generated as described in Section 5.1.3. This analyst aid indicates that the sources denoted R1 through R6 appear to be confirming each other's reporting, while sources U1 through U4 are either unconfirmed or discordant with the other sources. Examining the

dissimilarity scores in Table V-4 shows that U1 through U4 have relatively high scores which results from conflicting information. Examining the weightings in Table V-5, U1 through U4, the weightings are low indication that the information U1 through U4 provide is unconfirmed by other sources.

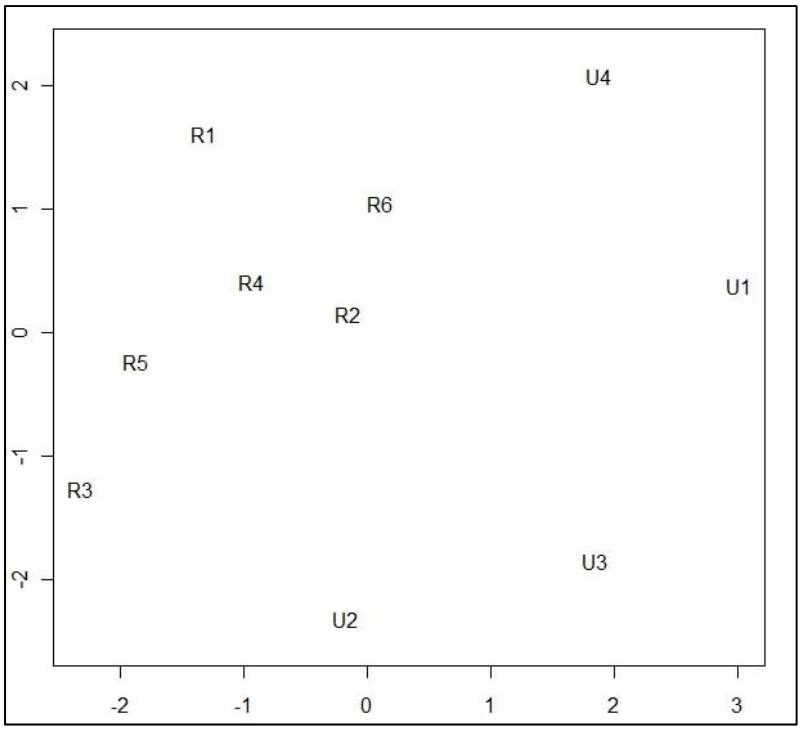

Figure V-6 Weighted MDS Visualization of Sources for the Example

As weighted MDS attempts to minimize an objective function of total normalized stress of the graph, given by Equation (5.2), one can examine every sources' contribution to the total normalized stress. Figure V-7 displays the sorted stress contributions by each source. As can be seen, sources U1 through U5 contribute more stress than the other sources for their placement on the weighted MDS visualization. The stress is a reflection of the difficulty of placing the source on the visualization as it is every sources'

contribution to the objective function of the weighted MDS, as shown in Equation (5.2). Sources possess high stress as a result of inconsistent similarity scores across all of their source pairings. Inconsistent similarity scores occur when a source possesses a high similarity score with one source and a low similarity score with another source, but those two sources possess a high or low similarity score. If this inconsistency occurs repeatedly across all source pairs, it is a potential indication of source unreliability.

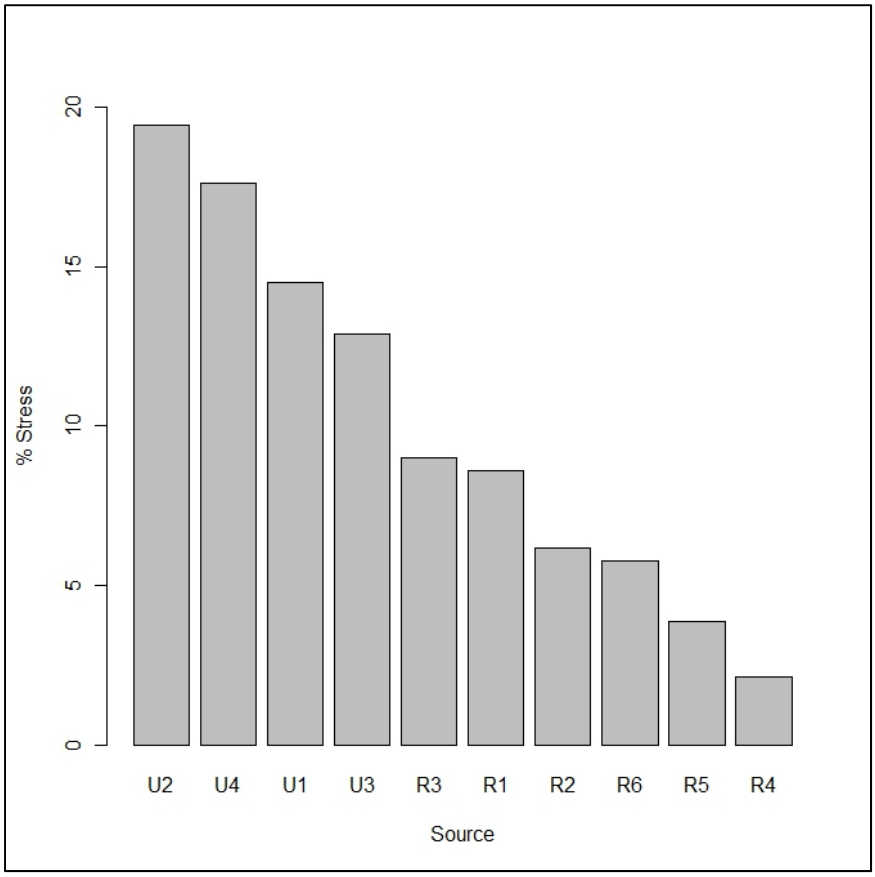

Figure V-7 Source Stress Contributions for the Example

5.5.4 Fuzzy Clustering.

The fuzzy clustering membership coefficients, as described in Section 5.2.3 are presented in Table V-6. Analogous to the weighted MDS visualization, the membership coefficients present the groupings of the sources. Table V-6 shows sources R1 through R5 with strong preferences to Group 1. R6's and U2's membership coefficients are close to 0.5 and could indicate placement in either group. U1, U3, and U4 indicate strong preference for Group 2. Please note, that the group number is incidental, of interest, is which information sources are being placed together in the same groupings.

Table V-6 Fuzzy Clustering Membership Coefficients for the Example

	Group 1	Group 2
R1	0.68	0.32
R2	0.66	0.34
R3	0.70	0.30
R4	0.80	0.20
R5	0.79	0.21
R6	0.55	0.45
U1	0.24	0.76
U2	0.51	0.49
U3	0.32	0.68
U4	0.32	0.68

The fuzzy clustering membership coefficients can be interpreted as a threshold for inclusion in the social network model. As Group 1 is internally consistent among its members, one can construct a ROC curve by using the membership coefficient for Group 1 as the parameter for inclusion. The ROC curve is displayed in Figure V-8 and its associated AUC is one. The AUC of one indicates that no matter where the threshold is specified, no unreliable source will be included in the social network model while a

reliable source is excluded. This means the methodology has associated all of the reliable sources with higher membership coefficients and unreliable sources possess lower membership coefficients. In other words, the reliable source with the smallest membership coefficient is still greater than the unreliable source with the largest membership coefficient. Applying the methodology developed in this dissertation does not always result in the trivial ROC curve as in Figure V-8 which indicates perfect classification performance.

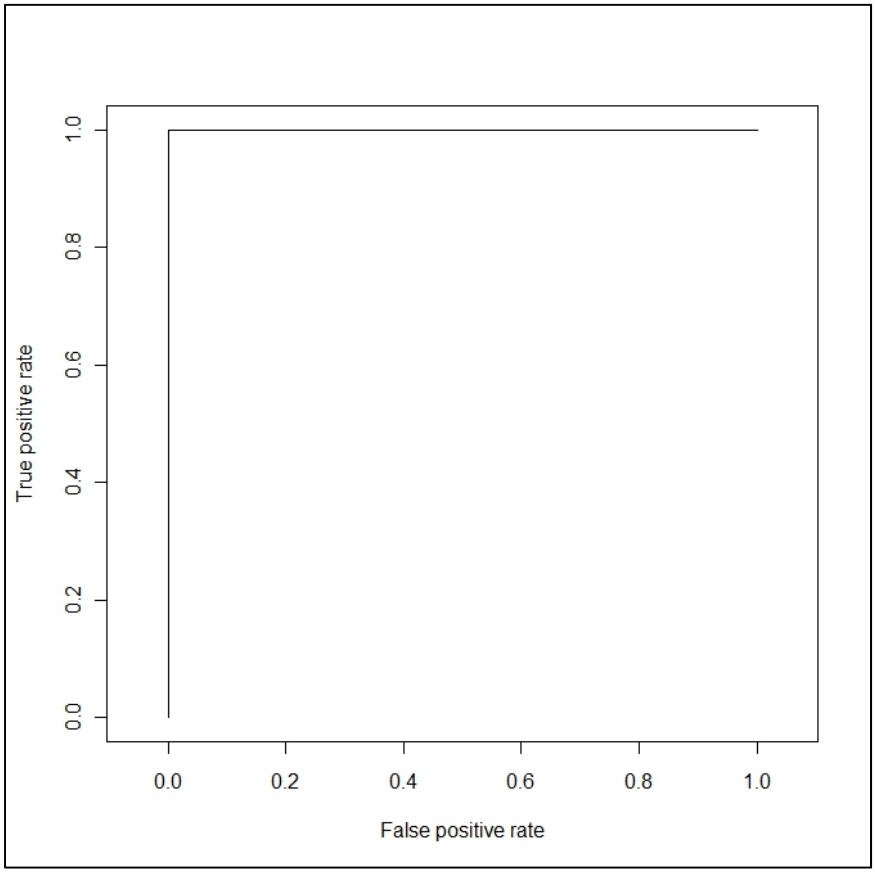

Figure V-8 ROC Curve for the Example

5.5.5 Analytical Conclusions for the Example.

Examining the fuzzy clustering membership coefficients, an analyst can derive that R1 through R5 are reporting different network information than U1, U3 and U4. By extending the analysis to consider the dissimilarity scores and the weightings, depicted by the weighted MDS visualization, the analyst can draw several conclusions. R1 through R5 are providing information that is confirmed by each other, inferable from their larger membership coefficients in Table V-6 and their close proximity to each other in Figure V-6. U1, U3 and U4 are presenting information that is generally unconfirmed by other sources, though there are disagreements with the other sources when confirmation is possible. This inference is from observing the low membership coefficients for Group 1 in Table V-6 and the large distances among U1, U3, and U4 in Figure V-6.

U2 is more difficult to assess than the other reporting sources. U2 possesses a membership coefficient near 0.5, in this case a 0.51, indicating it could be assigned to either group by the fuzzy clustering. What is informative is U2's placement on the weighted MDS visualization in Figure V-6, though it is susceptible to analyst interpretation. U2 appears to be far from all other sources indicating that its information is not confirmed. It closest source is U3, which has at this point already been identified as an unreliable source. Examining U2's dissimilarity scores and weightings in Table V-4 and Table V-5, respectively, U2 has relatively high dissimilarity scores and moderate weightings. This combination reflects the other information sources are reporting on the same network substructures to an extent, but they disagree from U2's reporting. U2 would most likely be assessed by the SNA analysts as an unreliable source. Depending upon the acceptable level of operational risk, U2's reported information's inclusion may

be based on the impact it has on the combined social network model and the specific SNA techniques being applied. Non-quantified information and, of course, SME judgment can be applied at this point to make the final determination.

5.5.6 Trusted Source Extension.

Extending the example to include a trusted source, somewhat simplifies the analysis. If source R4 is a trusted source, the SNA analyst can use that information to draw conclusions regarding information sources that are concordant with R4. Examining the weighted MDS in Figure V-6, information sources in close proximity to R4 can be assessed as reliable. Under this paradigm, sources R1, R2, R5, R6 and possible R3 would be assessed as reliable sources due to their concordance to trusted source R4. This in effect, uses trusted sources to vet other sources.

If U1 is a trusted source, the methodology enables an analytical interpretation that might be overlooked by SNA analysts not employing the methodology described here. Due to the large spatial distances between U1, the trusted source, and all of the other information sources, one could quickly draw the conclusion that the other information sources are either unreliable or reporting on different aspects of the social network. Discerning the clustering of R1, R2, R4, R5, and R6, an analyst could determine that despite the lack of reporting confirmation from the trusted source, U1, these clustered sources are confirming each other's reports while reporting on different substructures of the underlying social network. Since, the trusted source's reports are not being confirmed by any other sources, it may be an indication that the trusted source has been compromised, is being spoofed, or is collecting information based on a deception

operation. Regardless, in this example, employing the methodology enabled the SNA analyst to quickly and quantitatively assess the information sources' reliability and easily identify sources requiring further examination.

5.6 Initial Examination of Performance Measures

At this point the methodology has been presented and an example problem was examined in detail. To evaluate the efficacy of the methodology developed in this research, it will be tested according the design of experiments (DOE) described in Section 3.3. The DOE examines the methodology's performance under various conditions that are similar to real world problems faced by SNA analysts. The testing of the methodology begins with examining the seven binary similarity measures deemed suitabile for the source assessment application identified in Section 4.3.3.

The seven binary similarity measures, selected using the methodology implemented in Chapter IV, were applied to the data generated according to the DOE specified parameterization. The seven selected binary similarity measures are: *Cohen's Kappa*, *Gini coefficient*, *Hamann*, *Dispersion*, *Peirce-III*, *Anderberg*, and *Goodman & Kruskal's Maximum formula*. Among the seven measures, there are substantial differences in performance. The Area Under the Curve (AUC) generated from the fuzzy clustering membership coefficients was utilized as a performance measure. The algorithm was applied for all seven selected binary similarity measures and their corresponding AUCs were computed for all runs of the dataset.

Table V-7 provides descriptive statistics on the AUC values of the seven measures which were calculable in all cases. As evident in Figure V-9, several of the

measures exhibit substantial differences between their mean and their median, AUC values, represented by filled points and by the traditional box plot median line, respectively. Cohen's Kappa displayed superior performance in comparison to the other six binary similarity measures utilized in this experiment, with its median and mean bother showing substantially better performance than the other measures. The worst case observed using Cohen's Kappa in the methodology exhibiting better performance than 25% of the tests using the other measures. The first quantile of the AUC values generated from applying the methodology with Cohen's Kappa exceeds the third quantile of each of the other six binary similarity measures—the 25th percentile of the methodology executed using Cohen's Kappa exceeded 75% of the test results utilizing the other six binary similarity measures.

Table V-7 Seven Measures' AUC Values Descriptive Statistics

	Cohen	**Gini**	**Hamann**	**Disper.**	**Peirce**	**Anderberg**	**GK Max**
Min.	0.563	0.50	0.50	0.50	0.50	0.50	0.50
1st Quart.	0.833	0.548	0.542	0.548	0.535	0.519	0.507
Median	0.962	0.625	0.596	0.590	0.572	0.543	0.523
Mean	0.921	0.647	0.632	0.628	0.607	0.587	0.540
3rd Quart.	0.995	0.729	0.689	0.674	0.632	0.609	0.552
Max.	1.0	1.0	1.0	1.0	0.984	1.0	0.938

Disper. = Dispersion; Peirce = Peirce-III;
GK Max = Goodman & Kruskal Maximum

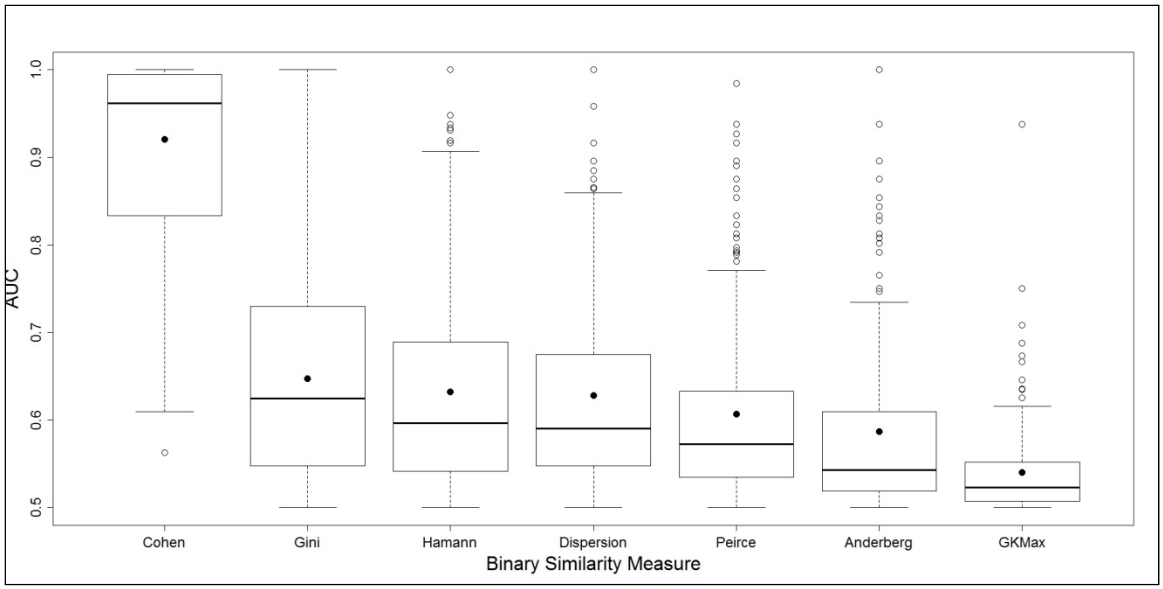

Figure V-9 Boxplots of 7 Similarity Measures' AUC Values

5.6.1 Cohen's Kappa.

One method of comparing sources' dyad nominations is interrater reliability. Several indices of interrater reliability have been created, as partially enumerated in Table V-8, though a common ratio has taken prominence in usage. Cohen's Kappa, Equation (5.4), utilizes a comparison of a given index's value, I_0, compared against the index value expected by chance selection by both parties, I_e. Cohen's Kappa, $\hat{\kappa}$, is equal to one in the case of complete agreement among the two sources, and will be greater than or equal to zero if the observed agreement is greater than expected chance agreement. In the case of the observed agreement being less than the expected chance agreement, Cohen's Kappa will be negative, bounded below at negative one, though in some cases the minimum obtainable values will be between negative one and zero. (Fleiss, Levin, & Paik, 2003, pp. 603-604).

> For most purposes, values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance. (Fleiss, Levin, & Paik, 2003, p. 604)

$$\hat{\kappa} = \frac{I_0 - I_e}{1 - I_e} \tag{5.4}$$

A number of indices of interrater reliability are available for computation of Cohen's Kappa based upon the comparison matrix as depicted in Table II-14. Conveniently, the commonly used indices displayed in Table V-8 all reduce to the same value of Cohen's Kappa as identified in Equation (5.5) (Fleiss, Levin, & Paik, 2003, p. 603).

Table V-8 Some Indices of Interrater Reliability

Interrater Reliability Index	Formula
Overall Proportion of Agreement	$p_0 = a + d$
Proportion of Specific Agreement (ignoring *d*)	$p_s = \frac{2a}{2a + b + c}$
Proportion of Specific Agreement (ignoring *a*)	$p_s' = \frac{2d}{2a + b + c}$
Averaged Proportion of Specific Agreement (Rogot and Goldberg [1966])	$A = \frac{1}{2}(p_s + p_s')$
Goodman's and Kruskal's Index of Agreement (1954)	$\lambda_r = \frac{2a - (b + c)}{2a + (b + c)}$

(Fleiss, Levin, & Paik, 2003, pp. 599-602)

$$\hat{\kappa} = \frac{2(ad - bc)}{(a + b)(b + d) + (c + d)(a + c)} \tag{5.5}$$

5.7 Examining the DOE Factors.

The chosen experimental design allows for analysis of the four experimental factors impact upon the methodology's performance in terms of AUC. The four factors, labeled A through D, and there associated coded values are presented in Table V-9. First, as described in Section 3.4.1, the response variable, AUC, needs to be transformed so the regression equation's predicted value falls within the [0.5, 1.0] range.

Table V-9 DOE Factors

Factors		-1	1
Network Size	A	200	1,500
# of Sources (% network size)	B	5%	10%
% Reliable Sources	C	60%	80%
Sampling %	D	10%	20%

5.7.1 Analysis of Variance (ANOVA).

The initial regression model incorporates the four factors' main effects and all interaction terms. The resulting model, Equation (5.6), whose ANOVA table is displayed in Table V-10, computed via R (2011), possesses a R^2 of 0.608 and an adjusted-R^2 of 0.590.

$$\begin{aligned} y = {} & 2.85 - 0.53A - 0.22B - 1.82C + 0.85D + 0.21A{:}B - 0.12A{:}C \\ & - 0.02B{:}C - 0.04A{:}D - 0.07B{:}D - 0.43C{:}D + 0.05A{:}B{:}C + 0.03A{:}B{:}D \\ & - 0.35A{:}C{:}D - 0.08B{:}C{:}D + 0.08A{:}B{:}C{:}D \end{aligned} \tag{5.6}$$

Table V-10 DOE Factors Full Model ANOVA

	DoF	Sum Sq	Mean Sq	F score	p-value
A	1	74.95	74.95	35.12	0.000
B	1	14.81	14.81	6.94	0.009
C	1	760.85	760.85	356.54	0.000
D	1	155.39	155.39	72.82	0.000
A:B	1	6.53	6.53	3.06	0.081
A:C	1	3.22	3.22	1.51	0.220
B:C	1	0.35	0.35	0.16	0.687
A:D	1	0.55	0.55	0.26	0.612
B:D	1	1.05	1.05	0.49	0.483
C:D	1	32.64	32.64	15.30	0.000
A:B:C	1	0.47	0.47	0.22	0.640
A:B:D	1	0.12	0.12	0.06	0.809
A:C:D	1	19.91	19.91	9.33	0.002
B:C:D	1	0.93	0.93	0.44	0.510
A:B:C:D	1	0.96	0.96	0.45	0.503
Error	324	691.41	2.13		
	$R^2 =$	0.608		adj. $R^2 =$	0.590

A reduced model is created by using backwards stepwise regression based on Akaike's Information Criteria (AIC), presented in Equation (5.7) with p representing the number of parameters (Crawley, 2007, pp. 353-354). Variables are removed that improve the overall model's AIC, while considering model hierarchy by keeping variables' lower order terms if their higher order interactions are currently in the model.

$$AIC = -2(\text{log-likelihood}) + 2(p+1) \tag{5.7}$$

The reduced model contained all four main effects, five of the six possible two-factor interactions, and two of the four possible three-factor interactions, as shown in Table V-11. The four-factor interaction was not present in the reduced model. The reduced model resulted in a R^2 of 0.606 and an adjusted-R^2 of 0.595, and since the

independent variables are orthogonal in a full factorial design, the coefficients are the same as in Equation (5.6). However, for both the full model and reduced models based on the initial DOE factors, the normality plots of the residuals show significant departures from normality.

Table V-11 DOE Factors Reduced Model ANOVA

	DoF	Sum Sq	Mean Sq	F score	p-value
A	1	74.95	74.95	35.59	0.000
B	1	14.81	14.81	7.03	0.008
C	1	760.85	760.85	361.27	0.000
D	1	155.39	155.39	73.78	0.000
A:B	1	6.53	6.53	3.10	0.079
A:C	1	3.22	3.22	1.53	0.217
A:D	1	0.59	0.59	0.28	0.598
C:D	1	32.95	32.95	15.65	0.000
A:C:D	1	19.86	19.86	9.43	0.002
Error	330	694.99	2.11		
	R^2 =	0.606		adj. R^2 =	0.595

5.7.2 Analysis of Covariance (ANCOVA).

The poor fit, in terms of error distribution assumptions, of the linear regression models based on the four factors comprising the DOE, suggests other variables are necessary to explain the methodology's performance.

5.7.2.1 Concomitant Variables.

Several SNA network measures were investigated for potential inclusion as concomitant variables into the regression model: density, diameter, mean path length (MeanPath), characteristic path length (CharPath), average clustering coefficient (AvgCC), degree correlation (DegreeCorr), and the degree distribution's power-law

exponent (alpha). These SNA measures were selected due to their common appearance and familiarity within the SNA literature. These measures were computed on all of the graphs representing the true underlying social networks used in the experimentation.

As with linear regression, multicollinearity can adversely affect ANCOVA results. An advantage of using the DOE approach, the factors represented in the full factorial design, center points, and space filling design are near-orthogonal by design; therefore, multicollinearity is not a concern among them. However, it is necessary to check the concomitant variables for multicollinearity. Several of the variables are expected be correlated with others. Density's definition is related to the number of nodes present in the graph, and thus can be eliminated from consideration due to correlation with number of nodes factor. The mean path length and the characteristic path length differ only slightly in definition, by use of a grand average as opposed to a median. Thus, the remaining concomitant variables are plotted pairwise, Figure V-10, and their corresponding correlations are computed, Table V-12, to investigate if multicollinearity is present.

As demonstrated by Figure V-10 and Table V-12, strong multicollinearity exists among several variables. As expected, the mean path length and the characteristic path length exhibit strong correlation, and both are strongly negatively correlated with the graph diameter. Thus, only one of these variables is required. For this analysis the mean path length was selected as it possesses the strongest positive and negative correlation, with the characteristic path length and the diameter, respectively. Additionally, the average clustering coefficient is eliminated as a potential concomitant variable as it is strongly correlated with the mean path length. The degree correlation is moderately

negatively correlated with the mean path length, but remained in this analysis as a concomitant variable. The examination of seven commonly applied SNA network measures resulted in only three measures being utilized in the ANCOVA: mean path length, degree correlation and the degree distribution's power-law exponent.

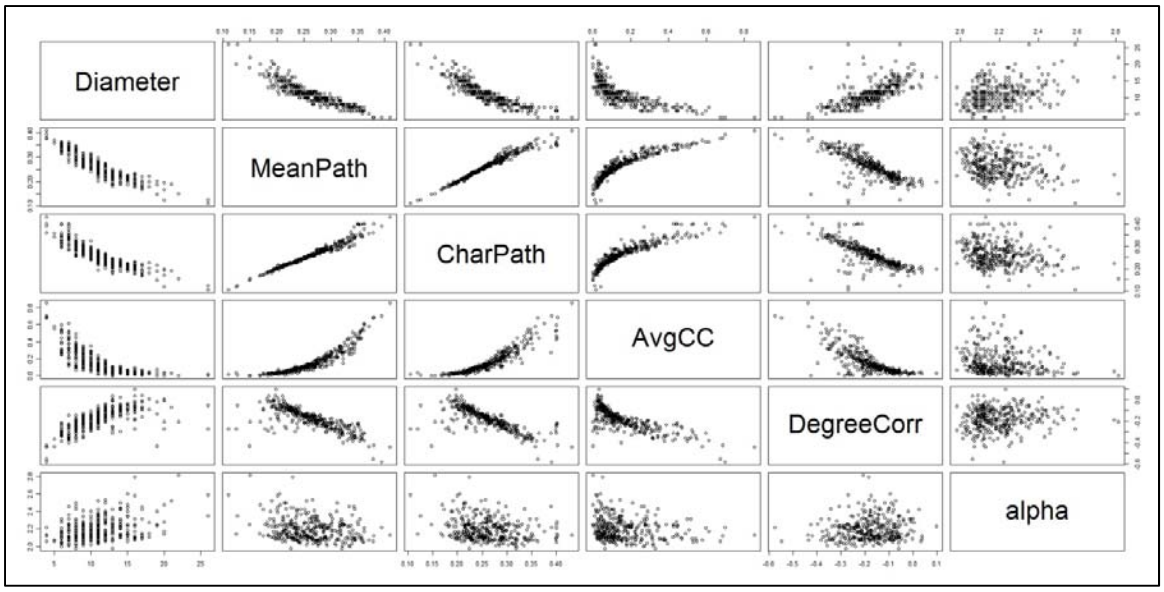

Figure V-10 Concomitant Variables Plot

Table V-12 Concomitant Variables Correlation

	Diameter	MeanPath	CharPath	AvgCC	DegreeCorr	alpha
Diameter	1.0	-0.907	-0.876	-0.738	0.677	0.346
MeanPath	-0.907	1.0	0.978	0.900	-0.757	-0.248
CharPath	-0.876	0.978	1.0	0.907	-0.747	-0.249
AvgCC	-0.738	0.900	0.907	1.0	-0.767	-0.177
DegreeCorr	0.677	-0.757	-0.747	-0.767	1.0	0.097
alpha	0.346	-0.248	-0.249	-0.177	0.097	1.0

5.7.2.2 ANCOVA Results.

The concomitant variables were all transformed to design space, i.e. centered and scaled, and the resultant ANCOVA model possessed a R^2 of 0.868 and adjusted-R^2 of 0.790, displaying substantial improvement of the ANOVA model's R^2 of 0.610 and adjusted-R^2 of 0.592. Performing backwards stepwise regression based on AIC, as was done for the ANOVA model, the reduced ANCOVA model only the single seven-factor and several six-factor interaction terms containing all factors and all concomitant variables. Introduction of the concomitant variables, substantially improved the explanatory power of the model, implying the graph to graph variation even with design runs has significant impact upon the methodology's performance.

5.7.2.3 ANCOVA Linear Regression Model Issues.

Despite the positive improvements in R^2 and adjusted-R^2 by incorporating the concomitant variables into the model, diagnostic analysis of the ANCOVA linear model expresses major concerns of assumption validity. The normality plot of the residuals, Figure V-11, appears to significantly depart from a normal distribution. The standardized residuals were tested for normalilty, with a preset critical p-value of 0.05, using the Anderson-Darling test (Anderson & Darling, 1952) available in the R nortest package (Gross, 2006), which resulted in an A test statistic value of 5.025 and an associated p-value of 1.9×10^{-12}, and the Shapiro-Wilk test (Shapiro & Wilk, 1965), which generated a W statistic of 0.942 and an associated p-value of 2.8×10^{-10}. This violates the assumption that the error terms are distributed on a standard normal distribution, $\varepsilon \sim N(0, \sigma^2)$.

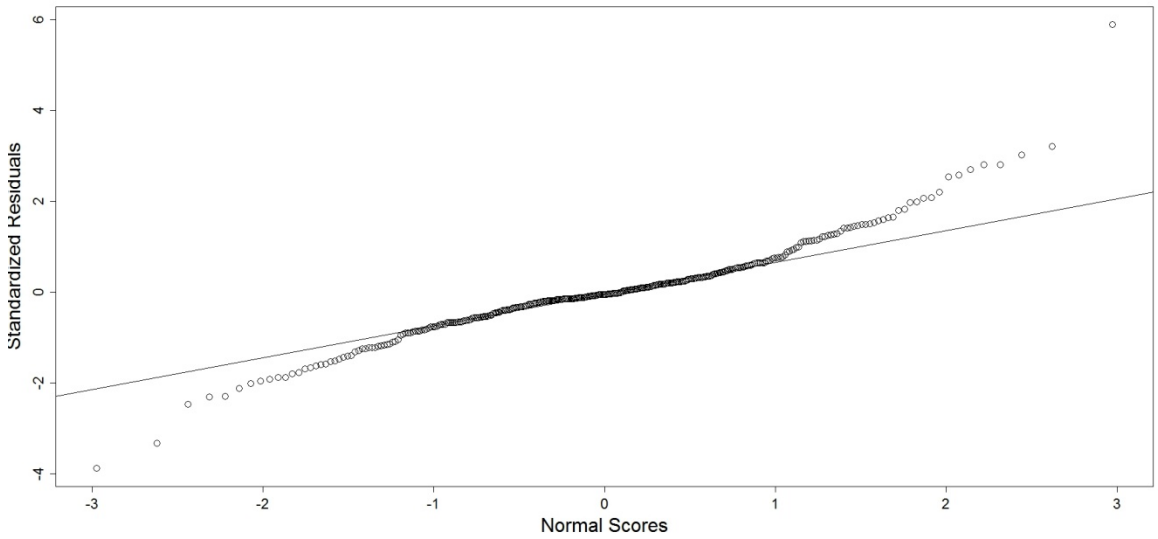

Figure V-11 Residuals Normal Plot from Full ANCOVA model

Additionally, there are concerns of the linear regression homoscedasticity assumption being violated. Figure V-12 depicts the fitted values versus the residuals illustrating that the variance does not appear to be constant. Figure V-12 also displays the presence of outliers, noted by the large residual deviations which are greater than three standard deviations.

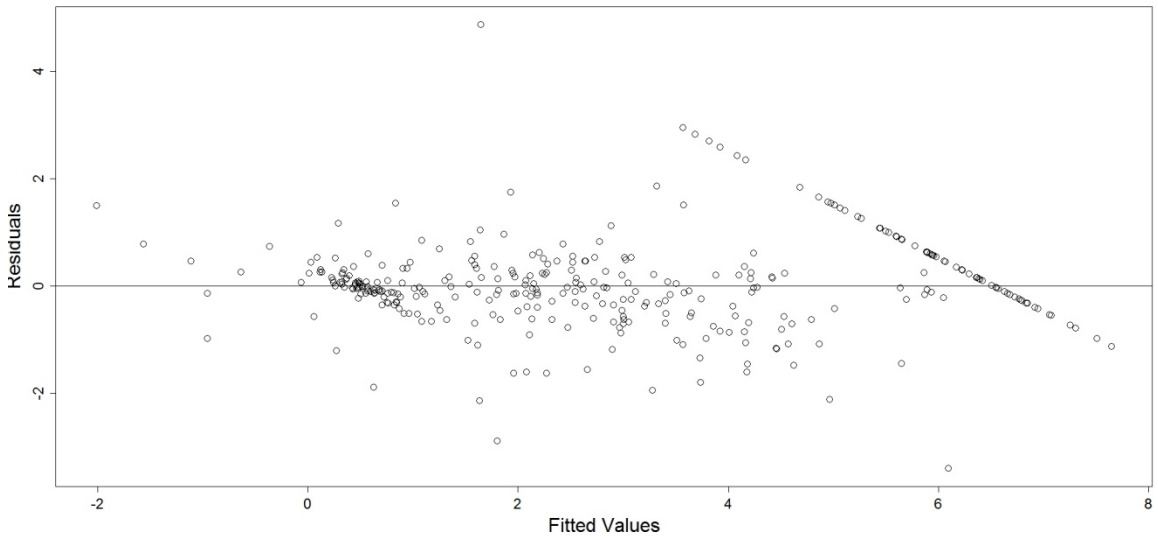

Figure V-12 Fitted Values vs. Residuals

5.7.2.4 Box-Cox Method.

The Box-Cox method was applied to attempt to correct the normality deficiencies and the heteroscedasticity (Montgomery, Peck, & Vining, 2006, p. 171). The response variable was examined across 100 points for the domain of [-2, 2] which determined the optimal value maximizing the log-likelihood function to be 1.43. This implies transforming the response variable according to Equations (5.8) and (5.9) produces the full model with greatest likelihood of explaining the data (Montgomery, Peck, & Vining, 2006, p. 171).

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases} \tag{5.8}$$

$$\dot{y} = \ln^{-1}\left(1/n \sum_{i=1}^{n} \ln y_i\right) \tag{5.9}$$

As Figure V-13 depicts, the log-likelihood function is optimized at the maximum investigate value, 2.0, with the 95% confidence interval indicated. The log-likelihood

function demonstrates the diminishing returns for transforming the response variable with its subsequent increase in model complexity and loss of interpretability. Additionally, performing the response variable transform as described in Equations (5.8) and (5.9) does not correct the non-normality of the error terms or the heteroscedasticity.

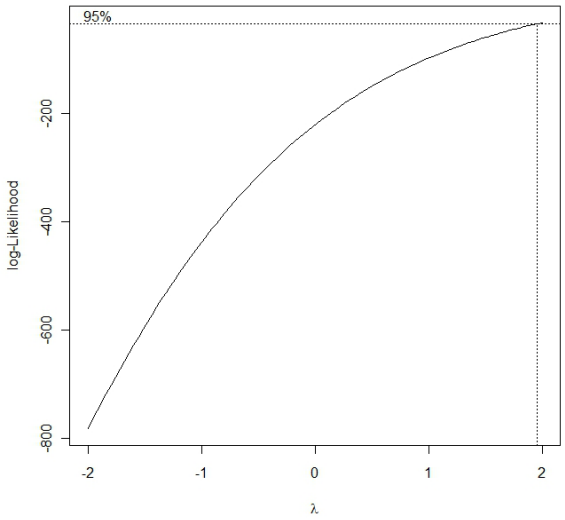

Figure V-13 Box-Cox Method of Log-Likelihood Maximization

5.7.2.5 Principal Component Regression.

Principal component regression was performed in an attempt to correct the error terms' non-normality and the heteroscedasticity. As multicollinearity was present on the covariates describing graph characteristics, converting the covariates into principal component scores will eliminate the multicollinearity (Montgomery, Peck, & Vining, 2006, pp. 355-357). The covariates, as well as factor A, which represents the number of nodes, were transformed into principal component scores via the orthogonal principal component loadings. This technique transforms the seven covariates and the single factor

into eight orthogonal components. The principal component loadings are presented in Table V-13. Performing the linear regression using the three DOE remaining factors and the eight variables based on principal component scores still exhibited non-normal distributed error terms and heteroscedasticity.

Table V-13 PCA Regression Loadings

	Component							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
numNodes	-0.12	0.60	0.06	-0.10	0.07	0.75	-0.22	-0.06
Density	0.22	-0.70	-0.06	0.06	-0.29	0.57	-0.22	-0.06
Diameter	-0.43	-0.05	-0.12	-0.59	-0.48	0.09	0.43	0.17
MeanPath	0.45	0.17	-0.02	0.21	-0.13	0.12	0.25	0.79
CharPath	0.44	0.19	-0.02	0.14	-0.24	0.10	0.59	-0.57
AvgCC	0.42	0.26	-0.11	-0.35	-0.50	-0.29	-0.53	-0.07
DegreeCorr	-0.40	0.12	0.27	0.62	-0.60	-0.08	-0.11	-0.01
alpha	-0.15	0.09	-0.95	0.27	-0.02	0.00	-0.04	-0.02
SS loadings	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Proportion Var	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%	12.5%
Cumulative Var	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%	87.5%	100.0%

5.7.2.6 Factor Regression.

Factor regression was also performed to attempt to correct the error terms' non-normality and the heteroscedasticity. Factor regression, similarly to principal components regression, reduces multicollinearity by transforming some of the covariates into latent factors (Curtis, 1976). Linear regression is then performed on the remaining covariates and the latent factors. Equation (5.10) displays the factor regression equation with dependent variable y, covariates X in design matrix format including intercept, regression coefficients β, factor loadings matrix Λ, and factors f, and error term ε (Carvalho, Chang, Lucas, Nevins, Wang, & West, 2008, p. 1439).

$$y = \beta X + \Lambda f + \varepsilon \quad (5.10)$$

Applying factor regression in this experiment, the concomitant variables and factor A, the number of nodes, were converted into Thomson regression factor scores, the alternative being the Bartlett factor scores which "treats the specific factors as random errors (Bartholomew, Deary, & Lawn, 2009, p. 577)." Three factors were kept according to the Kaiser-Guttman rule which keeps factors whose eigenvalues are greater than one. The factor loadings are presented in Table V-14. Linear regression using the three remaining DOE factors and the three variables based on factor analysis scores of the concomitant variables still exhibited non-normal distributed error terms and heteroscedasticity.

Table V-14 Factor Regression Loadings

	Factor 1	**Factor 2**	**Factor 3**
numNodes	-0.02	0.94	-0.12
Density	0.16	-0.92	0.17
Diameter	-0.67	0.17	-0.64
MeanPath	0.86	-0.07	0.49
CharPath	0.88	-0.04	0.44
AvgCC	0.96	0.01	0.15
DegreeCorr	-0.79	0.40	-0.10
alpha	-0.11	0.10	-0.32
SS loadings	3.544	1.937	1.02
Proportion Var	44.3%	24.2%	12.8%
Cumulative Var	44.3%	68.5%	81.3%

5.8 Quantile Regression

Due the failure of the techniques described above to correct the heteroscedasticity and non-normality of the error term's distribution, an alternative statistical approach than multivariate linear regression is required. As quantile regression does not assume a

distribution on the error terms, it was selected as the statistical analytic technique to investigate the four design factors and the concomitant variables.

Despite the failure to address the multiple linear regression model's assumption via the various techniques described in Section 5.7.2, some insights can be garnered from the analysis. The introduction of concomitant variables that address the underlying networks' characteristics substantially improved the R^2 and adjusted-R^2 values for the multivariate linear regression model. For the quantile regression analysis, it was decided to incorporate the same concomitant variables as in Section 5.7.2, to explore the effects of the design factors coupled with graph characteristics interactions. The following quantile regression analysis was conducted in R utilizing the quantreg package (Koenker, 2011).

5.8.1 Median Regression Model.

Quantile regression was used to create a median regression model with all of the design factors, denoted "factor QR model", and all possible interactions as regressors. The pseudo-R^2 was 0.522. Backwards stepwise regression was attempted, using improving AIC as a removal criteria; though, no terms were eliminated from the model. The factor QR model's coefficients are presented in Table V-15. The factor QR model identifies all of the main effects as statistically significant, as well as several higher order interaction terms.

Table V-15 Factor QR Model Coefficients

	Coefficient	Lower CI	Upper CI	Significant (α = .05)
(Intercept)	2.76	2.55	2.84	yes
A	-0.44	-0.48	-0.18	yes
B	-0.29	-0.38	-0.10	yes
C	-2.16	-2.26	-1.95	yes
D	0.73	0.64	0.98	yes
A:B	0.28	0.06	0.31	yes
A:C	0.29	0.08	0.35	yes
B:C	0.15	0.00	0.25	yes
A:D	0.12	-0.01	0.25	
B:D	0.08	-0.14	0.18	
C:D	-0.52	-0.75	-0.48	yes
A:B:C	-0.13	-0.19	0.09	
A:B:D	-0.07	-0.21	0.05	
A:C:D	-0.27	-0.35	-0.09	yes
B:C:D	-0.36	-0.40	-0.13	yes
A:B:C:D	0.34	0.16	0.44	yes

5.8.1.1 Factor QR Model Interpretation.

In contrast to the linear regression models and the other statistically techniques employed in Section 5.7, statistical significance of the regressor coefficients can be assessed. This is possible due to satisfying quantile regression's relaxed error assumptions, in comparison to the OLS regression attempts.

With the four-factor interaction deemed statistically significant, all interaction terms are included in the factor QR model due to the principle of model hierarchy. Interpreting the factor QR model begins by analyzing the main effects' coefficients. The strongest effect is factor C, the percentage of sources that are reliable. Factor C's coefficient sign is counter-intuitive; one would expect that with increasing percentages of reliable sources, the more accurately the methodology would perform. This counter-

intuition could result from factor C's involvement with several other statistically significant interaction terms. Factor C's wrong sign could also be an indication of omitting important regressors or some of the regressors' ranges may be too narrow (Montgomery, Peck, & Vining, 2006, p. 112).

There may be other explanations for Factor C's coefficient's unexpected sign. As the percentage of reliable sources increases within a set of information sources, it may become more difficult to correctly identify the few unreliable sources. It may be easier to correctly discern members of two approximately equally sized populations, than to identify correct associations of two imbalanced populations. Other classifiers have been noted in the literature to have degraded performance as a result of this phenomenon and it is an active research area (Japkowicz & Stephen, 2002, pp. 429, 432).

Another potential explanation involves the experimentation. Reliable sources were generated under a mechanism that intentionally causes them to fail to report relationships among actors to better mimic real world social network reporting empirical findings. As increasing percentages of reliable sources are present, each reliable source's minor errors can increase the difficulty of distinguishing unreliable sources who report greater amounts of erroneous information.

Factor D, the source sampling percentage, is the other design factor whose impact on the methodology can be anticipated. Factor D possess a positive coefficient, indicating at increased levels of sampling, the better the methodology performs. As the source sampling percentage impacts the amount of information given by each source, one would expect that the more information a source provides, the easier it is to correctly

classify them as reliable or unreliable. This intuition is reflected in factor D's positive coefficient.

Factors A's and B's main effects both possess negative coefficients, indicating that for their main effects' contributions increasing values diminish the methodology's performance. Factor A, the network size, indicates that larger networks inhibit correct assessment of source reliability. With larger networks, it is more likely the information sources are reporting on different aspects of the networks. This spreading of the sources' reporting across the network increases the chances that sources are commenting on unique relationships, which are not subject to confirmatory or disputing reporting. Factor B, the number of sources, indicates that the more sources reporting on the network, the more difficult it is to properly distinguish reliable from unreliable sources—more sources, more potential disagreement.

Several two-factor, three-factor interaction effects, along with the four-factor interaction, are statistically significant. This complexity of the median regression model is not surprising as one can imagine multiple scenarios for which interaction effects would affect. For instance, having many information sources (factor B) reporting little amounts of information (factor D), coupled with a large majority of the sources are reliable (factor C) may improve the methodology's performance. In fact, interaction BCD's effect coefficient is negative which aligns with the scenario just described.

5.8.1.2 Median Regression with Covariates.

Quantile regression was applied to create a median regression model encompassing all of the design factors, concomitant variables, and possible interactions.

This "full QR model" possesses a pseudo-R^2 value of 0.753. When backwards stepwise regression was attempted, the full model proved to possess the lowest AIC.

Extending the full QR model and factor WR model to every integer percentile within (5, 95), i.e. create individual quantile regression models for every percentile in the specified range, the relative-R^2 was plotted and is displayed in Figure V-14. Additionally, the pseudo-R^2 was computed for both the full and factor models and is displayed in Figure V-15. These figures indicate that the full QR model displayed substantial improvement over the factor QR model for all tested quantiles.

Figure V-14 displays the superiority in explanatory power of the full QR model over that of the factor QR model. The full QR model shows substantial improvement for the extreme quantile models. Figure V-15 displays the full QR model exhibiting more consistent model explanatory power across the various explored quantiles. The factor QR model appears inconsistent with notable decreases in pseudo-R^2 for the extreme quantile regression models.

The complete listing of the median regression coefficients and their 95% confidence interval bounds are provided in Table B-1 in Appendix A. Table V-16 provides the coefficients which are significant at the $\alpha = 0.05$ level, along with the associated 95% confidence interval upper and lower bounds, with interaction effects denoted with a ":" separating the variables. Examining the significant factors highlights the importance of the underlying network graph's structural characteristics on the methodology's performance. Only six terms, the intercept, factors B and C, the AB interaction, the BD interaction, and the ABD interaction do not involve a graph structural characteristic. All of the other significant terms, with the sole exception of alpha

exhibiting a main effect, are composed of design factors interactions with the concomitant graph characteristics.

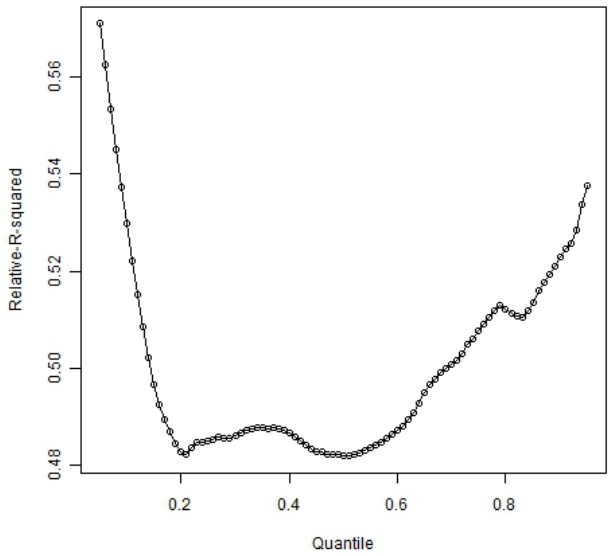

Figure V-14 Relative-R^2 of Full QR Model and Factor QR Model

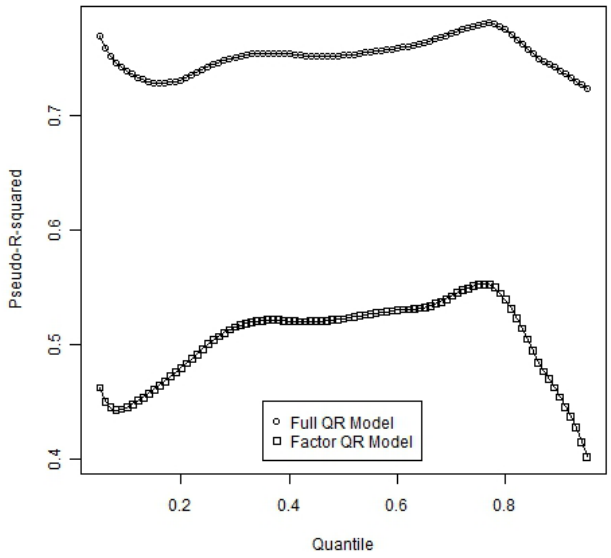

Figure V-15 Pseudo-R^2 of Full QR and Factor QR Models

The prominence of the concomitant variables as significant is indicative of the importance of underlying network structures on the algorithm's performance. Similar results were exhibited by the improvement in the R^2 and adjusted-R^2 in the ANCOVA analysis. However, as discussed in Section 2.10.1, the ANCOVA error distribution assumptions are violated and thus no definitive conclusions can be drawn from the ANCOVA analysis. Quantile regression's assumptions are met and provide conclusive statistical evidence for this data set that the underlying network structure and characteristics impact the algorithm's performance.

5.8.2 Quantile Regression Coefficients.

Next, an examination of the quantile regression coefficients was conducted. Quantile regression models were created for each integer percentile between the 9th and 91st quantiles. The large confidence intervals for the extreme quantiles, as found in this data set, limit their utility in examining their effects. Bootstrapping was then applied to generate 95% confidence intervals on each coefficient across the 81 quantiles. The ordinary least squares (OLS) coefficient with its 95% confidence levels were also computed according to the full model.

What follows is a series of figures depicting the quantile regression coefficients as computed across the 81 quantiles, denoted as dark points, and their associated 95% confidence intervals, identified by a gray area. Additionally for each regression term, the OLS coefficient and its associated 95% confidence interval is depicted. These are identifiable by a horizontal solid line with two dotted lines above and below marking the

Table V-16 Median Regression Significant (α = 0.05) Factors

	Coefficient	Lower CI	Upper CI
(Intercept)	2.55	2.43	2.70
A	-0.44	-0.61	-0.31
C	-1.97	-2.16	-1.84
D	0.67	0.53	0.91
MeanPath	-0.98	-1.18	-0.67
alpha	0.29	0.20	0.55
DegreeCorr	-0.69	-0.91	-0.36
A:C	0.19	0.08	0.39
C:D	-0.46	-0.66	-0.22
B:MeanPath	-0.40	-0.70	-0.04
C:MeanPath	0.28	0.06	0.54
A:DegreeCorr	0.62	0.04	0.82
A:B:MeanPath	0.58	0.05	0.98
A:C:MeanPath	0.61	0.22	1.02
B:C:MeanPath	-0.49	-0.79	-0.08
A:B:alpha	0.36	0.07	0.58
A:C:alpha	-0.40	-0.51	-0.04
A:D:alpha	0.37	0.18	0.51
B:D:alpha	0.28	0.10	0.44
A:C:DegreeCorr	0.43	0.10	0.70
A:D:DegreeCorr	0.35	0.12	0.68
B:D:DegreeCorr	0.45	0.11	0.81
C:D:DegreeCorr	-0.62	-0.98	-0.29
A:B:C:alpha	-0.30	-0.53	0.00
A:C:D:DegreeCorr	0.51	0.26	0.88
A:MeanPath:alpha:DegreeCorr	0.44	0.01	0.71
A:B:D:MeanPath:alpha	-0.77	-1.48	-0.28
A:C:D:MeanPath:alpha	-0.78	-1.33	-0.22
A:B:C:D:DegreeCorr	-0.49	-0.80	-0.10
B:C:D:MeanPath:DegreeCorr	-0.54	-0.94	-0.11
A:B:D:alpha:DegreeCorr	-0.81	-1.54	-0.30
B:C:D:alpha:DegreeCorr	0.75	0.15	1.57
C:D:MeanPath:alpha:DegreeCorr	0.41	0.10	0.74
A:B:C:D:MeanPath:DegreeCorr	0.78	0.30	1.06
A:B:C:MeanPath:alpha:DegreeCorr	-0.61	-1.23	-0.21
A:B:C:D:MeanPath:alpha:DegreeCorr	0.43	0.05	1.02

extent of the 95% confidence interval. The y-axis represents the regressor's coefficient value and the x-axis represents the quantile used in the quantile regression model.

Figure V-16 and Figure V-17 display the quantile regression main effect coefficients including the intercept. These figures highlight a capability of quantile regression when an analyst is interested in the tail behavior of a distribution. Notably, for this experimentation, of interest are the significant factors when the algorithm performs poorly. Thus, we are interested in the lower end of the response variable's distribution and the associated significant variables.

Examining the upper right quadrant in Figure V-16 which represents the coefficient associated with factor A, node size of the "true network graph", the OLS factor A coefficient appears to be significant at the $\alpha = 0.05$ level. Similarly in the median regression, factor A's coefficient is assessed as significant at the $\alpha = 0.05$ level across most quantile regression models. However, as depicted in Figure V-16, factor A's significance may be in question for lower quantile models. This implies the factor A may not possess a statistically significant effect on the methodology when it performs worse. Ignoring higher order interaction effects for the moment, when the algorithm performed worse, factor A possesses a statistically insignificant negative coefficient and when the algorithm exhibited good performance, factor A possesses a statistically significant negative coefficient with a stronger effect.

This ability to examine regressors' effects in the tail behavior of the response variable's distribution exemplifies a capability advantage of quantile regression over OLS regression. In this experimentation, the response variable's tail behavior corresponds to

when the methodology performs worse, which assists in characterizing worst case conditions for employment of the methodology.

Factor B, the number of sources appearing in the lower left panel of Figure V-16, appears to have a negative impact upon the response variable, again not taking into account the higher order interaction terms at this time. Factor B's coefficient seems to decrease for the higher performance levels of the algorithm. The coefficient decrease actually increases its effect, which is negative, which Factor B has on the response variable, implying that in the cases of superior algorithm performance Factor B may be more statistically significant.

Factor C, the percentage of sources that are reliable, displayed in the lower right panel of Figure V-16, has a negative effect on the methodology's proficiency at all levels of performance, which could result from class imbalance as previously discussed. The coefficient of Factor C, which varies for the various examined quantile regression models, is always statistically significant. Factor C exhibits greater influence on the methodology's performance when it performs well and becomes less influential as an explanatory factor when the methodology performs poorly, not accounting for higher order interactions which will be examined shortly.

Factor D, the sampling percentage, displayed in the upper left panel of Figure V-17, appears to be a significant factor at all levels of methodology performance. As the sampling percentage reflects the amount of information each source is providing, it is intuitive that this factor contributes positively to the methodology's performance. At higher levels of methodology performance, factor D's coefficient appears to be

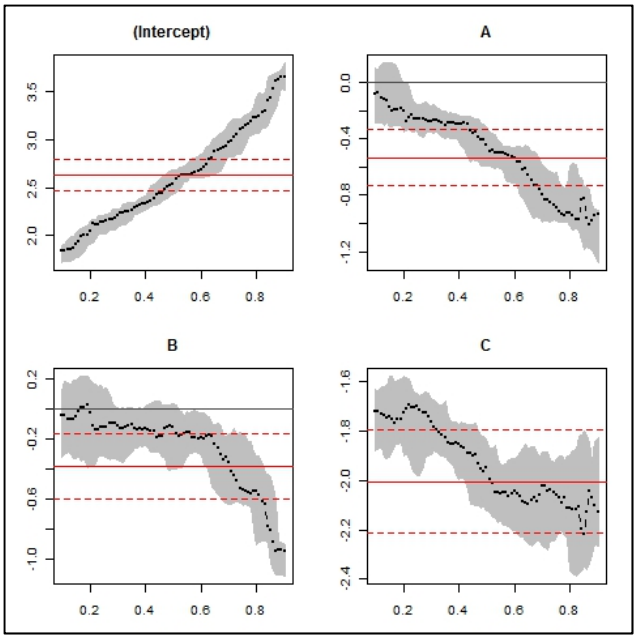

Figure V-16 Intercept and Main Effects' Coefficients' Confidence Intervals

increasing, implying factor D's increasing influence in the model; although, the confidence interval on the coefficient is also expanding.

The concomitant variables, mean path length and the degree correlation, appear to possess negative main effects that are statistically significant, Figure V-17. The concomitant computed alpha exponent of a power-law degree distribution variable, appears to possess a statistically significant positive main effect at almost all quantiles, as shown in the lower left panel of Figure V-17.

Proceeding in a similar fashion for the higher order interaction terms in the quantile regression models results in Table B-1 in Appendix A, which depicts the independent variables and interaction terms, their associated effects and their statistical significance at $\alpha = 0.05$. Since this is investigating multiple quantile regression models computed for a range of quantiles, the effects and statistical significance is summarized

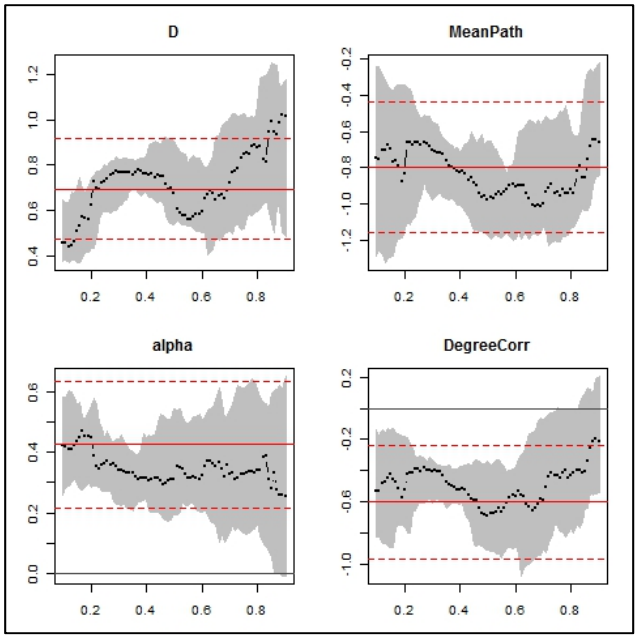

Figure V-17 Main Effects' Coefficients' Confidence Intervals (cont.)

by visual inspection of the graphs. Variables' and interaction terms' effects are assessed as positive or negative if the majority of the regressor's coefficients are observed to lie above or below zero on the y-axis, respectively. If a regressor appears to possess positive and negative coefficients for a substantial number of quantiles, the variable's effect was noted as mixed. A variable's statistical significance was noted as present, if, for at least a substantial number of the quantile regression models, the confidence interval of the coefficient did not overlap zero on the y-axis.

This visual analysis of the variables' effects across multiple quantile regression models is summarized in Table C-1 in Appendix C. Appendix D contains the plots, Figure D-1 through Figure D-18, upon which the visual inspection was accomplished. The plots depict each factor's or interactions' quantile regression coefficients across the range of examined quantiles.

Examining the variables' effects across multiple quantile regression models allows inference that spans across the quantile regression models. The statistical significance of higher order interaction terms is a testament to the complex interactions and effects, similar to observations in the ANCOVA model of Section 5.7.2. Multiple concomitant variables interactions with the design factors are statistically significant across the majority of quantile regression models, independent of the specific quantile. Table V-17 highlights regressors' whose coefficients appeared to be significant across a substantial number of quantiles.

Table V-17 Significant Regressors Across All Quantiles

Regressor	Effect		Regressor	Effect
(Intercept)	Positive		C:D:DegreeCorr	Negative
A	Negative		A:B:C:D	Mixed
B	Negative		B:C:D:MeanPath	Negative
C	Negative		A:C:D:alpha	Negative
D	Positive		A:C:D:DegreeCorr	Positive
MeanPath	Negative		A:MeanPath:alpha:DegreeCorr	Positive
alpha	Positive		B:C:D:MeanPath:DegreeCorr	Negative
DegreeCorr	Negative		A:B:D:alpha:DegreeCorr	Negative
A:C	Positive		A:C:D:alpha:DegreeCorr	Mixed
C:D	Negative		A:C:MeanPath:alpha:DegreeCorr	Negative
B:MeanPath	Negative		B:C:MeanPath:alpha:DegreeCorr	Negative
A:DegreeCorr	Positive		A:D:MeanPath:alpha:DegreeCorr	Mixed
B:C:MeanPath	Negative		C:D:MeanPath:alpha:DegreeCorr	Mixed
B:D:MeanPath	Positive		A:B:C:D:MeanPath:DegreeCorr	Positive
A:C:alpha	Negative		A:B:C:MeanPath:alpha:DegreeCorr	Negative
A:D:alpha	Positive		A:B:D:MeanPath:alpha:DegreeCorr	Mixed
A:C:DegreeCorr	Positive		A:C:D:MeanPath:alpha:DegreeCorr	Positive
A:D:DegreeCorr	Positive		A:B:C:D:MeanPath:alpha:DegreeCorr	Positive
B:D:DegreeCorr	Positive			

5.9 Analysis Summary.

The traditional statistical measure Fleiss' Kappa was employed to assess inter-rater reliability. As summarized in Table IV-1, the correlation between Fleiss' Kappa and the percentage of reliable sources was weakly positive. This illustrates some of the complexity involved with assessing social network information sources. Fleiss' Kappa only characterizes the entire collection of sources and in this experimentation showed poor performance in detecting the presence of unreliable sources in a collection of social network information sources.

As characterizing the entire collection of sources proved to be inadequate, pairwise source comparisons were conducted to determine source veracity. The methodology presented in this chapter, utilizing Cohen's Kappa, displayed good performance in correctly distinguishing between reliable and unreliable sources. Six other binary similarity measures were tested and performed worse than Cohen's Kappa.

Conducting the methodology with Cohen's Kappa resulted in a median AUC of 0.962 and an average of 0.921, indicating a strong ability to discriminate between reliable and unreliable information sources. The methodology appears robust to the variety expressed in the experimental design factors, with over 75% of examined cases delivering an AUC of greater than 0.83.

Detailed examination of the methodology's performance was conducted by statistical analysis of the DOE. The four DOE design factors' influence on methodology performance was initially analyzed with ANOVA. Unfortunately, there appeared to be substantial departures from normality and the presence of heteroscedasticity. This violation of assumptions precludes any statistical inference from the ANOVA.

Attempting to correct the deviations from the ANOVA assumptions, ANCOVA was employed, incorporating SNA measures that characterize the underlying social network structure. Seven SNA network measures were initially identified, but only three could be utilized due to multicollinearity among them. Accounting for this graph to graph variation did not improve the ordinary least squares regression model's adherence to its fundamental assumptions. Common statistical response variable transform techniques were applied, but to no avail. Continuing to incorporate the concomitant variables, PCA and factor regression were conducted to attempt to fit the ordinary least squares regression model while satisfying the statistical assumptions, though these techniques were unsuccessful.

Finally, quantile regression was applied, for its semi-parametric approach makes the error term's distribution irrelevant. A median regression model using the original design factors (factor QR model) was constructed, as well as another model containing the design factors and the three concomitant variables (full QR model). The models were compared, with the full QR model displaying superior performance for all tested quantiles in terms of relative-R^2. The pseudo-R^2 was found to be 0.753. The full QR model was extended to construct quantile regression models for all integer percentiles between the 9^{th} and 91^{st} quantiles. This allowed characterizing the regressors' impact across the spectrum of the methodology's performance. This enabled investigating regressors that contribute to good or poor performance of the methodology.

5.10 Analysis Results.

With the selection of the full quantile regression model as an appropriate model for the experimental data set, analysis was conducted examining the coefficients of the regressors. The model is complex with four design variables with an additional three concomitant variables, and the associated interaction terms. Despite this complexity, examining quantile regression models over a spectrum of quantiles enables generalizations to be made. The main effects for the design factors and the concomitant variables are statistically significant for most of the quantile regression models. As displayed in Table V-17, numerous interaction effects are statistically significant. Of interest, is that the majority of these regressors contain at least one of the concomitant variables as a component of the interaction.

The design variables' influence in the methodology meets expectations of their importance in assessing sources' reliability. It is of no surprise that the number of sources, the number of reliable sources, and the amount of source provided information are critical in discerning unreliable sources. However, the substantial improvements obtained in the models' goodness of fit with inclusion of network structural characteristics, implies that the underlying social network graph is an important consideration to the methodology's performance.

5.11 SNA Practical Results

These results show that for the factors examined in the experimentation, the methodology provides a quantitative technique for assessing social network information sources that displays good performance. The experimental design investigated factor

levels likely to be experienced by SNA analysts facing real world problems. These results are only valid for social network analytic conditions that are within the factors' ranges explored in the experimentation, but the DOE could easily be augmented to account for increased factors' ranges or even additional factors. The results provided in this chapter have only been shown to hold for the conditions provided in the experimental design. Due to the acceptable quantile regression model fit, interpolating the results to other points within the design space should provide similar results with little risk of extreme methodology performance deviance. Of course, employing the methodology under conditions that lie outside the design space investigated here assumes greater risk in actual methodology performance.

SNA analysts can employ the methodology to quantitatively assess social network information sources which combined with expert opinion and other subjective factors may substantial improve the likelihood of correctly identifying unreliable sources and excluding their information from the social network model. The methodology enables SNA analysts to quantitatively score the amount of concordance between two information sources by applying one or more binary similarity measures. SNA analysts can account for varying information sources perspectives of the social network via a weighting of pairwise comparisons among the sources based upon their overlap in reporting. Information sources can be quantitatively grouped into reliable and unreliable clusters via the weighted MDS and fuzzy clustering. The methodology provides a mechanism to provide visual representations of the concordance among the information sources' reports via the weighted MDS visualization. This visualization may highlight discrepancies or patterns that initiate further SNA analyst investigations and inquiries that

without application of the methodology developed in this research would potentially be overlooked.

5.12 Chapter Summary

This chapter developed the source comparison methodology with an example of the methodology being presented. The methodology was tested according to the experimental design detailed in Section 3.3. The results of the experimentation were analyzed using a variety of standard statistical techniques and nontraditional techniques that were introduced in Section 2.10. The next chapter presents an employment of the methodology in a case study format.

VI. Case Study

This chapter utilizes the developed methodology on an example based on real world data to demonstrate its effectiveness and utility for SNA analysts. First, a description of the real world data set is provided. Next, the experimentation process is discussed. Following, the methodology developed in this dissertation is employed against the case study. Finally, widely used SNA measures are applied to compare and contrast the results stemming from the different social network models to estimate the impact of imperfect information upon the analysis.

6.1 Data Set Description

This data set examined in this case study derives from Natarajan (2006). The data set is composed of a social network model generated via English translations of over 2,000 pages of a 1993 court case's transcripts documenting 2,408 conversations derived from the wiretapping of 21 phones (Natarajan, 2006, p. 176). The social network is a representation of

> "an international drug trafficking conspiracy, with links to mafia families, which had acquired $144 million in assets. The organization was said to be responsible for transporting, receiving and selling more than 200 hundred [sic] kilograms of heroin per year (approximately 193 kg were recovered during the investigation) (Natarajan, 2006, p. 173)."

Out of the total 2,408 conversations, only in 1,851 conversations were individuals able to be identified by name, resulting in 294 participants in the social network of interest. Out of the 294 identifiable participants, only 86 of the actors spoke with at least two members of the network. These 86 actors were further reduced to 38 "core

members…who had two or more contacts and were involved in five or more conversations (Natarajan, 2006, p. 179)."

Natarajan (2006) conducted a role analysis based upon a content analysis of a random sample of conversations among the 38 core members. Natarajan determined that the 38 core members performed one of four roles detailed in Table VI-1, although "individuals sometimes switched roles in the furtherance of particular deals (Natarajan, 2006, p. 187)." The description of each of these roles follows.

Table VI-1 Core Members' Role Composition

Role	Number of Core Members
Sellers	18
Retailers	8
Brokers	8
Secretaries	4
Total	**38**

(Natarajan, 2006, pp. 180-181)

> Sellers are "individuals who were mostly involved in selling drugs in quite large quantities (a pound or more), but who were also involved to a more limited extent in brokering deals (Natarajan, 2006, p. 180)."
>
> Retailers bought exclusively from sellers…[and] most deals involved between a quarter of an ounce to four ounces (Natarajan, 2006, p. 180).
>
> Brokers were middlemen between sellers as well as between sellers and retailers. They set up meetings to inspect drugs and helped negotiate the price (Natarajan, 2006, p. 181).
>
> Secretaries are women, "generally wives or girlfriends of the active members. They passed messages to buyers, sellers and brokers. They were well known to those involved in buying and selling (Natarajan, 2006, p. 181)."

The 38 core actors, labeled A through LL with their corresponding gender honorific, are displayed in Figure VI-1 as constructed by Natarajan (2006). Natarajan specifically identified the actors performing as sellers and brokers. The secretaries were assumed to represent the females depicted in Figure VI-1 in accordance with the description of the secretaries role. One female, Ms. Q, specifically identified by Natarajan as a seller, left the four remaining females, Ms. E, Ms. G, Ms. H, and Ms. CC, were assumed to be secretaries. The remaining network members were assessed as retailers. Table VI-2 lists the actors' assessed roles.

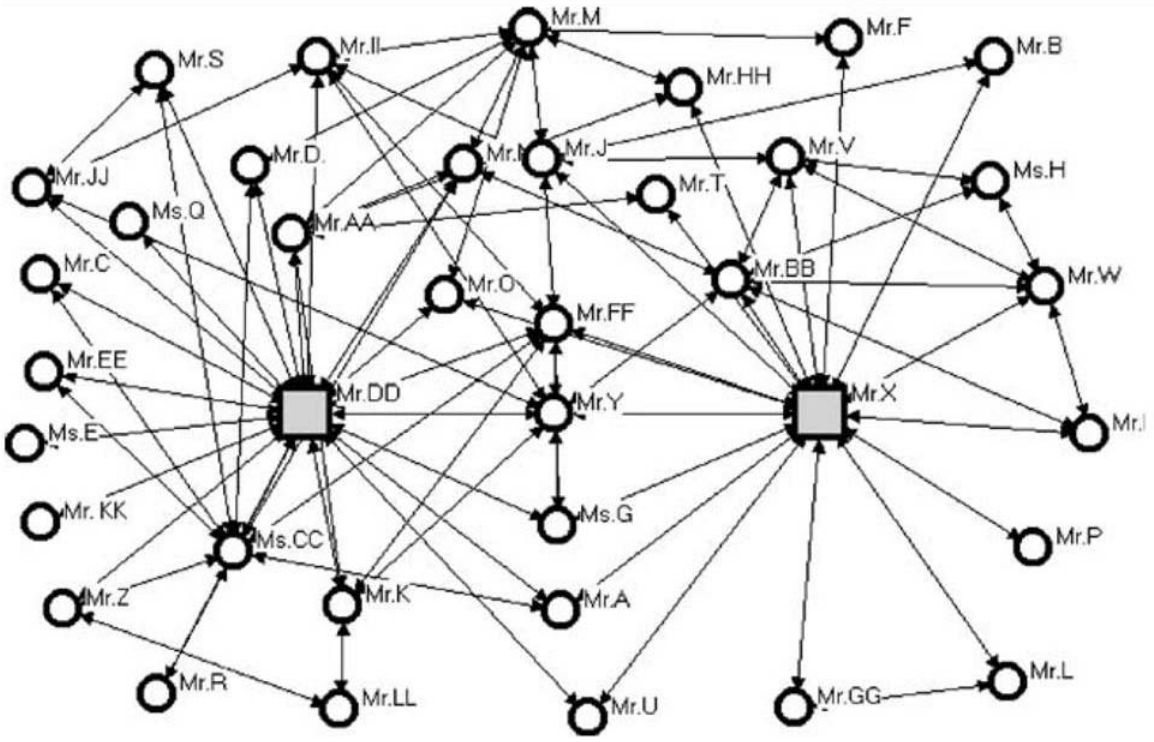

Figure VI-1 38 Core Members (Natarajan, 2006, p. 184)

Table VI-2 Core Members' Roles

Role	Actors
Sellers:	F, J, K, L, M, W, X, Y, Z, AA, BB, DD, EE, FF, GG, HH, JJ, LL
Retailers:	B, C, I, N, O, T, II, KK
Brokers:	A, D, P, Q, R, S, U, V (Q is identified as female in Figure VI-1.)
Secretaries:	E, G, H, CC (all identified as female in Figure VI-1.)

(Natarajan, 2006, pp. 184-185)

6.2 Experimentation

Despite the wealth of information contained in the numerous transcripts utilized in the trial of 35 defendants, this data set does not represent the information requirements faced by analysts, but reflects the results of an analysis. Analysts must sift through large amounts of data and discard irrelevant information. The information contained in the transcripts resulted from wiretaps, but even this monitoring is susceptible to generating imperfect social network information.

> [W]iretap interceptions must be authorized by the relevant court upon a detailed showing of probable cause. Specifically, the investigating officer must provide a detailed affidavit showing there is probable cause to believe the phone is being used to facilitate a specific, serious, indictable crime. Undoubtedly, chance will help to determine which phones are targeted by the investigators and there is no guarantee that they will succeed in identifying all the phones involved, or even the most critical ones (Natarajan, 2006, p. 177).

With this data set, the only information available is that which was deemed suitable for prosecution. The information that was discarded during the analysis phase of the investigation is unavailable. Thus, this case study, potentially, only presents a partial representation of the information used to construct the social network model. Additionally, with the data derived from wiretaps, it is likely that other interactions among the network's members were not observed and therefore, not incorporated into the

information used for the prosecutions. As a result, the social network model is a representation of the underlying relationships and interactions among members of the network and ultimately, the accuracy of its portrayal is unknown. However, this social network model does represent one of the few existing, publically available, dark network data sets. As such, this social network is used to demonstrate the employment of the developed methodology.

6.2.1 Generating Information Sources.

The social network of the 38 core members provided by Natarajan (2006) will serve as the true underlying social network. Hypothetical information sources were assumed and representative reporting was constructed. Unreliable sources will be generated as that information is unavailable from the original case. The methodology was employed to determine if it could correctly identify and distinguish the reliable information sources from the unreliable information sources.

6.2.2 Reliable Information Sources.

As mentioned in Section 2.3.2.4, dark network members practice OPSEC techniques and procedures to frustrate governmental efforts to inhibit the network's operations. For this case study, we will assume that the organization's members are carrying out good OPSEC practices. However, in this case, there are actors in the social network who are not members of the organization. Identified by Natarajan (2006) as secretaries, who are "wives or girlfriends of the active members (Natarajan, 2006, p. 181)", it can be expected that they may not employ OPSEC practices to the full extent as the organizations' members. For this case study, it will be assumed that the secretaries

are the weak link to the organization and this vulnerability is exploited by law enforcement. Thus, Ms. E, Ms. G, Ms. H, and Ms. CC, will serve as the reliable information sources. In this case, the secretaries do not have to be voluntary reliable information sources, but merely the source on which reliable information is being reported. We will assume that the secretaries' phones are being wiretapped and due to their poor practice of OPSEC, information regarding the underlying structure of the social network is being revealed. However, because of their positions as secretaries in the organization, there is a limit to the information that can be garnered from surveilling them. The four secretaries, as reliable information sources, reflect an approximate 10% of the overall organization composed of 38 actors.

6.2.2.1 Direct Relationships.

As the secretaries are being monitored, it can be assumed that their direct relationships with members of the network are obtainable. This assumes difficulties faced by law enforcement, such as speaking in foreign languages or in code, are not effective in obscuring organizational participation (Natarajan, 2006, p. 178). Thus any actor directly connected to one of the secretaries will be reported by that corresponding information source in this demonstration.

6.2.2.2 Indirect Relationships.

As the secretaries are wives and girlfriends of active members, it is not inconceivable that they would be aware of other network actors connected to their corresponding boyfriends and husbands. It is not reported in Natarajan (2006) which actors are the secretaries' boyfriends and husbands. Table VI-3 depicts the assumed

significant other(s) for each of the secretaries, noting that some secretaries possess several significant others. Any actor directly connected to a secretaries' significant other has a chance of being known and reported by the secretary. Or in this case of information based on wiretaps, the significant other's relationship with another member of the organization is discussed on a phone call with the secretary. For this research, these secretaries' indirect relationships will be probabilistically reported. A direct relationship between a secretary's significant other(s) and another member of the network will be reported by the secretary with probability 0.5.

Table VI-3 Secretaries' Significant Others

Secretary	Significant Other(s)
E:	DD
G:	X, Y, DD
H:	V, W, BB
CC:	DD

6.2.2.3 Triad Closure.

In some case, the secretary is directly connected to an actor who is also directly connected to the secretary's significant other. In these cases, it is more likely that the relationship between the actor and the significant other is known by the secretary or discussed in conversation. This sociological observed increased likelihood of triad closure was the inspiration for the clustering coefficients discussed in Section 2.2.1.9. In this case study, it will be assumed that relationships that constitute the third leg of a triad

for a secretary and her significant other will be reported with probability of 0.75, a substantial increase over the indirect relationship of 0.5.

6.2.2.4 Reliable Source Reporting.

Table VI-4 summarizes the probability of the true underlying social network relationships being reported. The probabilities ensure that each secretary's reporting is a reflection of her network perspective. For this reporting generation method, a secretary can only report relationships that are at most, two steps away from her.

Table VI-4 Relationship Reporting Probabilities

Relationship	Probability
Direct	1.0
Indirect	0.5
Triad Closure	0.75
All Others	0.0

Using the probabilities specified in Table VI-4, reports of the underlying social network were generated for each secretary and are displayed as edge lists in Figure VI-2 and visualized in Figure VI-3 through Figure VI-6.

E		G		H		CC	
E	DD	A	X	H	V	A	CC
K	DD	B	X	H	W	C	CC
O	DD	D	DD	H	BB	C	DD
Q	DD	E	DD	I	W	D	CC
Y	DD	G	X	J	V	D	DD
Z	DD	G	Y	I	BB	G	DD
AA	DD	G	DD	V	W	N	CC
CC	DD	N	DD	V	BB	O	DD
DD	II	O	X	W	BB	R	CC
DD	JJ	O	DD	X	BB	S	CC
		P	X			S	DD
		Q	DD			Z	CC
		V	X			CC	DD
		W	X			CC	EE
		X	Y			CC	FF
		X	BB			DD	EE
		Z	DD			DD	FF
		DD	JJ			DD	II
		DD	KK			DD	JJ

Figure VI-2 Edge List of Each Secretary's Reports

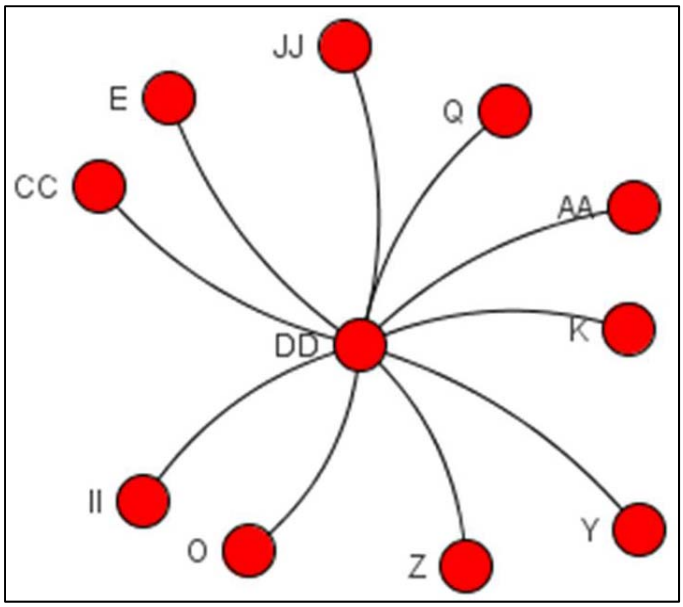

Figure VI-3 Visualization of Ms. E's Report

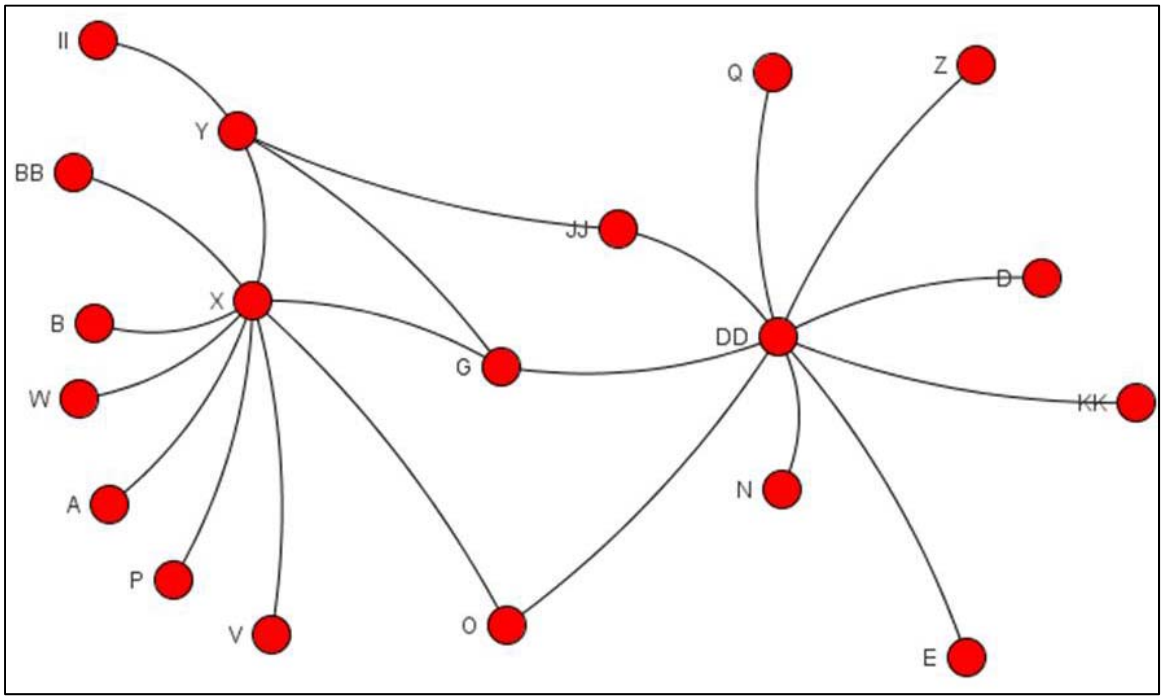

Figure VI-4 Visualization of Ms. G's Report

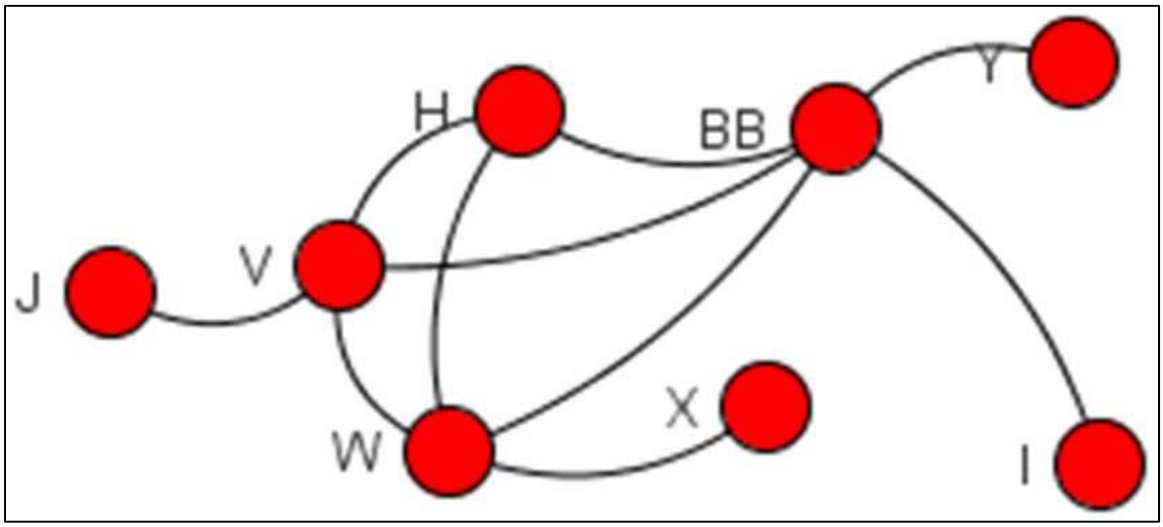

Figure VI-5 Visualization of Ms. H's Report

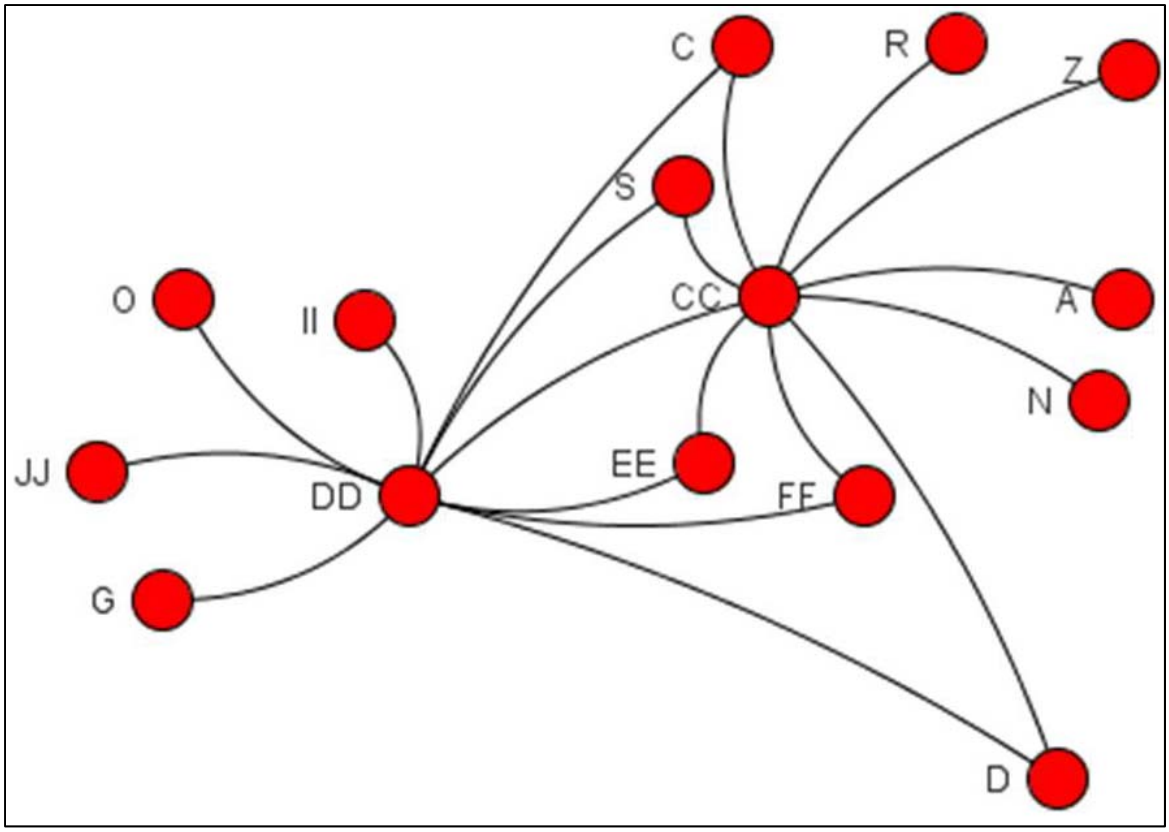

Figure VI-6 Visualization of Ms. CC's Report

6.2.3 Unreliable Information Sources.

Unreliable sources need to be randomly generated as any detected in the real world data set were discarded and excluded from presentation at trial. With four reliable sources, it was decided to generate two unreliable sources, U1 and U2, resulting in one third of the information sources being unreliable. The unreliable sources should appear similar to reliable sources. The number of edges reported by each secretary is presented in Table VI-5. The two unreliable sources were selected to report 32 and 21 edges by a random drawing of two integers from the uniform distribution on the range [14 , 44],

bounded by the minimum and maximum number of edges reported by the reliable information sources.

Table VI-5 Number of Reported Edges by Secretary

Secretary	Number of Reported Edges
E	20
G	44
H	14
CC	29

With the number of edges specified for each of the two unreliable sources, U1 and U2, random edges to be reported as present were generated by uniformly at random selecting the first actor from the set of 38 actors and then uniformly at random selecting a second actor from the remaining 37 actors. This process produced the unreliable sources reporting the edge lists displayed in Figure VI-7 and the visualizations depicted in Figure VI-8 and Figure VI-9.

U1		U2	
Z	P	W	H
O	K	Q	GG
EE	P	H	F
F	DD	Z	M
F	FF	Z	A
J	U	W	U
R	I	LL	E
U	J	G	HH
FF	J	W	HH
BB	KK	GG	JJ
L	N	J	DD
AA	GG	Z	I
T	EE	Y	U
Q	D	P	L
B	CC	J	F
DD	AA	H	J
HH	F	C	FF
EE	KK	JJ	R
O	W	D	O
U	KK	K	DD
V	Q	KK	W
N	D		
V	P		
V	JJ		
B	F		
Z	M		
A	II		
T	K		
DD	X		
HH	G		
Z	I		
B	JJ		

Figure VI-7 Edge List of Unreliable Sources' Reports

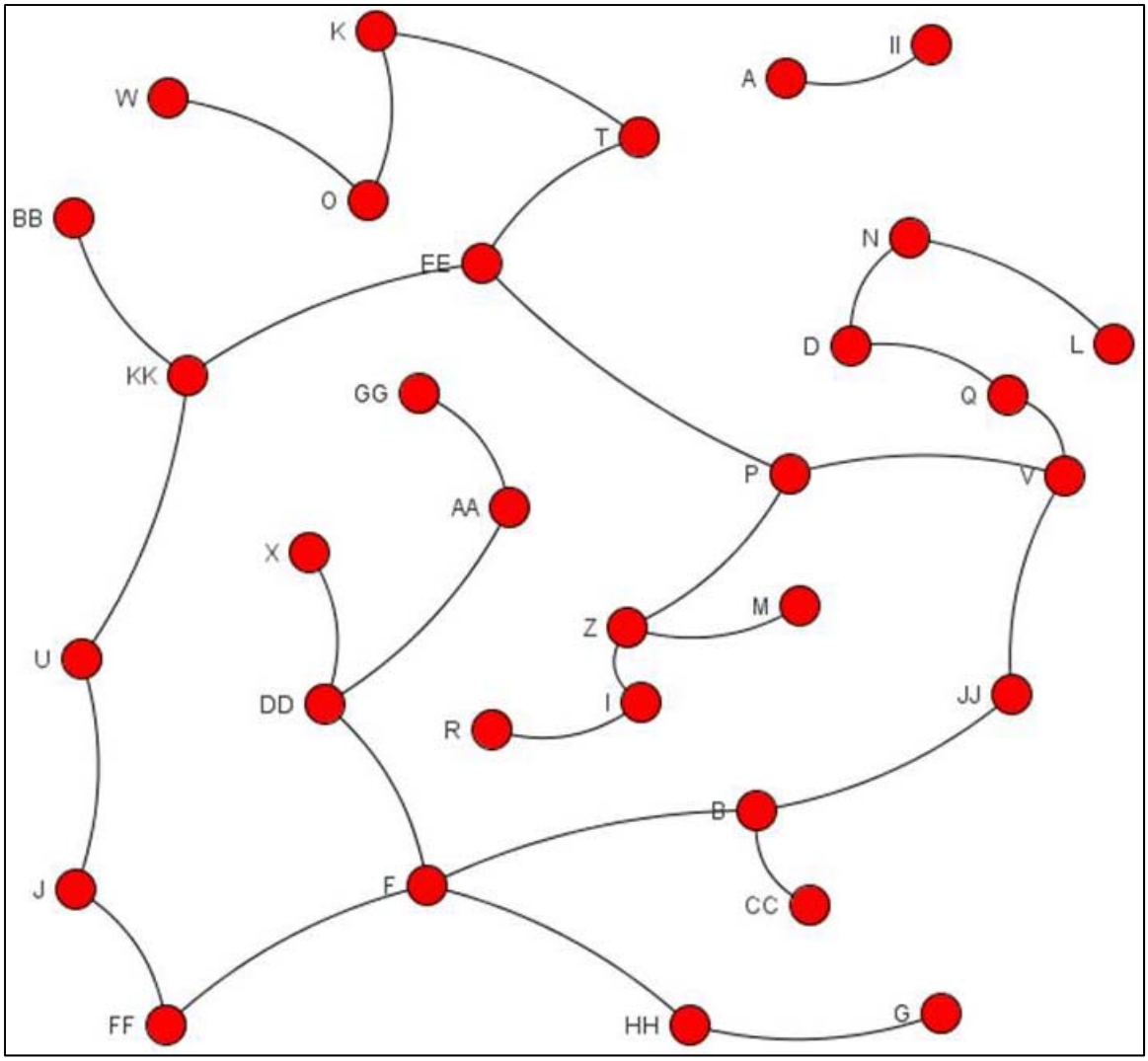

Figure VI-8 Visualization of U1's Report

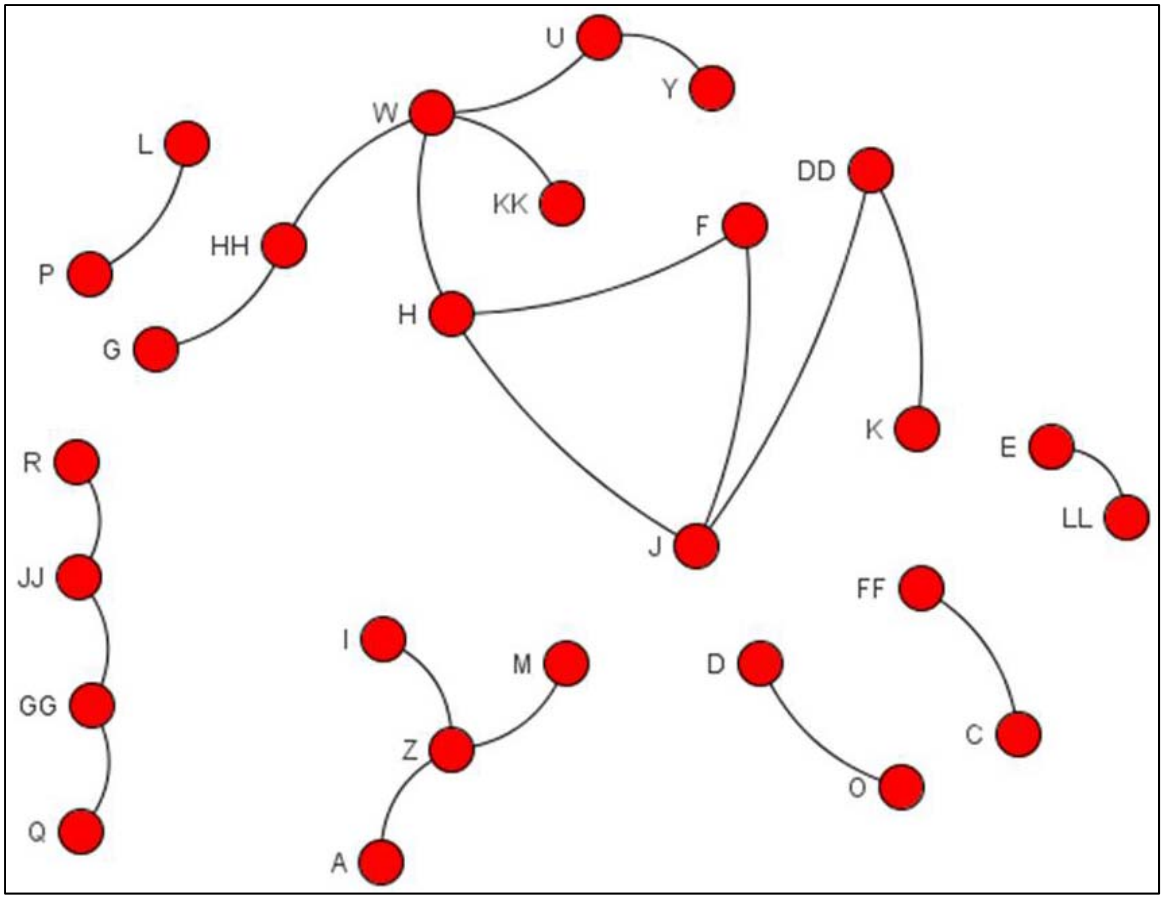

Figure VI-9 Visualization of U2's Report

6.3 Methodology Employment

The methodology developed in Chapter V was implemented using the binary similarity measure selected in Chapter IV, Cohen's kappa. Table VI-6 displays the computed pairwise dissimilarity scores and Table VI-7 displays the source weightings.

Table VI-6 Source Dissimilarity Scores

	E	G	H	CC	U1	U2
E		0.38	1.5	0.3	0.88	0.8
G	0.38		1.4	0.32	1.11	1.06
H	1.5	1.4		1.5	1	0.38
CC	0.3	0.32	1.5		1.05	1.12
U1	0.88	1.11	1	1.05		0.85
U2	0.8	1.06	0.38	1.12	0.85	

Table VI-7 Source Weightings

	E	G	H	CC	U1	U2
E		0.36	0	0.3	0.26	0.27
G	0.36		0.23	0.36	0.5	0.39
H	0	0.23		0	0.18	0.17
CC	0.3	0.36	0		0.38	0.31
U1	0.26	0.5	0.18	0.38		0.59
U2	0.27	0.39	0.17	0.31	0.59	

6.3.1 Inferences Based on Sources' Scores and Weights.

By examining the sources' scores found in Table VI-6, several inferences can be drawn. Information source E appears to strongly agree with the reports made by sources G and CC and disagrees with U1's and U2's reports. Source H is reporting on distinctly different parts of the social network compared against sources E and CC, identifiable due to the minimum weighting of zero between sources H and E and sources H and CC in Table VI-7. Source CC possesses a similar pattern and reporting relationships as source E. Source G follows the patterns of sources E and CC with one exception. Source G is in strong disagreement with source H, with a dissimilarity score of 1.4, and their reporting has weak to moderate overlap with a weighting of 0.23. This leads to the

inference that sources E, G, and CC are concordant and are providing confirmation of each other's reports. With those three sources being identified as concordant, an SNA analyst is likely to assess sources E, G, and CC as reliable. With source G being deemed a reliable source, its strong disagreement with source H implies source H is unreliable.

Information source U1 possess moderate to strong disagreement with all other reporting sources, as evidenced by its high dissimilarity scores in Table VI-6. Source U1's reporting is not unique and partially aligns with other information sources as its weightings are non-zero. As U1 is discordant with all other sources, it will likely be assessed as unreliable. U2 possesses high dissimilarity scores will all information sources with the exception of source H, though there is weak overlap with a weighting of 0.17, the weakest weighting in Table VI-7. Sources U2 and H appear to be concordant with each other, but discordant with all other information sources with which their social network reports overlap. Information source H's strong disagreement with source G implies that it is unreliable. Information source U2 is only concordant with source H, indicating that it is also an unreliable source. However, the weak weightings of information sources U2 and H with the other reporting sources signifies that they are reporting on different aspects of the social network.

In this case, the SNA analyst faces a dilemma. Information sources E, G, and CC are concordant and are reporting on similar aspects of the social network. Information source U1 is discordant with all other information sources and can be assessed as unreliable. Information sources U2 and H may be unreliable due to source H's strong disagreement with source G, or an alternative explanation is that U2 and H are reliable sources and E, G, and CC are unreliable. However, information sources U2 and H appear

to be reporting on different aspects of the social network in comparison to sources E, G, and CC.

6.3.2 Source Reporting Visualization.

The visualization developed in this research enables a SNA analyst to pictorially depict the information contained in Table VI-6 and Table VI-7. In this case study, the limited number of information sources allows for inspection of the source dissimilarity score and weightings matrix. With larger number of sources, the visualization may prove more useful to SNA analysts as large matrices become unmanageable and unwieldy for visual inspections and comparisons. Even with a limited number of information sources, as the six in this case study, the visualization can provide insight.

The weighted MDS visualization of the information sources is provided in Figure VI-10. As illustrated, the strong concordance of information sources E, G, and CC is readily apparent. U1's discordance with all other sources is visually depicted by its isolation on the graph. Sources U2 and H appear concordant, but provide an alternative depiction of the network, visible from their separation from the cluster containing sources E, G, and CC.

6.3.3 Information Sources Assessment.

Faced with the raw data presented in Figure VI-2 and Figure VI-7, SNA analysts' did not possess many options for assessing information sources based on quantitative methods before the introduction of the methodology presented in this dissertation. Applying the methodology against the real world data present in this case study, albeit

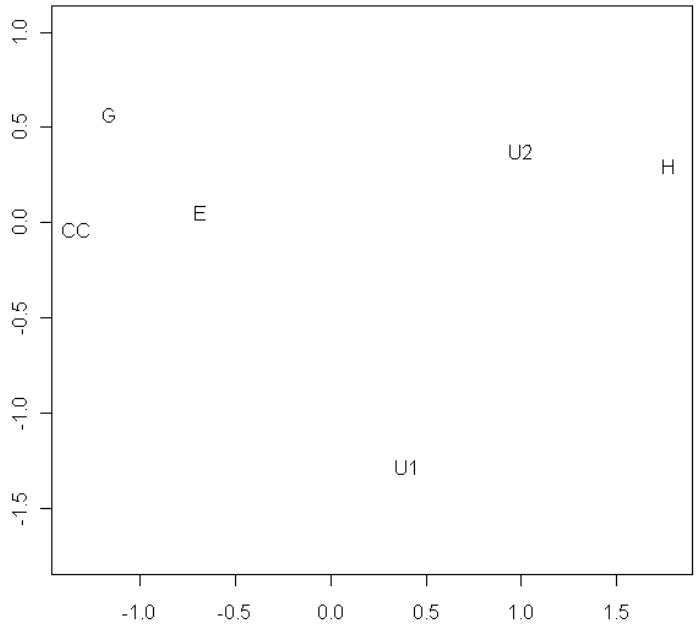

Figure VI-10 Visualization of Information Sources' Reporting

with hypothetical information sources, highlighted the developed methodology's capability to aid SNA analysts in quantitatively assessing information sources.

In this case study, the methodology quickly detected the unreliability of U1, which was a true unreliable source as a result of the generation mechanism. The methodology also quickly determined the strong concordance in reporting among information sources E, G, and CC. Information sources U2 and H were also concordant, though this is a function of the random generation method of source U2. In this case study, reliable sources G and H were shown to be discordant. As reliable information sources can provide errors in their reports, discordance among reliable sources can be expected to occur on real world data sets with some frequency.

One view of the results of this case study may be that not discerning source H as a definite reliable source is an error. However, the point of the methodology is to focus

SNA analysts' attention to discrepancies and patterns existing with the social network data collected. In this case study, the methodology identifies two sets of sources, generally reporting on different aspects of the social network, but when examining the minor overlap in their reporting, there exists strong disagreement. Utilizing the developed methodology, a SNA analyst conducting analysis on this case study quantitatively identifies the discrepancies in the information sources' reports, leading to further investigation, potentially of sources H and U2, to draw conclusions on the reliability of the sources.

6.4 SNA Impact of Imperfect Information

This section investigates the impact of utilizing imperfect social network data when conducting social network analysis. Social Network Analysis measures, described in detail in Section 2.2, are applied to various social network models to demonstrate the potential difference in analytical results and conclusions. Several social network models are used to compare results. The full 38 core member network from Figure VI-1 is used as the true underlying social network, denoted as the Ground Truth model. A social network model composed of all six information sources, an All Sources model, is used to represent the traditional SNA practice of utilizing all of the social network data provided by any information source. A Reliable Sources model, constructed from the four reliable information sources, Ms. E, Ms. G, Ms. H, and Ms. CC, represents the ideal case, the best model that can be constructed with this set of information sources. Finally, a social network model constructed using the three information sources, Ms. E, Ms. G, and Ms. CC, identified as reliable by utilizing the methodology. This model resulting from

application of the developed methodology is referred to as the Selected Sources model. Table VI-8 summarizes the four social network models being compared in this section, visualized in Figure VI-10 through Figure VI-14. Comparison among these social network models will demonstrate the difference in results from using all social network information sources and utilizing the methodology.

Table VI-8 Models' Description

Model	Description
Ground Truth	Social network model (Figure VI-1) as presented in Natarajan 2006. It represents the true underlying social network model for this demonstration.
All Sources	All sources' (E, G, H, CC, U1, and U2) reporting included. Inclusion of all social network information sources (in the absence of *a priori* information) is a traditional, common approach employed by SNA analysts.
Reliable Sources	Only reporting from E, G, H, and CC included. This is the best possible model from this collection of social network information sources. All reports from reliable information sources are included, and unreliable sources' reports are discarded.
Selected Sources	Only reporting from E, G, and CC included. This is the social network model constructed from applying the methodology and selecting information sources E, G, and CC as reliable.

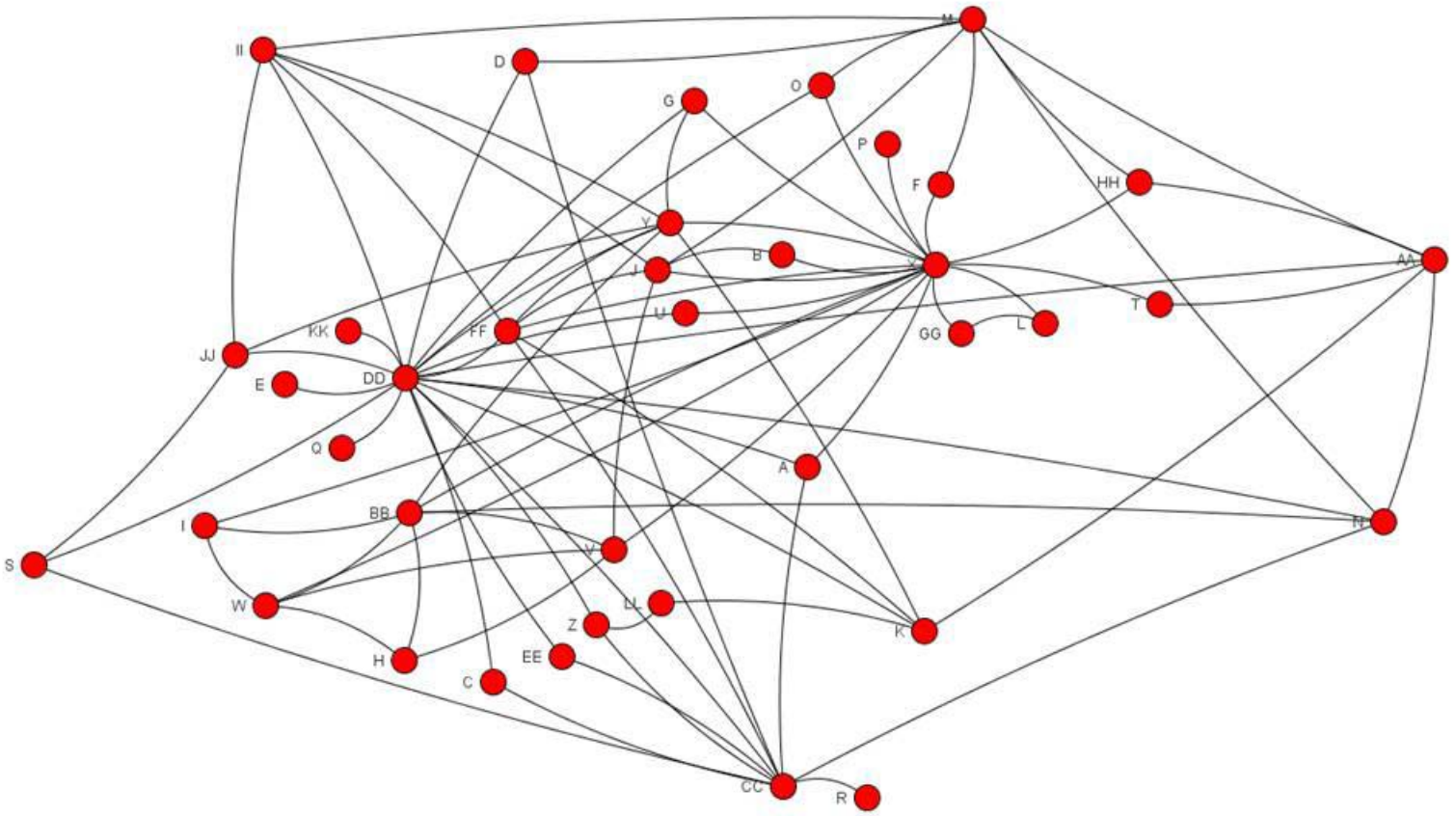

Figure VI-11 Visualization of Ground Truth Social Network Model

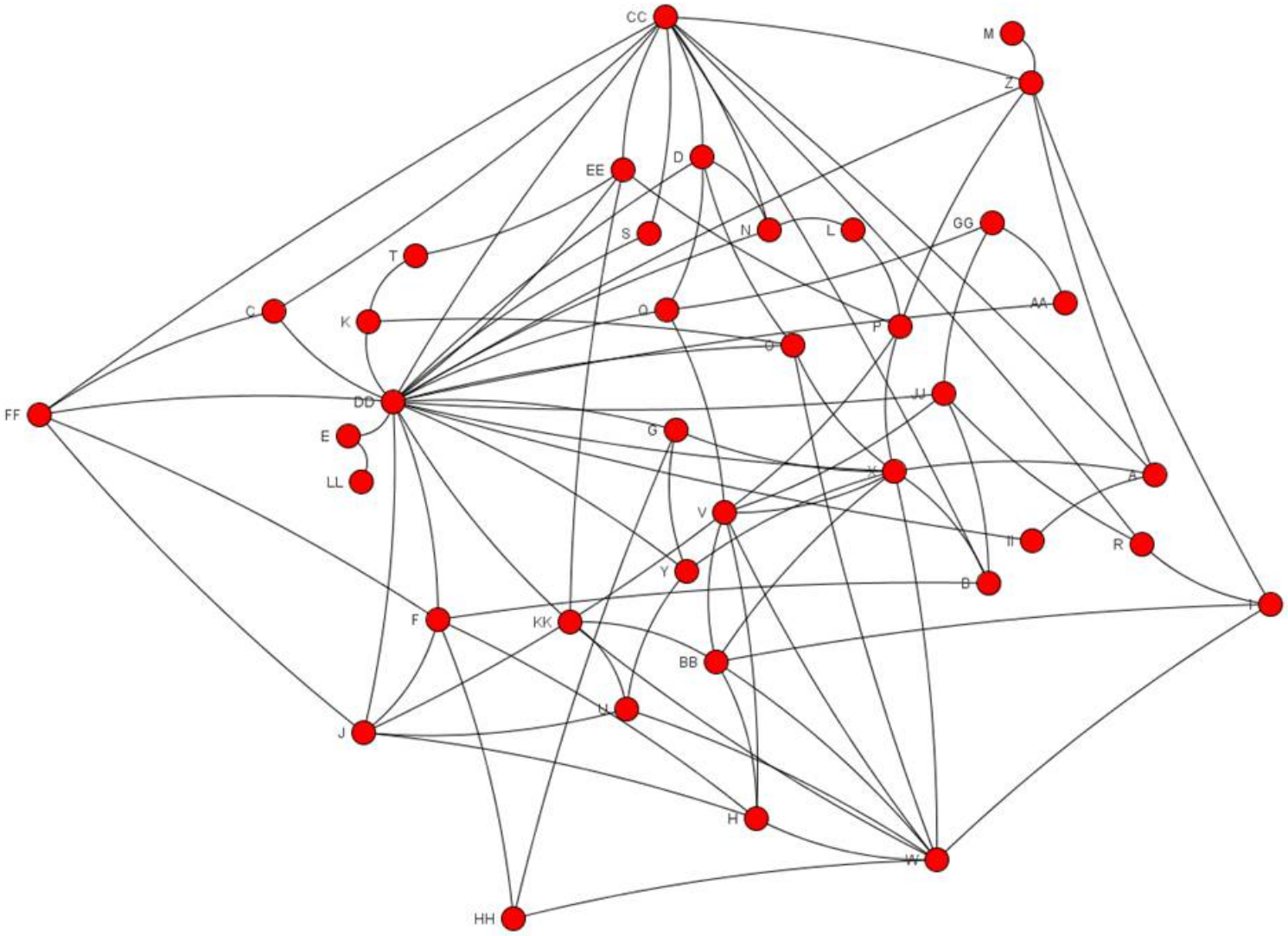

Figure VI-12 Visualization of All Sources Social Network Model

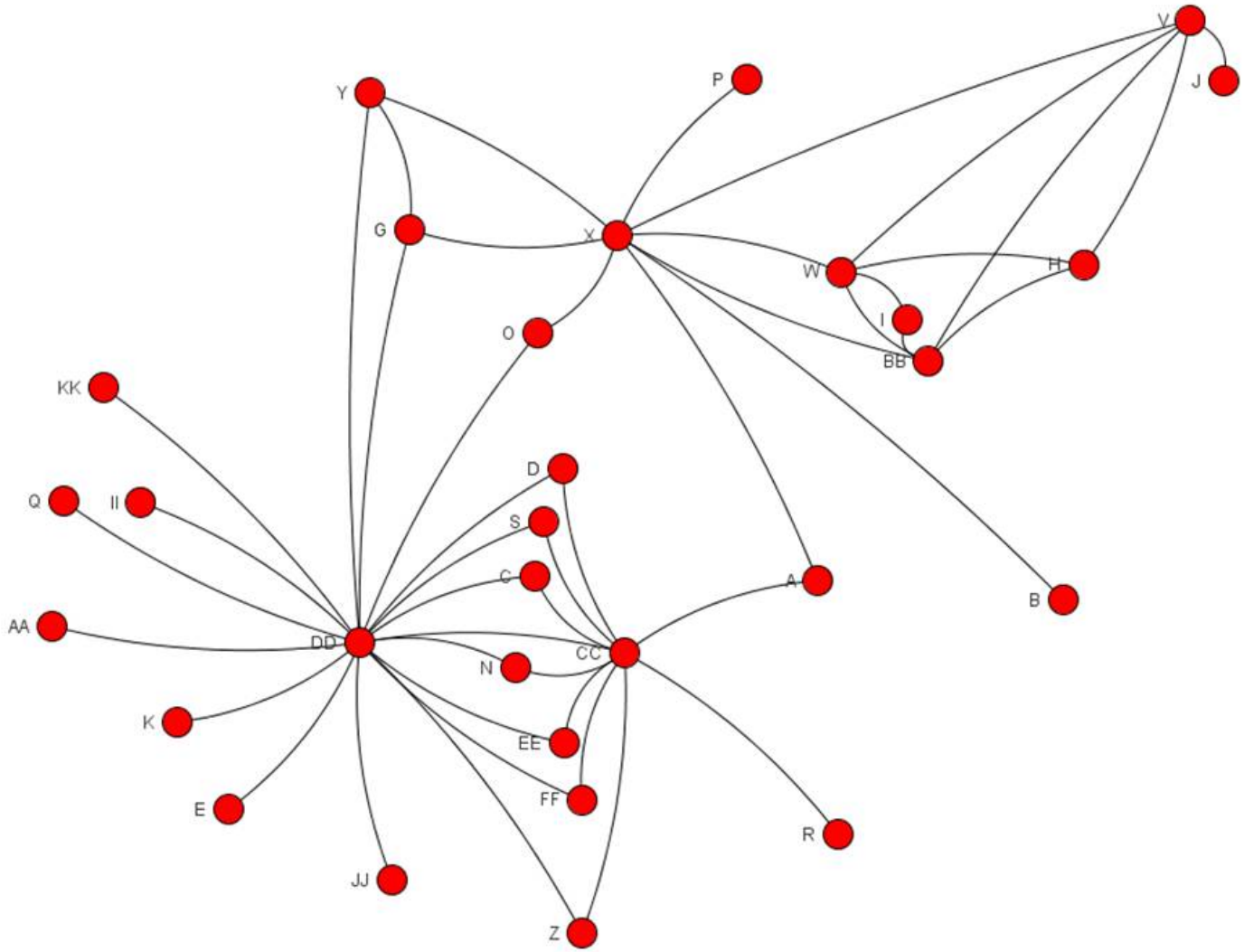

Figure VI-13 Visualization of Reliable Social Network Model

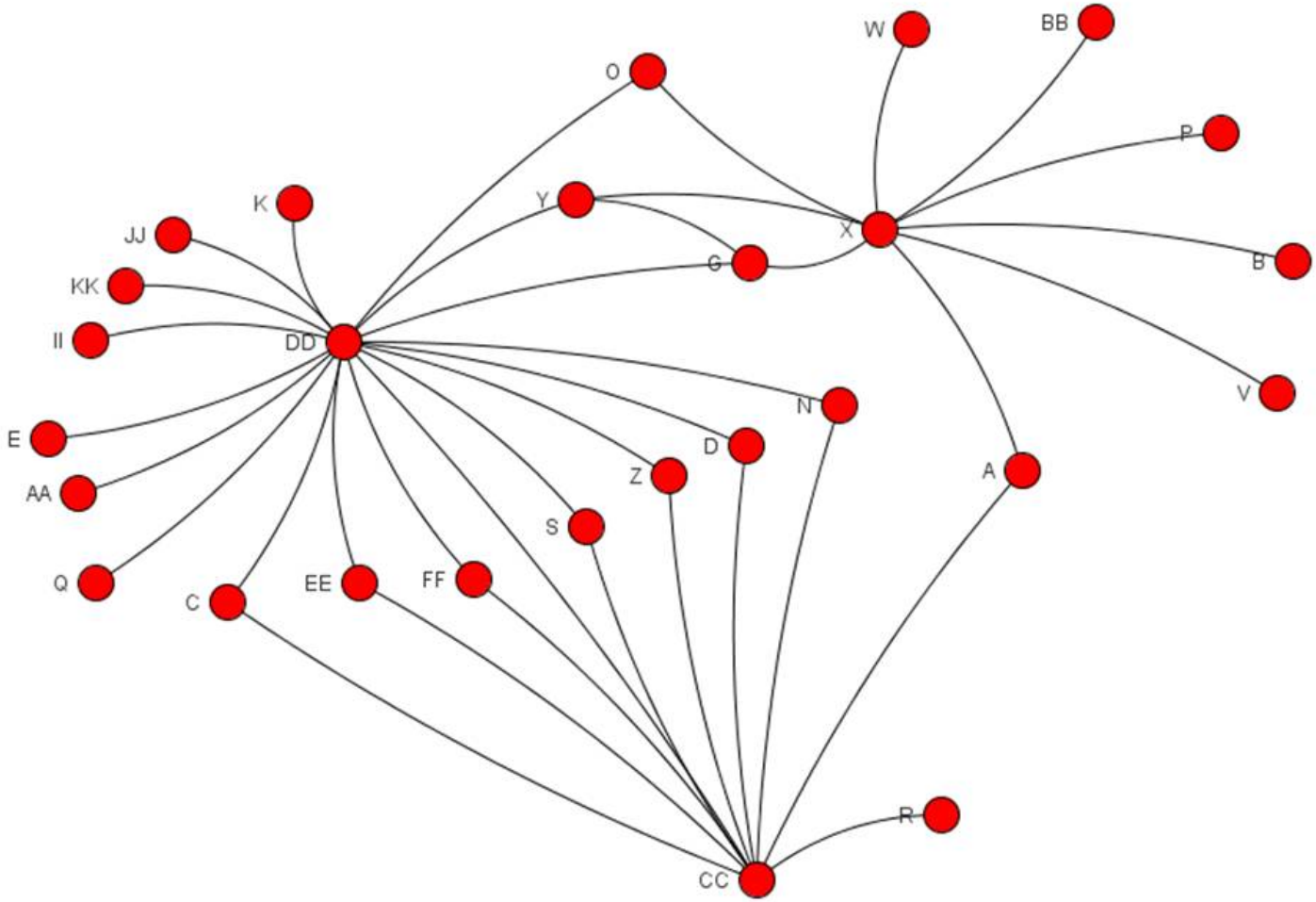

Figure VI-14 Visualization of Selected Social Network Model

6.4.1 SNA Network Measures.

Table VI-9 displays a comparison of SNA network measures' values for the four examined social network models. The SNA network measures are defined in Section 2.2.2. The number of nodes network varies between the four models, as can be expected. For this case study, it is worth noting, that extraneous actors not relevant to the underlying social network can not be included in any of the models. Conversely, it is possible for unreliable information sources to report extraneous relationships as evident by the All Sources model containing more edges that the Ground Truth model. The remaining SNA network measures investigated in Table VI-9 appear to produce similar results across the models with the exceptions of average clustering coefficient and degree correlation. All of the social network models show substantial departures from the Ground Truth model. These SNA network measures' results may be indicative of the various measures' sensitivities to imperfect data, with the average clustering coefficient affected greater than the other measures.

Table VI-9 Model Comparison with SNA Network Measures

	Social Network Models			
SNA Network Measures	**Ground Truth**	**All Sources**	**All Reliable**	**Selected Sources**
Number of Nodes	38	38	30	27
Number of Edges	85	92	46	37
Density	0.121	0.131	0.106	0.105
Diameter	4	4	5	4
Mean Path Length	0.433	0.456	0.395	0.440
Characteristic Path Length	0.421	0.443	0.367	0.426
Avg. Clustering Coefficient	0.457	0.258	0.650	0.612
Degree Correlation	-0.486	-0.219	-0.637	-0.729

6.4.2 SNA Nodal Measures.

SNA analysts generally use SNA nodal measures to rank the importance of actors participating in the social network. As discussed in Section 2.2.1, commonly used centrality measures include: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. This section examines how individual actors score across the social network models. The All Reliable and Selected Sources network models do not possess all of the actors and those nodes are identified as missing.

SNA analysts are generally more concerned about an actor's ranking in comparison to the other actors for a given centrality measure and not the actors' raw scores. The various rankings of the actors across the different measures can be examined statistically through the use of rank correlation measures, specifically Spearman's ρ and Kendall's τ. Spearman's ρ is equivalent to the Pearson correlation on the ranks without the presence of ties. If ties are present, the tied values are assigned the average of the ranks (Conover, 1971, pp. 245-246). Kendall's τ is based upon comparisons of pairs of observations drawn from both samples (Conover, 1971, p. 249). Spearman's ρ generally produces a larger value than Kendall's τ (Conover, 1971, p. 251).

6.4.2.1 Degree Centrality.

Examining the actors' rankings according to degree centrality as presented in Table VI-10, it appears that degree centrality is rather robust to deviations from the underlying social network. The actors possessing the largest degree centrality scores, reflected in a ranking of 1, 2, 3, …, and so forth, are relatively consistently identified by all four models. However, there are noted exceptions. The All Sources model identified

actor F of possessing the 6th highest degree centrality score. In Ground Truth model actor F is tied for the 25th ranking. This discrepancy results from the unreliable sources associating multiple false relationships to actor F. Actor F's degree centrality rankings demonstrate of how false reporting can increase, whether intentionally or mistakenly, the relative importance of an actor.

The developed methodology can not detect if inappropriate information is included in the social network model, but attempts to prevent its inclusion be assisting the SNA analyst in assessing the information sources providing the data. The methodology developed here only aids assessment information sources, with the intention of accepting or rejecting all data provided by reliable or unreliable sources respectively. Assessing data elements individually in the social network model would enable construction of the ideal model. Without utilizing *a priori* information, there is no currently available methodology that assesses individual data elements of a social network model and determines the impact upon overall conclusions regarding actor importance and other SNA results.

Examining the rank correlations of degree centrality across the four social network models, presented in Table VI-11, shows some interesting patterns. Of interest is that the All Sources model, representing the traditional SNA analyst practice, possesses the lowest rank correlation with the Ground Truth model, of 0.34 and 0.27 for Spearman's ρ and Kendall's τ, respectively. Additionally, the Selected Sources model, generated by employing the methodology, possesses a moderate to high rank correlation,

Table VI-10 Degree Centrality Actor Rankings by Social Network Model

	Social Network Model			
Actor	**Ground Truth**	**All Sources**	**Reliable Sources**	**Selected Sources**
DD	1	1	1	1
X	2	3	3	3
CC	3	2	2	2
M	4	37	Missing	Missing
Y	4	18	7	4
BB	6	6	4	15
FF	6	10	10	6
J	8	6	20	Missing
AA	8	31	20	15
II	8	31	20	15
K	11	26	20	15
N	11	18	10	6
V	11	5	4	15
W	11	4	4	15
JJ	15	10	20	15
A	16	18	10	6
D	16	10	10	6
G	16	18	7	4
H	16	10	7	Missing
I	16	18	10	Missing
O	16	10	10	6
S	16	31	10	6
Z	16	6	10	6
HH	16	26	Missing	Missing
B	25	18	20	15
C	25	26	10	6
F	25	6	Missing	Missing
L	25	31	Missing	Missing
T	25	31	Missing	Missing
U	25	18	Missing	Missing
EE	25	10	10	6
GG	25	26	Missing	Missing
LL	25	37	Missing	Missing
E	34	31	20	15
P	34	10	20	15
Q	34	18	20	15
R	34	26	20	15
KK	34	10	20	15

0.69 and 0.68, to the Reliable Sources model, the best model possible with this collection of sources for this example.

Comparing current SNA practice against performance obtained from employing the methodology, the Selected Sources model possesses greater rank correlations with Ground Truth than the All Sources model in this instance. The Selected Sources model possesses rank correlations of 0.43 and 0.37, for Spearman's ρ and Kendall's τ respectively, to the Ground Truth, which is greater than the All Sources model's rank correlations of 0.34 for Spearman's ρ and 0.27 and Kendall's τ. For this example, employing the methodology generates a more accurate representation of actors' importance in terms of degree centrality in comparison with the common SNA practice of including data from all available information sources.

Table VI-11 Degree Centrality Rank Correlations

	Ground Truth	**All Sources**	**Reliable Sources**	**Selected Sources**
Ground Truth		0.34	0.55	0.43
All Sources	0.27		0.61	0.33
Reliable Sources	0.48	0.53		0.69
Selected Sources	0.37	0.28	0.68	

upper triangle: Spearman's ρ
lower triangle: Kendall's τ

6.4.2.2 Closeness Centrality.

In contrast to degree centrality's rank results, closeness centrality displayed greater variability in the actors' ranking across the four social network models as illustrated in Table VI-12. Actor DD was identified as the most important actor, by

closeness centrality scores, by all four social network models. The Reliable Sources and Selected Sources models both exhibited numerous ties beginning at the 8th place ranking. Otherwise, examining the top ten individuals in the Ground Truth model, one can see how the Selected Sources model identified the majority of those actors, though in a different order. The All Sources model identified actors DD, X, CC, and O as correctly being in the top ten individuals. However, the All Sources model failed to identify the remaining top ten members.

Of interest is Actor A. Actor A was identified as rank 5, 6 and 7 by the Ground Truth, and the Selected Sources and Reliable Sources models, respectively. In contrast, the All Sources assessed Actor A of being tied for 25th in closeness centrality scores. This is of interest due to the Reliable Sources and the Selected Sources models are subsets of the reporting contained in the All Sources model. Thus, the extraneous false information provided by the unreliable sources resulted in Actor A's closeness centrality score to drop substantially. This is a reflection of the impact of false information in enabling important nodes to appear relatively unimportant through the inclusion of extraneous edges in the social network data.

Examining the closeness centrality rank correlations in Table VI-13, the All Sources model again exhibits the worst correlation with the Ground Truth. The Reliable Sources and Selected Sources models possess moderate rank correlation with the Ground Truth. Of note, the Selected Sources model possesses an extremely high rank correlation with the Reliable Sources model, which represents the best possible case with this particular data set.

Table VI-12 Closeness Centrality Actor Rankings by Social Network Model

	Social Network Model			
Actor	**Ground Truth**	**All Sources**	**Reliable Sources**	**Selected Sources**
DD	1	1	1	1
X	2	2	6	7
Y	3	11	2	3
FF	4	11	8	8
A	5	25	7	6
N	5	18	8	8
O	5	4	5	5
G	8	11	2	3
CC	8	3	2	2
U	10	32	Missing	Missing
J	11	7	30	Missing
AA	11	25	15	15
BB	11	20	22	23
II	11	25	15	15
K	15	23	15	15
JJ	16	7	15	15
M	17	37	Missing	Missing
HH	17	33	Missing	Missing
T	19	36	Missing	Missing
D	20	11	8	8
V	21	11	22	23
F	22	10	Missing	Missing
Z	22	4	8	8
S	24	23	8	8
W	24	16	22	23
C	26	20	8	8
EE	26	6	8	8
B	28	19	26	23
I	28	29	29	Missing
E	30	25	15	15
Q	30	17	15	15
KK	30	7	15	15
L	33	34	Missing	Missing
GG	33	35	Missing	Missing
P	35	20	26	23
H	36	29	28	Missing
R	37	29	25	22
LL	37	38	Missing	Missing

For this example, the Selected Sources model again showed greater rank correlation than the All Sources model. The Selected Sources model possesses a Spearman's ρ correlation of 0.67 compared against the All Source model's correlation of 0.44. Similarly for Kendall's τ, the Selected Sources model's correlation is 0.67, which is greater than the All Sources model's correlation of 0.50. Similar to the results for degree centrality, the methodology produced Selected Sources model more accurately characterizes actors for closeness centrality in comparison to the All Sources model.

Table VI-13 Closeness Centrality Rank Correlations

	Ground Truth	**All Sources**	**Reliable Sources**	**Selected Sources**
Ground Truth		0.44	0.68	0.67
All Sources	0.33		0.54	0.50
Reliable Sources	0.51	0.42		0.99
Selected Sources	0.50	0.39	0.96	

upper triangle: Spearman's ρ
lower triangle: Kendall's τ

6.4.2.3 Betweenness Centrality.

The actor rankings based on betweenness centrality of Table VI-14 showed a substantial departure in rankings agreement between the All Sources model and the Ground Truth. Actors DD, X, and CC were correctly identified as important by the All Sources model. However, excluding those three identifications, the remaining rankings substantially differ from the Ground Truth. In this instance, the false data in the All Sources model is obscuring the true importance, with respect to betweenness centrality, of a majority of the actors in the network. The Reliable Sources model and the Selected

Sources models both possess numerous ties at ranks 11 and 8, respectively. Both of these models correctly identified the top four actors in the network, with respect to betweenness centrality, but due to the numerous ties, drawing conclusions beyond the top four for the Selected Sources model and the top 10 for the Reliable Sources is difficult.

The betweenness centrality rank correlations in Table VI-15 again show the All Sources model exhibiting the lowest rank correlation with the Ground Truth. The Selected Sources model shows high correlation with the Reliable Sources model's results though this can be impacted by the numerous ties in the betweenness centrality scores.

For this example, the difference between the rank correlations of the Selected Sources model and the All Sources model with the Ground Truth were the largest observed on the SNA nodal measures examined here. The Selected Sources model's rank correlations of 0.60 and 0.52, for Spearman's ρ and Kendall's τ respectively, were substantially larger than the All Sources model's rank correlations of 0.21 and 0.15. In this case, for betweenness centrality, there are large discrepancies between relative actor importance between the methodology produced Selected Sources model and the common practice All Sources model, with the Selected Sources model exhibiting greater accuracy in terms of alignment with the Ground Truth.

Table VI-14 Betweenness Centrality Actor Rankings by Social Network Model

	Social Network Model			
Actor	**Ground Truth**	**All Sources**	**Reliable Sources**	**Selected Sources**
DD	1	1	1	1
X	2	4	2	2
CC	3	2	3	3
Y	4	21	4	4
FF	5	27	11	8
BB	6	20	9	8
AA	7	22	11	8
N	8	14	11	8
M	9	35	Missing	Missing
K	10	18	11	8
J	11	12	11	Missing
A	12	23	8	7
II	13	32	11	8
O	14	17	4	4
Z	15	3	11	8
G	16	19	4	4
U	16	29	Missing	Missing
V	18	7	7	8
HH	19	31	Missing	Missing
D	20	26	11	8
W	21	6	9	8
T	22	34	Missing	Missing
JJ	23	9	11	8
F	24	11	Missing	Missing
S	25	35	11	8
LL	26	35	Missing	Missing
B	27	25	11	8
C	27	35	11	8
E	27	5	11	8
H	27	28	11	Missing
I	27	16	11	Missing
L	27	33	Missing	Missing
P	27	10	11	8
Q	27	15	11	8
R	27	24	11	8
EE	27	8	11	8
GG	27	30	Missing	Missing
KK	27	13	11	8

Table VI-15 Betweenness Centrality Rank Correlations

	Ground Truth	All Sources	Reliable Sources	Selected Sources
Ground Truth		0.21	0.60	0.60
All Sources	0.15		0.38	0.32
Reliable Sources	0.52	0.31		0.88
Selected Sources	0.52	0.26	0.85	

upper triangle: Spearman's ρ
lower triangle: Kendall's τ

6.4.2.4 Eigenvector Centrality.

Eigenvector centrality exhibited the greatest variability in rankings between the four social network models as illustrated in Table VI-16. All of the models agreed upon Actor DD possessing the largest eigenvector score. The Reliable Sources and Selected Sources models both possessed numerous ties in actor scores.

Table VI-17 displays the rank correlations for eigenvector centrality. Eigenvector centrality was the only SNA nodal measure examined here where the All Sources model possessed a higher rank correlation to the Ground Truth than the other social network models for this example. The Selected Sources model possesses an extremely high correlation with the Reliable Sources model. The All Sources model only differs from the Reliable Sources model in that it possesses false information. That this inclusion of false information improves the All Sources' eigenvector centrality rank correlation with Ground Truth calls into question the effectiveness of eigenvector centrality in the presence of imperfect social network data.

For this data set, eigenvector centrality was the only occurrence where the All Sources model exhibited better rank correlation with the Ground Truth than the Selected

Table VI-16 Eigenvector Centrality Actor Rankings by Social Network Model

	Social Network Model			
Actor	**Ground Truth**	**All Sources**	**Reliable Sources**	**Selected Sources**
DD	1	1	1	1
X	2	2	12	12
Y	3	15	10	10
FF	4	7	3	3
CC	5	3	2	2
II	6	29	14	14
J	7	6	30	Missing
K	8	27	14	14
N	9	17	3	3
BB	10	14	23	23
M	11	37	Missing	Missing
AA	12	30	14	14
JJ	13	19	14	14
G	14	16	10	10
A	15	24	21	21
O	16	8	13	13
V	17	5	25	23
D	18	9	3	3
S	19	25	3	3
W	20	4	23	23
U	21	26	Missing	Missing
Z	22	11	3	3
C	23	18	3	3
EE	23	13	3	3
HH	25	31	Missing	Missing
I	26	28	29	Missing
B	27	22	27	23
F	28	10	Missing	Missing
T	29	36	Missing	Missing
E	30	33	14	14
Q	30	21	14	14
KK	30	12	14	14
H	33	20	26	Missing
L	34	35	Missing	Missing
GG	34	34	Missing	Missing
P	36	23	27	23
LL	37	38	Missing	Missing
R	38	32	22	22

Table VI-17 Eigenvector Centrality Rank Correlations

	Ground Truth	All Sources	Reliable Sources	Selected Sources
Ground Truth		0.52	0.41	0.39
All Sources	0.38		0.34	0.37
Reliable Sources	0.29	0.27		1.00
Selected Sources	0.29	0.28	0.99	

upper triangle: Spearman's ρ
lower triangle: Kendall's τ

Sources model. The All Sources model exhibited a Spearman's ρ correlation of 0.52 with the Ground Truth, compared against the Selected Sources model's rank correlation of 0.39. For Kendall's τ, The All Sources model's correlation of 0.38 was greater than the Selected Sources model's rank correlation of 0.29. For this case, the All Sources model possessed the greatest rank correlation with the Ground Truth of all examined models.

6.5 Conclusions

Some of these patterns in the rank correlations may result from inherent characteristics of the SNA nodal measures (Guzman, 2012). Degree centrality appears rather robust to imperfect social network data in comparison to the other SNA nodal measures for this data set. In contrast, eigenvector centrality appears to be the most susceptible to imperfect social network information for this case study. Regardless, the impact of imperfect social network information is displayed in the variability of the different models' actor rankings in comparison to the Ground Truth representing the true underlying social network.

For the four SNA nodal measures examined here, the Selected Sources model, derived from employing the methodology, exhibited greater rank correlation with the

Ground Truth than the All Sources model, representing a traditional practice of SNA analysts, for three of the measures, with the exception of eigenvector centrality. This calls into question the effectiveness of basing decisions upon SNA results stemming from models constructed from data provided by unvetted information sources. Unvetted information sources can be an avenue for imperfect social network information to be included in the model. As discussed in Section 2.4, the inclusion of imperfect social network information impacts SNA nodal measures' scoring of actors within the network (Costenbader & Valente, 2003; Sterling, 2004; Borgatti, Carley, & Krackhardt, 2006; Kossinets, 2006; Kim & Jeong, 2007).

6.6 Chapter Summary

This chapter presented an employment of the developed methodology in a case study format based on real world social network data. Unfortunately, as the raw information used to create the social network model, particularly the discarded data not included in the model, are unavailable, social network information sources had to be artificially generated. The reliable sources were generated in a manner consistent with OPSEC considerations of dark network organizations. The unreliable sources were generated using a random network generation technique, similar in nature to creating an Erdös-Rényi random graph. Using the generated social network information sources, the methodology was employed to test its ability to discern reliable and unreliable sources. The impact of the various possible social network models that could be constructed from this data was investigated. Overall, the social network model developed from application of the methodology displayed greater correlation to the Ground Truth model in this

example when examining the actor rankings by various, commonly applied SNA nodal measures as compared against the traditional SNA analyst practice of including information from all social network information sources. The next chapter presents the conclusions and recommendations for this research.

VII. Conclusions and Recommendations

This chapter summarizes the contributions of the methodology developed in this research. First, the methodology's assumptions and limitations are discussed. The theoretical contributions are then reviewed. Following the theoretical contributions, the practical contributions of this research to a SNA analyst are highlighted. Future recommended work that continues and extends the research conducted here is identified. Finally, other potential applications are explored.

7.1 Assumptions and Limitations

The methodology utilizes several assumptions commonly applied to informant accuracy in social network analysis as overviewed in Section 2.7.3. These assumptions precipitate several inherent limitations of the methodology. However, these limitations are present in every technique currently available to SNA analysts facing the information source assessment problem. This research has produced developments that address some of the limiting aspects of these assumptions.

7.1.1 Assumptions.

The four assumptions presented by Romney and Weller (1984) apply to the methodology and influence several limitations upon SNA analysts employing the methods developed in this research.

7.1.1.1 Single Underlying Social Network.

The methodology assumes that information sources are reporting data of a true underlying social network. This corresponds to the first assumption as presented by

Romney and Weller (1984): "There exists an objective set of "facts" or reality pertaining to the pattern of interaction of the group under investigation (Romney & Weller, 1984, p. 61)." This assumption can be violated if an analyst improperly defines the social network under investigation. Improper or imprecise boundary specification can cause irrelevant sources to be incorporated into the collection of information sources or relevant social network actors to be excluded from the reporting and subsequent analysis. These additional sources would be considered unreliable sources as the information they are reporting does not pertain to the social network model the SNA analyst is attempting to construct. The methodology may be able to correctly identify these sources as unreliable, but the inclusion of these inappropriate sources increases the non-zero probability of an incorrect classification. In the experimentation, one of the factors was the percentage of sources that are reliable. The inclusion of non-relevant sources in effect, decreases the percentage of reliable sources in the collection of sources.

Improper boundary specification can occur in another manner. Sources may be reporting on the correct underlying social network, but may be reporting on different relation types, as presented in Table II-1, that are outside the specified boundary. These sources could be reporting reliable information, but confounding it with information that is unreliable due to the boundary specification. For example, if a criminal organization is under investigation, and some sources may report relationships among actors that are based on common participation in criminal activities and other sources may report information based upon cooperation in criminal activities and familial relationships. If familial relations are not a basis for the criminal organization's structure, the sources reporting familial relations may be deemed unreliable as they could be presenting

relationships unconfirmed by other information sources. Including source provided irrelevant information may cause misclassification of sources during the source comparison phase of the methodology.

However, an advantage of employing the methodology developed in this research is the potential ability to evaluate the efficacy of different data collection means, techniques, and technologies. For example, a criminal organization is under investigation and social network data is being provided by informants and electronic communications monitoring. If the data being reported is properly screened to consider the boundary specification, i.e. communications only relevant to the workings of the criminal organization are included, the methodology developed here can assess the reliability of the informants and the communications monitoring. Utilizing the methodology, the information sources' reports are used to determine the likelihood of reliability for the informants and the electronic communications. Specific informants may be discovered to be providing information with the objective of misleading investigating authorities. It is also possible that the criminal organization's members are spoofing the electronic communications monitoring by conducting fake conversations with irrelevant actors. Assessing the reliability of the sources through the methodology may determine that for some dark network organizations certain data collection methods, such as electronic monitoring is inappropriate due to reliability issues when constructing the social network model.

7.1.1.2 Sources' Network Perspective.

It is assumed that information sources only possess partial knowledge of the underlying social network. It is assumed no one source has a complete picture of the social network. Additionally, not only do information sources only have partial information, the data they report varies in amount and in the portions of the social network they report upon. "Individuals vary in the extent to which they know all the facts or reality pertaining to the pattern of interaction of the group. We refer to this as knowledge (Romney & Weller, 1984, p. 61)." The methodology developed in this research, specifically accounted for this assumption. The pairwise source comparisons procedure only utilizes information that two sources have in common when determining their concordance. The score weighting component of the methodology accounts for the different social network perspective each source possesses. If this assumption is violated and sources are reporting on the same structural aspects of the social network, the methodology's performance may show substantial improvement as the sources' similarity scores will be based on greater amounts of data.

7.1.1.3 Independence of Sources.

The independence of the information sources' reporting is a critical assumption for the methodology: "The knowledge of each individual about the group is assumed to be independent of the knowledge of every other individual (Romney & Weller, 1984, p. 61)." Sources' assessed reliability is ultimately based on the confirmation and dissention of reported dyads. If the information sources are not independent, then the dyad confirmations and dissentions can be called into question. An example of this which

would adversely impact the methodology's performance is when a single source is delivering reports that are being attributed to multiple sources. For example, if an informant is reporting through multiple channels into an intelligence collection system, it is conceivable that the reports can be construed as initiating from different information sources. As these multiple sources would be compared against the other sources and each other in the pairwise source comparison component of the methodology, the informant's reports would be compared against themselves to assess concordance. The informant's reports would be scored as very concordant represented by a high similarity score and will also be highly weighted since they are reporting on the same aspects of the underlying social network structure. Thus, the informant's reports will be confirmed…by the informant's own reports! This confirmation coupled with similar network perspective increases the likelihood of the informant to be assessed as a reliable source. Though, it is likely that any information source assessment methodology would be susceptible and the resulting classification performance would be quite sensitive to this kind of error.

7.1.1.4 Source Reporting Overlap.

> The correlation of knowledge between any two subjects is a function of the extent to which each has knowledge of the objective reality. Specifically, the correlation of knowledge between individual A and individual B is the product of the correlation of individual A with the "truth" and of individual B with the "truth". (Romney & Weller, 1984, p. 61)

Social network information source assessment is based upon confirming or rejecting supplied information. The confirmation or dissention can be based upon an existing social network model, or a probabilistic model using Bayesian inference as in Section 2.6. The methodology developed in this research utilizes other sources' reporting

to confirm or question an individual source's reports. The source comparison is accomplished on a dyad by dyad basis. Inherent in this approach is that a source's information is able to be compared against other sources' reports. Without overlap of sources' reporting, the methodology presented here, consensus structure aggregation and Butt's Bayesian approach are ineffective.

7.1.2 Limitations of Conducting SNA on Dark Networks.

Conducting SNA on dark network organizations is accompanied by a unique set of issues that are not prevalent in traditional applications of SNA. Dark network actors may be conducting several techniques to frustrate data collection efforts. These dark network actor applied OPSEC techniques exacerbate the imperfect data problem normally associated with SNA.

7.1.2.1 Deceptive Information Sources.

One problem that is probably quite rare in traditional SNA applications is deceptive information sources. Deceptive sources' objective is the intentionally provide information to the data collection agent that is incorrect. To maximize the long term effects, deceptive sources begin by providing correct information in order to gain the data collector's trust. After the data collector's trust has been obtained, the deceptive source then introduces false information to deceive the analysis. The deceptive source may initially introduce false information with true information to obscure the false information. Over time, the ratio between false and true information can gradually shift to where the majority of the information being provided is incorrect.

An alternative model of a deceptive source is to provide true information until a significant time or event. At that point, false information is introduced with the goal of initiating a major disruption to the analysis. This strategy, a one big score, has the duplicitous informant waiting until an opportune time before providing false information to the data collector.

Both of these deceptive source strategies generate two effects. First, they confound the analysis with false data with the goal of inhibiting effective decision-making. Second, they attempt to corrupt the data collection system by discrediting reliable sources. As deceptive sources initially appear to be reliable, over time, they become a basis by which other sources are vetted. By initially assessing these deceptive sources as reliable, information sources acquired later maybe rejected as unreliable due to the false information being provided by the deceptive sources.

7.1.2.2 Practical Worst Case.

The worst case in practice for a SNA data collection effort on dark networks is one that faces multiple colluding deceptive sources. With collusion, deceptive sources are confirming each others reports. This confirmation leads SNA analysts to believe that the deceptive sources are reliable and they may even discount reliable sources due to dissentions with the deceptive sources. The multiple colluding deceptive sources can, in some instances, co-opt the data collection system, in that, all of the information it is acquiring originates from the colluding deceptive sources and it actively ignores reporting from other sources.

7.1.2.3 Mitigating Factor.

There exists a mitigating factor—if proper OPSEC has been conducted, information sources are generally not aware of what social network information the SNA analyst possesses. The reporting sources are providing information and do not have access to, or even be aware of, the other reporting sources. The reporting sources are providing information in a vacuum. They are unaware of the data collector's assessment of their reliability, and thus, it is difficult for the information sources to tailor their reporting to attempt to influence SNA analysts' conclusions.

7.2 Theoretical Contributions

The research conducted in this dissertation lead to several theoretical contributions. The developed methodology presented here was designed to address gaps in social network analysis methodology for constructing social network models in the presence of imperfect information sources. Thus, the methodology addresses needs for SNA and these practical contributions are discussed in Section 7.3. During the course of conducting the research several theoretical advancements were achieved and are depicted in Figure VII-1. This section describes those theoretical contributions.

7.2.1 Binary Similarity Measures Selection Methodology.

Binary similarity and dissimilarity measures have been introduced and developed for use in various academic disciplines. This has led to a wide selection of available measures with 105 being identified in this research. For comparison, the most complete listings of binary similarity measures discovered in the literature stems from a 2008 PhD dissertation cataloguing 76 measures (Choi, 2008) and a 2011 book containing 75

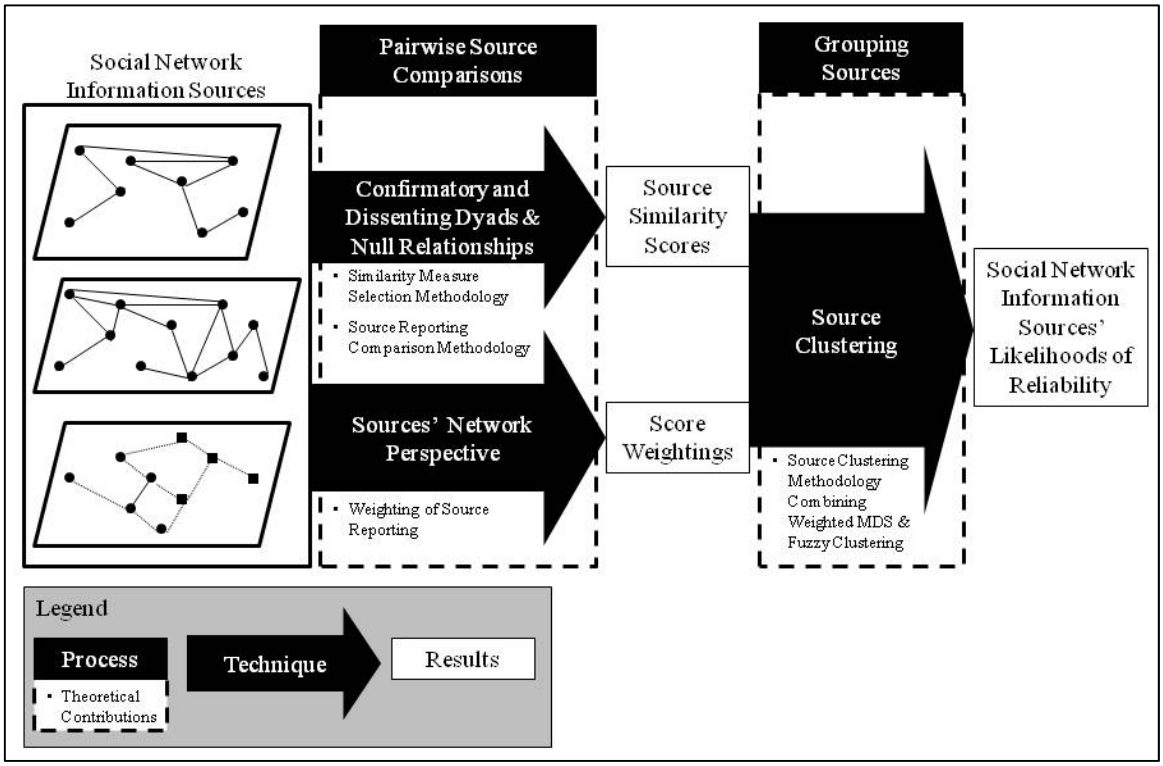

Figure VII-1 Methodology with Theoretical Contributions

measures (Eidenberger, 2011). This study assembled and identified 105 unique binary similarity and dissimilarity measures. It also determined measures that are algebraically equivalent to other measures, and identified measures referenced by multiple names in the literature spanning several academic disciplines. In the course of this research, a methodology was developed to select appropriate binary measures dependent upon the intended application. Prior to this development, measures were selected according to predominantly subjective assessments of the measures characteristics. The methodology was demonstrated in Chapter IV, reducing the 105 measures to a subset of seven measures suitable for social network information source assessment. This methodological framework can be used to select appropriate binary similarity measures

in applications other than the social network information source assessment function discussed in this dissertation.

7.2.2 Source Reporting Comparison Methodology.

This research was conducted in pursuit of comparing and assessing information sources reporting social network information. This research led to the development of a new methodology based upon pairwise comparisons that examined independent portrayals of the overall network and evaluated the concordance among these reports. Network models are common representations of many real world systems, yet little research has been done on proper data collection to construct these models. It is assumed that many applications are similar to social network analysis in that they draw data from multiple information sources. Information source assessment to determine information reliability is crucial to proper network model construction.

The methodology assesses information sources' reliability based solely upon the network information they report. The lack of required *a priori* information or estimates makes this methodology's approach novel. It conducts comparisons to other information sources' reports to identify concordance and dissentions among the sources. This information is presented in a visual depiction to the analyst to aid their decision of which information sources to ultimately include in the network model representation. Additionally, quantification of the sources' concordance is accomplished and sources' information can be included in the network model if an analyst specified threshold is achieved.

7.2.3 Weighting of Source Reporting.

Information sources reporting network information may only present data reflecting a portion of the overall network. This can result from the information source's perspective of the network, in which the source only has access to specific subcomponents of the overall graph. This research suggests a source weighting mechanism to deal with information sources' perspective. Its inclusion in the developed methodology ensures that sources reporting on unique portions of the network are not penalized due to non-concordance with other sources providing information. It additionally enables network substructures that benefit from several sources' confirmations to serve as a litmus test for evaluating sources reporting on the same structures.

7.2.4 Information Sources Clustering Methodology.

The developed methodology presented in this research enabled clustering of information sources in accordance with their levels of pairwise concordance. This was accomplished by the novel approach of utilizing fuzzy clustering on a weighted multidimensional scaling (MDS) visualization. Weighted MDS enhances the SNA analysts' immersion in the data, by allowing them to visualize the information sources' concordance. The methodology's capability to visualize information sources' concordance is a new development in the fields of social network analysis and network science. Applying fuzzy clustering allows various thresholds for information source inclusion into the final social network model to be examined. Before the development of the methodology in this research, information source inclusion and exclusion into a

network model was based on arbitrary, sometimes subjective, rules specified by the analyst. The methodology developed in this research is the first known non-Bayesian quantitative technique for network models.

7.2.5 DOE with Quantile Regression.

The analysis of the experimentation utilized design of experiments (DOE) with quantile regression. Applying quantile regression to DOE has not been documented in the literature. This is not surprising as quantile regression is a relatively new introduction as a statistical technique. However, the relaxation of error distributional assumptions normally accepted by applying ordinary least squares regression shows a great promise of utilizing quantile regression in DOE where applicable. For the network sciences discipline, this research demonstrated the inappropriateness of the normality assumption of the error terms in a linear regression model. While, it is not known how widespread this normality assumption fallacy extends in the network sciences, this research is the first acknowledgement that it may be present. The experimentation detected the normality assumption violations, and proposed and implemented a solution, quantile regression, applying it in the analysis phase of the research. This research is the first known documented application of quantile regression to network models as well as the introduction of quantile regression accompanying DOE.

7.2.6 Examining Classifier Performance.

The methodology developed in this research can essentially be distilled to a classification problem of determining the reliability of information sources. The traditional area under the response operating curve was the performance measure under

observation. The average and in this case, median performance, was statistically analyzed in Chapter V. However, unique to this analysis was examining the extreme performance of the classifier. Utilizing quantile regression, the factors contributing to poor performance of the classifier were examined. This focus on what constitutes and drives poor classification performance is a unique and new application of quantile regression. This novel application extends quantile regression while contributing a new statistical technique to examining classifiers' performance.

7.2.7 Random Social Network Generation.

In the course of this research, a random network generator was created to produce graphs with desired characteristics found in real world social networks. The Prescribed Node Degree, Connected Graph (PNDCG Algorithm) builds upon previous algorithms used to generate random SNs for testing metrics and algorithms in SNA. The PNDCG Algorithm has several advantages over the other algorithms in use today. One of these is the reduction in assortative mixing prevalent in graphs generated by the other algorithms. Another advantage is the user's ability to specify *a priori* the degree distribution for the nodes (a characteristic some other algorithms also possess).

Perhaps the greatest advantage of the PNDCG Algorithm over the other available algorithms is its ability to generate a weakly connected graph. As most analytical techniques used in SNA require the use of fully connected graphs, this provides an efficient way to generate random networks for testing. The algorithm can be executed several times to generate components that can be combined to form a disconnected

network. Because the user generates each component separately, s/he can prescribe the degree distribution of each of the components.

Two main extensions provide important additional functionality to the PNDCG Algorithm. The first allows the creation of clusters, relational "triangles" in the generated random graph. This is a property common among SNs and among clandestine organizations in particular. The second extension allows for the inclusion of degree adjacency information through the use of an exemplar network. This adaptation was driven by the lack of assortative mixing exhibited in cellular organizations. The tendency for other algorithms to connect hubs made it unlikely that the algorithms would generate networks with cellular structures. Through a parameter and an exemplar network, the PNDCG Algorithm is able to generate random networks with characteristics similar to the exemplar network. Thus, the generated networks can be built to exhibit desired levels of assortative mixing. The PNDCG Algorithm enables the generation of random graphs that better imitiate characteristics found in real world social networks and makes the subsequent experimentation results more relevant to SNA analysts.

7.3 Practical Contributions

There are several practical contributions for SNA analysts implementing the methodology presented in this research. The procedures defined herein begin to address an analytic gap in SNA methodology. Proper construction of social network models is essential before SNA techniques and methodologies can be applied to generate meaningful and valid analytical results and conclusions, particularly when applying SNA to the dark networks problem set.

7.3.1 SNA Analyst Aid.

First and foremost, the methodology developed in this research is an aid to the SNA analyst. As SNA investigates human behavior, there are numerous complexities that are not completely represented by any mathematical formulation of a social network model. Externalities, the analytical objective, and a host of other considerations must be addressed when conducting a social network analysis. Examination of the data collection techniques and the associated information sources, up to the development of this methodology, has been traditionally accomplished through subjective analyst expertise. The development of this methodology presents a quantitative procedure which can be applied to aid the SNA analyst in assessing the information sources. This quantitative methodology shows promise of substantial improvement of the traditional subjective methods, as it reduces the uncertainty of dealing with idiosyncrasies of expert opinion. In the end, however, subject matter expertise insight should not be ignored.

7.3.1.1 Quantitatively Compare Source Reporting.

This methodology develops techniques to quantitatively compare social network information sources' reporting. There are no limitations to the number or type of sources that can be compared. The sources' reporting can be assessed as long as some confirmations and dissentions with other sources are available. This methodology's lack of restrictions on sources' reporting enables a level of flexibility and widespread applicability to SNA analysts.

7.3.1.2 Sources' Network Perspective Characterization.

Unique to this methodology, sources' perspectives of the underlying social network are taken into account when assessing their suitability for incorporating their reporting into the social network model. Dark networks can be large or composed of numerous sub-structures. As a result of their clandestine nature, it is conceivable that information sources are limited in their view of the network's actors and their relationships. Sources reporting on different aspects of the network should not be expected to favorably compare against sources examining other network components. Conversely, sources reporting on the same aspects of the network are expected to provide similar reporting. Accounting for these phenomenons quantitatively captures heuristics subjectively employed by SNA analysts. Again, the quantitative methods developed in this research aid and augment the critical thinking processes of SNA analysts and are not considered a replacement for analysts' heuristics, methods, and intuition.

7.3.1.3 Operational Risk Considerations.

An advantage of the methodology presented here is that SNA analysts can specify a threshold for information sources' data inclusion into the social network model. Specifying a threshold allows the SNA analyst to account for the operational risk associated with the decisions that will result from the analysis. Varying the threshold will enable the construction of several social network models, allowing sensitivity analysis to be conducted on the applied SNA methodologies and techniques. Especially when dealing with dark networks, the decisions resulting from SNA can result in a substantial expenditure in resources, time, and possibly lives.

7.3.2 Each SNA is a Unique SNA.

Every SNA is a unique analysis and subject matter expertise may prove to be the deciding factor between deriving correct assessments or false conclusions. The analytical focus varies and each social network possesses unique characteristics. It is important to note that SNA analysts do not control the social network's characteristics, but merely observe and account for them in the subsequent analysis. The experimentation conducted on the methodology in this research examined a wide range of factors that affect the methodology's performance. The SNA analyst can not control these factors, and in most cases, will have difficulty even estimating the factors' values. The design of experiments used in this study examined viable ranges for the factors characterizing social networks encountered by SNA analysts when dealing with real world dark social networks.

7.3.3 Trusted Information Sources.

The methodology possesses an easy to implement adaptation to account for trusted information sources. The methodological procedures can be accomplished without initially identifying the trusted sources. At the completion of the methodology, information sources that are classified as concordant with trusted sources can be assessed as reliable and discordant sources can be discounted as unreliable. However, a unique aspect of this methodology is the ability to assess trusted sources. A source may be assumed to be trustworthy based on *a priori* information. Although, if other information sources are concordant with each other but discordant with the trusted source, it may be indicative that the assumption of trustworthiness needs to be reexamined.

7.4 Recommendations for Future Research

While conducting this research, several areas were identified that show potential promise in future developments. These stem from theoretical gaps in the literature and from capability needs in social network analysis. First, improvements to the developed methodology are discussed leading to general potential research contributions to the discipline.

7.4.1 Methodological Improvements.

There are several components of the developed methodology discussed in this section that may present potential performance improvements if additional developments occur. These identified areas address specific aspects of the developed methodology that appear to be fertile ground for theoretical research that could lead to practical application betterments.

7.4.1.1 Incorporation of a priori Information.

The methodology developed in this research enables assessment of information sources based solely on the information they provide. In real world applications, frequently data regarding the information sources is available. In some instances, this data can provide indications of a source's reliability. The methodology presented here could be extended by incorporation *a priori* information regarding the reporting sources to more accurately classify an information source.

7.4.1.2 Additional Statistical Clustering Techniques.

Fuzzy clustering was selected as the clustering technique for its capability of providing an indication of each member's inclusion likelihood for each cluster. A key

parameter of the technique is specifying the number of clusters. In this experimentation, two clusters were selected and exhibited good performance in the overall classification methodology. Numerous statistical clustering techniques exist and others may exhibit enhanced methodological performance for this application. Additional research in determining the number of clusters to include in the methodology may substantially improve the methodology's performance under certain conditions.

7.4.1.3 Investigating Other Weighting Mechanisms.

During the course of developing the methodology, pairwise source weightings were utilized to represent the sources' varying perspectives of the social network. The weightings account for sources reporting on different substructures within the social network. The weighting mechanism developed in this dissertation measures the overlap in reporting between sources, computed by the dividing the number of nodes in common by the total number of nodes between two information sources. Other weighting mechanisms could be constructed, perhaps incorporating other data such as considerations of the social network's underlying structure or *a priori* information regarding the information sources.

7.4.1.4 Investigating Higher Dimensions of Weighted MDS.

The weighted MDS was applied to provide a visualization to aid the social network analyst in indentifying reliable and unreliable information sources. The weighted MDS was conducted in a 2-dimensional space, although mathematically higher dimensions can easily be accommodated. Two dimensions were selected to facilitate visualization for analyst consumption. However, higher dimensions may provide greater

insight into source reporting concordance and should be explored if the greater complexity increases analyst understanding of the collected data.

7.4.2 General Discipline Gaps.

In the course of conducting this research, it became apparent that a major source of variation in the methodology's performance stemmed from variations in the underlying graphs. Network measures were applied to capture this graph to graph variation, but it is believed that further developmental opportunities exist in this area.

7.4.2.1 Characterizing Networks.

While conducting the statistical analysis of the experimentation, the importance of network characteristics as covariates became apparent. These variables that represented structural characteristics of the underlying social network increased the statistical models' explanatory power. The network measures selected to represent the graph's characteristics were chosen due to their prevalence within the network literature. However, most of these network measures are relatively recent additions to the field. With their relatively recent introduction in the literature, these network measures may not accurately fully characterize graph structures. The analysis also identified several strong correlations existing among the network measures. A result of these strong correlations, and the novelty of these network measures, highlights a potential field of future research. Identifying new network measures and characterizing the relationships among them may lead to a better understanding of graph theory, network science, and potentially uncovering attributes associated with real world data sets. The analysis conducted in this research identified the significance of several of these network measures to the

methodology's performance. It is conceivable that other SNA methodologies', measures', and techniques' performance and interpretation is a function of network characteristics and structural attributes. Further research in identifying suitable network measures and their impact upon SNA should be conducted as they could result in substantial theoretical and practical application improvements.

7.4.2.2 Network Population Estimation.

One defining characteristic of graphs is the number of nodes composing the network. When modeling social networks, boundary specification is an essential initial step which defines the social network as a finite set of actors, usually restricted by a common actor characteristic, specific relations, or associations of interest. When modeling bright networks, the size of the social network graph is usually known or quickly determined during the data collection phase. When constructing social network models of dark networks, the number of actors composing the network is likely unknown due to OPSEC efforts by the adversary. If the size of the network is known, it is substantially easier to determine if enough information has been collected to begin SNA techniques to assess the network. If the total number of actors is unknown, it is difficult to determine how representative the network model is of the true underlying social network.

Social network information sources generally only report on a portion of the underlying social network, and therefore, a single source is unlikely to present an accurate count of the total number of involved actors in any but modest sized networks. When considering the entire collection of information sources, all actors within the social

network may, or may not, be identified and accounted. It is very likely that with dark networks, only limited information sources will be available and will only provide at best, a partial presentation of the social network. Methodologies that estimate the total number of actors within a social network based upon various information sources' reports would provide bounds by which to judge if the amount of information is sufficient to conduct specific SNA techniques appropriately.

7.5 Other Potential Applications

The methodology developed in this research may be relevant to other applications beyond network model construction. This research focused on evaluating the reliability of information sources providing social network data. The general methodology could be extended to address information sources providing other types of data. For example, information sources can report event data which details an event occurrence, temporal data, and geospatial locations. Information sources could be assessed on whether there is concordance among the sources on the details of the events. If pairwise source concordance measurement can occur, the methodology developed in this research can be applied to visualize and cluster sources based on their concordance in reporting.

Another potential application is the increasing use of recommender systems for online marketing. These recommender systems use customer information to make product and service recommendations to market to the customer. These recommender systems are generally proprietary, and it is possible that several recommender systems are active to generate customer recommendations. Evaluating the efficacy of multiple recommender systems by examining their performance of generating a customer purchase

based upon a recommendation could lead to substantial improvement in business' personalized marketing efforts. The recommender systems could be evaluated on their recommendations' concordance with other recommender systems and with consumer purchases. This could lead to determining which recommender system is the most effective on a customer by customer basis.

7.6 Conclusions

This research culminated in developing a methodology to assess information sources providing data used to construct social network models. This addresses a real world problem that is faced by the DoD and other agency analysts on a frequent basis. As SNA's acceptance and usage continues to increase, it will become more integrated into decision-making processes. The decisions will be based on analysis resulting from applying SNA methods and techniques, but ultimately derives from a social network model. The construction of this social network model has to this point received sparse attention despite its importance as a preliminary step of the analytical process.

The methodology developed in this research provides a unique method to assess information sources. Utilizing this technique, SNA analysts will be able to construct more accurate social network models. Additionally, SNA analysts can provide feedback to the data collection system, as the methodology enables them to quantitatively assess information sources and potentially identifying reliable and unreliable sources. Closing this communication loop will improve the data collection process and enhance the resulting analysis—ideally leading to better decision making when dealing with dark networks.

Appendix A Binary Similarity and Dissimilarity Measures

Table A-1 Binary Similarity Measures

Similarity Measures (alternative names)	Equation	Range
Anderberg [1]	$\frac{8a}{8a+b+c}$	[0, 1]
Anderberg's D [1]	$\frac{\sigma-\sigma'}{2n}$ where $\sigma = \max(a,b) + \max(c,d) + \max(a,c) + \max(b,d)$ and $\sigma' = \max(a+c, b+d) + \max(a+b, c+d)$	[0, 1)
Baroni-Urbani & Buser-I [2]	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	[0, 1]
Baroni-Urbani & Buser-II [2]	$\frac{\sqrt{ad}+a-(b+c)}{\sqrt{ad}+a+b+c}$	[-1, 1]
Batagelj & Bren [3]	$\frac{bc}{ad}$	[0, ∞)
Benini (1901) [4]	$\frac{a-(a+b)(a+c)}{a+\min(b,c)-(a+b)(a+c)}$	[1, 2]
Braun-Blanquet[5]	$\frac{a}{\max(a+b, a+c)}$	[0, 1]
Browsing	$a-bc$	(-∞, ∞)
Clement [6]	$\frac{a(c+d)}{(a+b)}+\frac{d(a+b)}{(c+d)}$	(0, ∞)
Cohen's κ [7]	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	[-½, 1]
Cole-I [8]	$\frac{ad-bc}{\min((a+b)(a+c),(b+d)(c+d))}$	[-1, ∞)
Cole-II [8]	$\frac{ad-bc}{(a+b)(b+d)}$	[-1, 1]
Cole-III [8]	$\frac{ad-bc}{(a+c)(c+d)}$	[-1, ∞)
Cosine Ochiai-I [9] Otsuka [Look in Ochiai paper] Driver & Kroeber [10] Fowlkes & Mallows [11] (Gower & Legendre XII) [12]	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0, 1]
d Specific Agreement	$\frac{2d}{2a+b+c}$	[0, ∞)
Dennis [13]	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	[-1, ∞)
Dice-I [14] Wallace [15] Post & Snijders [16]	$\frac{a}{a+b}$	[0, 1]

Similarity Measures (alternative names)	Equation	Range
Dice-II [14] Wallace [15] Post & Snijders [16]	$\frac{a}{a+c}$	[0, 1]
Digby [17]	$\frac{(ad)^{\frac{3}{4}}-(bc)^{\frac{3}{4}}}{(ad)^{\frac{3}{4}}+(bc)^{\frac{3}{4}}}$	[-1, 1]
Dispersion	$\frac{ad-bc}{(a+b+c+d)^2}$	[-⅓, ⅓]
Doolittle [18] Pearson (1926)	$\frac{(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$	[0, 1]
Eyraud [19]	$\frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$	(-∞, ∞)
Fager & McGowan [20]	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2\sqrt{\max(a+b,a+c)}}$	[-½, 1)
Faith [21]	$\frac{a+0.5d}{a+b+c+d}$	[0, 1]
Fleiss [22]	$\frac{(ad-bc)[(a+b)(b+d)+(a+c)(c+d)]}{2(a+b)(a+c)(b+d)(c+d)}$	(-∞, 1]
Forbes-I [23]	$\frac{na}{(a+b)(a+c)}$	[0, ∞)
Fossum [24] Jones & Curtis [25]	$\frac{n(a-0.5)^2}{(a+b)(a+c)}$	(0, ∞)
Gilbert [26] (Ratio of Success)	$\frac{a-\frac{(a+b)(a+c)}{n}}{a+b+c-\frac{(a+b)(a+c)}{n}}$	[-⅓, 1]
Gilbert & Wells [27]	$\log a-\log n-\log(\frac{a+b}{n})-\log(\frac{a+c}{n})$	[0, ∞)
Gini [28]	$\frac{a-(a+b)(a+c)}{\sqrt{(1-(a+b)^2)(1-(a+c)^2)}}$	$[-4/3, 0]$
Gleason [29] Dice [14] Sørenson [30] (Coincidence Index) (Quotient Similarity) Czekanowski [31] Nei & Li [32] (Genetic Coefficient) (Gower & Legendre VII) [12]	$\frac{2a}{2a+b+c}$	[0, 1]
Goodman & Kruskal Max [33]	$\frac{a+d-\max(a,d)-\frac{b+c}{2}}{1-\max(a,d)-\frac{b+c}{2}}$	[-1, 1]
Goodman & Kruskal Min [33]	$\frac{2\min(a,d)-b-c}{2\min(a,d)+b+c}$	[-1, 1]
Goodman & Kruskal Probability [33]	$\frac{\max(a,c)+\max(b,d)-\max(a+b,c+d)}{1-\max(a+b,c+d)}$	[-⅓, 0]

Similarity Measures (alternative names)	Equation	Range
Goodman & Kruskal's Lambda [33]	$\frac{\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)-\max(a+c,b+d)-\max(a+b,c+d)}{2-\max(a+c,b+d)-\max(a+b,c+d)}$	[0, 1]
Goodman & Kruskal's Tau [33]	$\frac{\frac{(a-(a+b)(a+c))^2+(b-(a+b)(b+d))^2}{(a+b)}+\frac{(c-(a+c)(c+d))^2+(d-(b+d)(c+d))^2}{(c+d)}}{1-(a+c)^2-(b+d)^2}$	(-∞, -2]
Gower [34]	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	[0, 1.5]
Hamann [35] Holley & Guilford [36] Hubert [37] (Gower & Legendre IX) [12]	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[-1, 1]
Harris & Lahey [38]	$\frac{a((c+d)+(b+d))}{2(a+b+c)}+\frac{d((a+b)+(a+c))}{2(b+c+d)}$	[0, ∞)
Hawkins & Dotson [39]	$\frac{1}{2}(\frac{a}{a+b+c}+\frac{d}{b+c+d})$	[0, 1]
Inner Product (Hamming Complement) [40]	$a+d$	[0, ∞)
Intersection	a	[0, ∞)
Jaccard [41] Gilbert [26] (Ratio of Verification) Tanimoto [42] (Cosine Coefficient) (Gower & Legendre III) [12]	$\frac{a}{a+b+c}$	[0, 1]
Jaccard-3W	$\frac{3a}{3a+b+c}$	[0, 1]
Johnson (1967)[43]	$\frac{a}{a+b}+\frac{a}{a+c}$	[0, 2]
Kent & Foster-I [44]	$\frac{-bc}{b(a+b)+c(a+c)+bc}$	[-⅓, 0]
Kent & Foster-II [44]	$\frac{-bc}{b(c+d)+c(b+d)+bc}$	[-⅓, 0]
Köppen [45]	$\frac{(a+b)(1-a-b)-c}{(a+b)(1-a-b)}$	(-∞, ∞)
Köppen [46]	$a+\frac{b+c}{2}$	[0, ∞)
Kuder & Richardson [47] Cronbach [48]	$\frac{4(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)+2(ad-bc)}$	[-2, 1]
Kuhns [49]	$\frac{2(ad-bc)}{n(2a+b+c)}$	[-½, 1]
Kuhns Proportion [49]	$\frac{ad-bc}{n\left(1-\frac{a}{(a+b)(a+c)}\right)\left(2a+b+c-\frac{(a+b)(a+c)}{n}\right)}$	[-⅓, 1)
Kulczyński-I [50] (Gower & Legendre I) [12]	$\frac{a}{b+c}$	[0, ∞)
Kulczyński-II [50] Driver & Kroeber [10] (Gower & Legendre X) [12]	$\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$	[0, 1]

Similarity Measures (alternative names)	Equation	Range
Loevinger's H [51] [52] Forbes-II [23] Mokken [53] Sijtsma & Molenaar [54]	$\frac{na-(a+b)(a+c)}{n\min(a+b,a+c)-(a+b)(a+c)}$	[-1, 1]
Maron & Kuhns [55]	$\frac{ad-bc}{(a+b+c+d)}$	(-∞, ∞)
Maxwell & Pilliner [56]	$\frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	[-1, 1]
McConnaughey [57]	$\frac{a^2-bc}{(a+b)(a+c)}$	[-1, 1]
Michael [58]	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	[-1, 1]
Modified Gini [28]	$\frac{a-(a+b)(a+c)}{1-\frac{\lvert b-c\rvert}{2}-(a+b)(a+c)}$	$[0, {}^4/_3]$
Mountford [59]	$\frac{a}{0.5(ab+ac)+bc}$	[0, 2]
Pearson & Heron-II [60]	$\cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)$	[-1, 1]
Pearson-I [61] (Coefficient of Chi-square Contingency)	χ^2 where $\chi^2=\frac{n(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$	[0, ∞)
Pearson-II [61] (Coefficient of Mean Square Contingency)	$\sqrt{\frac{\chi^2}{n+\chi^2}}$ where $\chi^2=\frac{n(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$	$[0, \sqrt{{}^1/_2})$
Pearson-III [62] (Coefficient of Racial Likeness)	$\sqrt{\frac{\rho}{n+\rho}}$ where $\rho=\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	$[0, \sqrt{{}^1/_3})$
Peirce-I [63]	$\frac{ad-bc}{(a+b)(c+d)}$	[-1, 1]
Peirce-II [63]	$\frac{ad-bc}{(a+c)(b+d)}$	[-1, 1]
Peirce-III [63]	$\frac{ab+bc}{ab+2bc+cd}$	[0, 1]
Phi Coefficient Yule [64] Pearson & Heron-I [60] (Fourfold point correlation) (binary version of Pearson's Product Moment Correlation Coefficient) [61] (Gower & Legendre XIV) [12]	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	[-1, 1]
Relative Decrease of Error Probability	$\frac{\max(a,b)+\max(c,d)-\max(a+c,b+d)}{1-\max(a+c,b+d)}$	[-1, 0]
Rogers & Tanimoto [65] Farkas [66] (Gower & Legendre VI) [12]	$\frac{a+d}{a+2(b+c)+d}$	[0, 1]
Rogot & Goldberg [67]	$\frac{a}{(a+b)+(a+c)}+\frac{d}{(c+d)+(b+d)}$	[0, 1]

Similarity Measures (alternative names)	Equation	Range
Russell & Rao [68] (dot product) (inner product) (Gower & Legendre II) [12]	$\frac{a}{a+b+c+d}$	[0, 1]
Scott [69]	$\frac{4(ad-bc)-(b-c)^2}{(2a+b+c)(b+c+2d)}$	[-1, 1]
Simpson [70] (Ecological Coexistence Coefficient)	$\frac{a}{\min(a+b,a+c)}$	[0, 1]
Sokal & Michener [71] (Simple Matching Coefficient) Rand [72] Brennan & Light [73] (Gower & Legendre IV) [12]	$\frac{a+d}{a+b+c+d}$	[0, 1]
Sokal & Sneath-I [74] (Gower & Legendre V) [12]	$\frac{a}{a+2b+2c}$	[0, 1]
Sokal & Sneath-II [74] (Gower & Legendre VIII) [12]	$\frac{2(a+d)}{2a+b+c+2d}$	[0, 1]
Sokal & Sneath-III [74]	$\frac{a+d}{b+c}$	[0, ∞)
Sokal & Sneath-IV [74] (Gower & Legendre XI) [12]	$\frac{\frac{a}{(a+b)}+\frac{a}{(a+c)}+\frac{d}{(b+d)}+\frac{d}{(c+d)}}{4}$	[0, 1]
Sokal & Sneath-V [74] Ochiai-II [9] (Gower & Legendre XIII) [12]	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	[0, 1]
Sorgenfrei [75] Cheetham & Hazel [76] (Correlation Ratio)	$\frac{a^2}{(a+b)(a+c)}$	[0, 1]
Stiles [77]	$\log_{10}\frac{n(\lvert ad-bc\rvert-\frac{n}{2})^2}{(a+b)(a+c)(b+d)(c+d)}$	(-∞, ∞)
Stuart's τ_c [78]	$2(ad-bc)$	(-∞, ∞)
Tarantula [79] Ample [80]	$\frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}}=\frac{a(c+d)}{c(a+b)}$	[0, ∞)
Tarwid [81]	$\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$	[-1, 1)
Tversky [82] (Feature Contrast Model)	$a-b-c$	(-∞, ∞)
Warrens-I [83]	$\frac{2a-b-c}{2a+b+c}$	[-1, 1]
Warrens-II [83]	$\frac{2d}{b+c+2d}$	[0, 1]
Warrens-III [83]	$\frac{2d-b-c}{b+c+2d}$	[-1, 1]
Warrens-IV [83]	$\frac{4ad}{4ad+(a+d)(b+c)}$	[0, 1]
Warrens-V [83]	$\frac{ad-bc}{\min((a+b)(a+c),(c+d)(b+d))}$	[-1, ∞)

Similarity Measures (alternative names)	Equation	Range
Yule Q [84] (Coefficient of Association) Montgomery & Crittenden [85] (Gower & Legendre XV) [12]	$\frac{ad-bc}{ad+bc}$	[-1, 1]
Yule Y [64] (Coefficient of Colligation)	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	[-1, 1]

Table A-2 Binary Dissimilarity Measures

Dissimilarity Measures (alternative names)	Equation	Range
Chord [86]	$\sqrt{2(1-\frac{a}{\sqrt{(a+b)(a+c)}})}$	[0, √2]
Euclidean (Pythagorean metric)	$\sqrt{b+c}$	[0, ∞)
Hamming [40] Squared-Euclidean Canberra [87] Manhattan CityBlock Minkowski	$b+c$	[0, ∞)
Hellinger [88]	$2\sqrt{1-\frac{a}{\sqrt{(a+b)(a+c)}}}$	[0, 2]
Lance & Williams [89] Bray & Curtis [90]	$\frac{b+c}{(2a+b+c)}$	[0, 1]
Mean Manhattan	$\frac{b+c}{(a+b+c+d)}$	[0, 1]
Pattern Difference	$\frac{4bc}{(a+b+c+d)^2}$	[0, 1]
Shape Difference Baulieu [91]	$\frac{n(b+c)-(b-c)^2}{(a+b+c+d)^2}$	[0, 1]
Size Difference Baulieu [91]	$\frac{(b-c)^2}{(a+b+c+d)^2}$	[0, 1]
Variance	$\frac{(b+c)}{4(a+b+c+d)}$	[0, 0.25]
Yule Q dissimilarity [84]	$\frac{2bc}{ad+bc}$	[-1, 1]

A.1 Binary Measures References

[1] M. R. Anderberg, Cluster Analysis for Applications, Monographs and Textbooks on Probability and Mathematical Statistics, New York: Academic Press, Inc., 1973.

[2] C. Baroni-Urbani and M. W. Buser, "Similarity of Binary Data," *Systematic Zoology,* vol. 25, no. 3, pp. 251-259, 1976.

[3] V. Batagelj and M. Bren, "Comparing Resemblance Measures," *Journal of Classification,* vol. 12, pp. 73-90, 1995.

[4] R. Benini, Principii di Demographia. No. 29 of Manuali Barbera di Science Giuridiche Sociali e Politiche, Firenzi: G. Barbera, 1901.

[5] J. Braun-Blanquet, Plant Sociology: the Study of Plant Communities, New York: McGraw-Hill, 1932.

[6] P. W. Clement, "A Formula for Computing Inter-Observer Agreement," *Psychological Reports,* vol. 39, pp. 257-258, 1976.

[7] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement,* vol. 20, pp. 37-46, 1960.

[8] L. C. Cole, "The Measurement of Interspecific Association," *Ecology,* vol. 30, pp. 411-424, 1949.

[9] A. Ochiai, "Zoogeographic Studies on the Soleoid Fishes Found in Japan and its Neighbouring Regions," *Bulletin of the Japanese Society for Fish Science,* vol. 22, pp. 526-530, 1957.

[10] H. E. Driver and A. L. Kroeber, "Quantitative Expression of Cultural Relationships," *The University of California Publications in American Archaeology and Ethnology,* vol. 31, pp. 211-256, 1932.

[11] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association,* vol. 78, pp. 553-569, 1983.

[12] J. C. Gower and P. Legendre, "Metric and Euclidean Properties of Dissimilarity Coefficients," *Journal of Classification,* vol. 3, pp. 5-48, 1986.

[13] S. F. Dennis, "The Construction of a Thesaurus Automatically from a Sample of Text," in *Statistical Association Techniques for Mechanized Documentation: Symposium Proceedings*, Washington, DC, 1965.

[14] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology,* vol. 26, pp. 297-302, 1945.

[15] D. L. Wallace, "A Method for Comparing Two Hierarchical Clusterings: Comment," *Journal of the American Statistical Association,* vol. 78, pp. 569-576, 1983.

[16] W. J. Post and T. A. B. Snijders, "Nonparametric Unfolding Models for Dichotomous Data," *Methodika,* vol. 7, pp. 130-156, 1993.

[17] P. G. N. Digby, "Approximating the Tetrachoric Correlation Coefficient," *Biometrics,* vol. 39, pp. 753-757, 1983.

[18] M. H. Doolittle, "The Verification of Predictions," *Bulletin of the Philosophical Society of Washington,* vol. 7, pp. 122-127, 1885.

[19] H. Eyraud, "Les Principes de la Mesure des Correlations," *Annales de l'Universite de Lyon, Series A,* vol. 3, no. 1, pp. 3-47, 1936.

[20] E. W. Fager and J. A. McGowan, "Zooplankton Species Groups in the North Pacific," *Science,* vol. 140, pp. 453-460, 1963.

[21] D. P. Faith, P. R. Minchin and L. Belbin, "Compositional Dissimilarity as a Robust Measure of Ecological Distance," *Vegetatio,* vol. 69, pp. 57-68, 1987.

[22] J. L. Fleiss, "Measuring Agreement Between Two Judges on the Presence or Absence of a Trait," *Biometrics,* vol. 31, pp. 651-659, 1975.

[23] S. A. Forbes, "On the Local Distribution of Certain Illinois Fishes: An Essay in Statistical Ecology," *Bulletin of the Illinois State Laboratory for Natural History,* vol. 7, pp. 273-303, 1907.

[24] E. G. Fossum, Optimization and Standardization of Information Retrieval Language and Systems, Springfield, VA: Clearinghouse for Federal Scientific and Technical Information, 1966.

[25] P. E. Jones and R. M. Curtice, "A Framework for Comparing Document Term Association Measures," *American Documentation,* vol. 18, pp. 153-161, 1967.

[26] G. K. Gilbert, "Finley's Tornado Predictions," *American Meteorological Journal,* vol. 1, pp. 166-172, 1884.

[27] N. Gilbert and T. C. E. Wells, "Analysis of Quadrat Data," *Journal of Ecology,* vol. 54, no. 3, pp. 675-685, 1966.

[28] C. Gini, "Variabilita e Mutabilita," *Studi Economico-Giuridici dell'Universita di Cagliari,* vol. 3, pp. 1-158, 1912.

[29] H. A. Gleason, "Some Applications of the Quadrant Method," *Bulletin of the Torrey Botanical Club,* vol. 47, pp. 21-33, 1920.

[30] T. Sorenson, "A Method of Stabilizing Groups of Equivalent Amplitude in Plant Sociology Based on the Similarity of Species Content and its Applications to Analyses of the Vegetation on Danish Commons," *Kongelige Danske Videnskabernes Selskab Biologiske Skrifter,* vol. 5, pp. 1-34, 1948.

[31] J. Czekanowski, "Coefficient of Racial Likeness und Durchschnittliche Differenz," *Anthropologischer Anzeiger,* vol. 9, pp. 227-249, 1932.

[32] M. Nei and W.-H. Li, "Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases," *Proceedings of the National Academy of Sciences,* vol. 76, no. 10, pp. 5269-5273, 1979.

[33] L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classifications," *Journal of the American Statistical Association,* vol. 49, no. 338, pp. 732-764, 1954.

[34] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics,* vol. 27, no. 4, pp. 857-871, 1971.

[35] U. Hamann, "Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen," *Willdenowia,* vol. 2, pp. 639-768, 1961.

[36] J. W. Holley and J. P. Guilford, "A Note on the G-Index of Agreement," *Educational and Psychological Measurement,* vol. 24, pp. 749-753, 1964.

[37] L. J. Hubert, "Nominal Scale Response Agreement as a Generalized Correlation," *British Journal of Mathematical and Statistical Psychology,* vol. 30, pp. 98-103, 1977.

[38] F. C. Harris and B. B. Lahey, "A Method for Combining Occurrence and Nonoccurrence Interobserver Agreement Scores," *Journal of Applied Behavioral Analysis,* vol. 11, pp. 523-527, 1978.

[39] R. P. Hawkins and V. A. Dotson, "Reliability Scores That Delude: An Alice in Wonderland Trip Through the Misleading Characteristics of Inter-Observer Agreement Scores in Interval Recording," in *Behavior Analysis: Areas of Research and Application*, E. Ramp and G. Semb, Eds., Englewood Cliffs, N.J., Prentice-Hall, 1968.

[40] R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell System Technical Journal,* vol. 26, no. 2, pp. 147-160, 1950.

[41] P. Jaccard, "The Distribution of the Flora in the Alpine Zone," *The New Phytologist,* vol. 11, pp. 37-50, 1912.

[42] T. T. Tanimoto, "IBM Internal Report," 1957.

[43] S. C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika,* vol. 32, no. 3, pp. 241-254, 1967.

[44] R. N. Kent and S. L. Foster, "Direct Observational Procedures: Methodological Issues in Naturalistic Settings," in *Handbook of Behavioral Assessment*, A. R. Ciminero, K. S. Calhoun and H. E. Adams, Eds., New York, John Wiley & Sons, 1977, pp. 279-328.

[45] W. Koppen, "Die Aufeinanderfolge der Unperiodischen Witterungserscheinungen nach den Grundsatzen der Wahrscheinlichkeitsrechnung," in *Repertorium fur Meteorologie*, Petrograd, Akademiia Nauk, 1870-1, pp. 189-238.

[46] W. Koppen, "Eine Rationelle Methode zur Prufung der Wetterprognosen," *Meteorologische Zeitschrift,* vol. 1, pp. 397-404, 1884.

[47] G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika,* vol. 2, pp. 151-160, 1937.

[48] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika,* vol. 16, no. 3, pp. 297-334, 1951.

[49] J. L. Kuhns, "The Continuum of Coefficients of Association," in *Statistical Association Methods for Mechanized Documentation*, S. e. al., Ed., Washington, National Bureau of Standards, 1965, pp. 33-39.

[50] S. Kulczynski, "Die Pflanzenassociationen der Pienenen," *Bulletin International de L'Academie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles, Serie B, Supplement II,* vol. 2, pp. 57-203, 1927.

[51] J. A. Loevinger, "A Systematic Approach to the Construction and Evaluation of Tests of Ability," *Psychological Monograph,* vol. 61, no. 4, 1947.

[52] J. A. Loevinger, "The Technic of Homogeneous Tests Compared with Some Aspects of Scale Analysis and Factor Analysis," *Psychological Bulletin,* vol. 45, pp. 507-530, 1948.

[53] R. J. Mokken, A Theory and Procedure of Scale Analysis, The Hague, The Netherlands: Mouton, 1971.

[54] K. Sijtsma and I. W. Molenaar, Introduction to Nonparametric Item Response Theory, Thousand Oaks: Siege, 2002.

[55] M. E. Maron and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of the ACM,* vol. 7, pp. 216-224, 1960.

[56] A. E. Maxwell and A. E. G. Pilliner, "Deriving Coefficients of Reliability and Agreement for Ratings," *British Journal of Mathematical and Statistical Psychology,* vol. 21, pp. 105-116, 1968.

[57] B. H. McConnaughey, "The Determination and Analysis of Plankton Communities," *Marine Research, Special No, Indonesia,* pp. 1-40, 1964.

[58] E. L. Michael, "Marine Ecology and the Coefficient of Association: A Plea in Behalf of Quantitative Biology," *Journal of Animal Ecology,* vol. 8, pp. 54-59, 1920.

[59] M. D. Mountford, "An Index of Similarity and Its Application to Classificatory Problems," in *Progress in Soil Zoology*, London, Butterworths, 1962, pp. 43-50.

[60] K. Pearson and D. Heron, "On Theories of Association," *Biometrika,* vol. 9, pp. 159-315, 1913.

[61] K. Pearson, "Mathematical Contributions to the Theory of Evolution, XIII: On the Theory of Contingency and Its Relation to Association and Normal Correlation," in *Draper's Company Research Memoirs, Biometric Series II*, Cambridge University Press, 1904.

[62] K. Pearson, "On the Coefficient of Racial Likeness," *Biometrika,* vol. 9, pp. 105-117, 1926.

[63] C. S. Peirce, "The Numerical Measure of the Success of Predictions," *Science,* vol. 4, pp. 453-454, 1884.

[64] G. U. Yule, "On the Methods of Measuring Association Between Two Attributes," *Journal of the Royal Statistical Society, Series A,* vol. 75, no. 6, pp. 579-652, 1912.

[65] J. S. Rogers and T. T. Tanimoto, "A Computer Program for Classifying Plants," *Science,* vol. 132, pp. 1115-1118, 1960.

[66] G. M. Farkas, "Correction for Bias Present in a Method of Calculating Interobserver Agreement," *Journal of Applied Behavior Analysis,* vol. 11, p. 188, 1978.

[67] E. Rogot and I. D. Goldberg, "A Proposed Index for Measuring Agreement in Test-Retest Studies," *Journal of Chronic Disease,* vol. 19, pp. 991-1006, 1966.

[68] P. F. Russel and T. R. Rao, "On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras," *Journal of Malaria Institute India,* vol. 3, pp. 153-178, 1940.

[69] W. A. Scott, "Reliability of Content Analysis: The Case of Nominal Scale Coding," *Public Opinion Quarterly,* vol. 19, pp. 321-325, 1955.

[70] G. G. Simpson, "Mammals and the Nature of Continents," *American Journal of Science,* vol. 241, pp. 1-31, 1943.

[71] R. R. Sokal and C. D. Michener, "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Science Bulletin,* vol. 38, pp. 1409-1438, 1958.

[72] W. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association,* vol. 66, pp. 846-850, 1971.

[73] R. L. Brennan and R. J. Light, "Measuring Agreement When Two Observers Classify People into Categories Not Defined in Advance," *British Journal of Mathematical and Statistical Psychology,* vol. 27, pp. 154-163, 1974.

[74] R. R. Sokal and P. H. Sneath, Principles of Numerical Taxonomy, San Francisco: W. H. Freeman and Company, 1963.

[75] T. Sorgenfrei, Molluscan Assemblages From the Marine Middle Miocene of South Jutland and Their Environments, Copenhagen: Reitzel, 1958.

[76] A. H. Cheetham and J. E. Hazel, "Binary (Presence-Absence) Similarity Coefficients," *Journal of Paleontology,* vol. 43, no. 5, pp. 1130-1136, 1969.

[77] H. E. Stiles, "The Association Factor in Information Retrieval," *Journal of the Association for Computing Machinery,* vol. 8, pp. 271-279, 1961.

[78] A. Stuart, "The Estimation and Comparison of Strengths of Association in Contingency Tables," *Biometrika,* vol. 40, pp. 105-110, 1953.

[79] J. Jones, M. J. Harrold and J. Stasko, "Visualization for Fault Localization," in *Proceedings of the Workshop on Software Visualization, 23rd International Conference on Software Engineering*, Toronto, Ontario, 2001.

[80] V. Dallmeier, C. Lindig and A. Zeller, "Lightweight Defect Localization for Java," in *Proceedings of the 19th European Conference on Object-Oriented Programming, ECOOP 2005*, 2005.

[81] K. Tarwid, "Szacowanie Zbieznosci Nisz Ekologicznych Gatunkow Droga Oceny Prawdopodobienstwa Spotykania Sie Ich W Polowach," *Ecologia Polska, Series B,* vol. 6, pp. 115-130, 1960.

[82] A. Tversky, "Features of Similarity," *Psychological Review,* vol. 84, no. 4, pp. 327-352, 1977.

[83] M. J. Warrens, "Similarity Coefficients for Binary Data," Leiden, Netherlands, 2008.

[84] G. U. Yule, "On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society," *Philosophical Transactions of the Royal Society, Series A,* vol. 194, pp. 257-319, 1900.

[85] A. C. Montgomery and K. S. Crittenden, "Improving Coding Reliability for Open-Ended Questions," *Public Opinion Quarterly,* vol. 41, pp. 235-243, 1977.

[86] L. Orloci, "An Agglomerative Method for Classification of Plant Communities," *Journal of Ecology,* vol. 55, no. 1, pp. 193-205, 1967b.

[87] G. N. Lance and W. T. Williams, "Computer Programs for Classification," in *Proceedings of the ANCAC Conference*, Canberra, Australia, 1966c.

[88] C. R. Rao, "A Review of Canonical Coordinates and an Alternative to Correspondence Analysis Using Hellinger Distance," *Questiio (Quaderns d'Estadistica i Investigacio Operativa),* vol. 19, pp. 23-63, 1995.

[89] G. N. Lance and W. T. Williams, "Mixed-Data Classificatory Programs I -- Agglomerative Systems," *Australian Computer Journal,* vol. 1, no. 1, pp. 15-20, 1967.

[90] J. R. Bray and J. T. Curtis, " An Ordination of the Upland Forest of the Southern Wisconsin," *Ecological Monographs,* vol. 27, no. 4, pp. 325-349, 1957.

[91] F. B. Baulieu, "A Classification of Presence/Absence Based Dissimilarity Coefficients," *Journal of Classification,* vol. 6, pp. 233-246, 1989.

Appendix B Median Regression Coefficients

Table B-1 Median Regression Coefficients

Regressor	Coefficient	Lower CI	Upper CI	Significant (α = .05)
(Intercept)	2.55	2.43	2.70	yes
A	-0.44	-0.61	-0.31	yes
B	-0.14	-0.41	0.04	
C	-1.97	-2.16	-1.84	yes
D	0.67	0.53	0.91	yes
MeanPath	-0.98	-1.18	-0.67	yes
alpha	0.29	0.20	0.55	yes
DegreeCorr	-0.69	-0.91	-0.36	yes
A:B	0.02	-0.08	0.27	
A:C	0.19	0.08	0.39	yes
B:C	0.01	-0.13	0.24	
A:D	-0.01	-0.33	0.12	
B:D	-0.08	-0.34	0.13	
C:D	-0.46	-0.66	-0.22	yes
A:MeanPath	0.47	-0.12	0.65	
B:MeanPath	-0.40	-0.70	-0.04	yes
C:MeanPath	0.28	0.06	0.54	yes
D:MeanPath	0.06	-0.20	0.32	
A:alpha	0.14	-0.06	0.30	
B:alpha	-0.09	-0.32	0.11	
C:alpha	-0.08	-0.42	0.07	
D:alpha	-0.01	-0.16	0.21	
MeanPath:alpha	-0.01	-0.50	0.47	
A:DegreeCorr	0.62	0.04	0.82	yes
B:DegreeCorr	0.10	-0.23	0.38	
C:DegreeCorr	-0.03	-0.30	0.27	
D:DegreeCorr	-0.10	-0.33	0.13	
MeanPath:DegreeCorr	-0.07	-0.22	0.07	
alpha:DegreeCorr	-0.08	-0.60	0.38	
A:B:C	0.16	-0.17	0.27	
A:B:D	-0.03	-0.24	0.23	
A:C:D	-0.13	-0.38	0.09	
B:C:D	0.07	-0.30	0.23	
A:B:MeanPath	0.58	0.05	0.98	yes
A:C:MeanPath	0.61	0.22	1.02	yes
B:C:MeanPath	-0.49	-0.79	-0.08	yes
A:D:MeanPath	0.02	-0.18	0.42	
B:D:MeanPath	0.25	-0.13	0.73	

Regressor	Coefficient	Lower CI	Upper CI	Significant (α = .05)
C:D:MeanPath	-0.12	-0.39	0.18	
A:B:alpha	0.36	0.07	0.58	yes
A:C:alpha	-0.40	-0.51	-0.04	yes
B:C:alpha	-0.05	-0.25	0.26	
A:D:alpha	0.37	0.18	0.51	yes
B:D:alpha	0.28	0.10	0.44	yes
C:D:alpha	-0.17	-0.32	0.22	
A:MeanPath:alpha	0.41	-0.42	0.78	
B:MeanPath:alpha	-0.02	-0.85	0.45	
C:MeanPath:alpha	-0.15	-0.94	0.33	
D:MeanPath:alpha	0.02	-0.67	0.38	
A:B:DegreeCorr	0.10	-0.26	0.42	
A:C:DegreeCorr	0.43	0.10	0.70	yes
B:C:DegreeCorr	-0.25	-0.47	0.21	
A:D:DegreeCorr	0.35	0.12	0.68	yes
B:D:DegreeCorr	0.45	0.11	0.81	yes
C:D:DegreeCorr	-0.62	-0.98	-0.29	yes
A:MeanPath:DegreeCorr	-0.05	-0.34	0.17	
B:MeanPath:DegreeCorr	-0.06	-0.18	0.18	
C:MeanPath:DegreeCorr	0.14	-0.16	0.34	
D:MeanPath:DegreeCorr	0.04	-0.04	0.20	
A:alpha:DegreeCorr	0.29	-0.45	0.66	
B:alpha:DegreeCorr	0.27	-0.21	0.65	
C:alpha:DegreeCorr	-0.22	-0.97	0.37	
D:alpha:DegreeCorr	-0.35	-0.87	0.15	
MeanPath:alpha:DegreeCorr	0.00	-0.15	0.20	
A:B:C:D	0.11	-0.05	0.43	
A:B:C:MeanPath	-0.06	-0.41	0.42	
A:B:D:MeanPath	0.09	-0.26	0.52	
A:C:D:MeanPath	0.09	-0.16	0.51	
B:C:D:MeanPath	-0.58	-1.15	0.03	
A:B:C:alpha	-0.30	-0.53	0.00	yes
A:B:D:alpha	-0.18	-0.39	0.06	
A:C:D:alpha	-0.07	-0.37	0.04	
B:C:D:alpha	-0.05	-0.24	0.15	
A:B:MeanPath:alpha	0.25	-0.39	0.94	
A:C:MeanPath:alpha	0.24	-0.41	0.81	
B:C:MeanPath:alpha	0.18	-0.43	0.90	
A:D:MeanPath:alpha	-0.11	-0.63	0.76	
B:D:MeanPath:alpha	0.06	-0.65	0.73	
C:D:MeanPath:alpha	-0.33	-1.12	0.37	
A:B:C:DegreeCorr	-0.29	-0.66	0.07	
A:B:D:DegreeCorr	-0.31	-0.51	0.17	

Regressor	Coefficient	Lower CI	Upper CI	Significant (α = .05)
A:C:D:DegreeCorr	0.51	0.26	0.88	yes
B:C:D:DegreeCorr	0.10	-0.41	0.30	
A:B:MeanPath:DegreeCorr	0.18	-0.09	0.36	
A:C:MeanPath:DegreeCorr	-0.22	-0.44	0.13	
B:C:MeanPath:DegreeCorr	-0.10	-0.37	0.14	
A:D:MeanPath:DegreeCorr	-0.13	-0.32	0.03	
B:D:MeanPath:DegreeCorr	-0.24	-0.46	0.12	
C:D:MeanPath:DegreeCorr	0.25	-0.14	0.46	
A:B:alpha:DegreeCorr	-0.09	-0.78	0.89	
A:C:alpha:DegreeCorr	0.19	-0.42	0.83	
B:C:alpha:DegreeCorr	0.07	-0.69	0.46	
A:D:alpha:DegreeCorr	-0.06	-0.62	0.68	
B:D:alpha:DegreeCorr	0.11	-0.64	0.80	
C:D:alpha:DegreeCorr	-0.19	-0.92	0.59	
A:MeanPath:alpha:DegreeCorr	0.44	0.01	0.71	yes
B:MeanPath:alpha:DegreeCorr	-0.11	-0.38	0.12	
C:MeanPath:alpha:DegreeCorr	0.12	-0.28	0.35	
D:MeanPath:alpha:DegreeCorr	-0.04	-0.27	0.13	
A:B:C:D:MeanPath	0.03	-0.83	0.38	
A:B:C:D:alpha	-0.02	-0.26	0.10	
A:B:C:MeanPath:alpha	0.34	-0.54	0.97	
A:B:D:MeanPath:alpha	-0.77	-1.48	-0.28	yes
A:C:D:MeanPath:alpha	-0.78	-1.33	-0.22	yes
B:C:D:MeanPath:alpha	0.40	-0.30	1.06	
A:B:C:D:DegreeCorr	-0.49	-0.80	-0.10	yes
A:B:C:MeanPath:DegreeCorr	-0.10	-0.41	0.22	
A:B:D:MeanPath:DegreeCorr	0.17	-0.18	0.34	
A:C:D:MeanPath:DegreeCorr	-0.12	-0.36	0.21	
B:C:D:MeanPath:DegreeCorr	-0.54	-0.94	-0.11	yes
A:B:C:alpha:DegreeCorr	0.01	-1.09	0.64	
A:B:D:alpha:DegreeCorr	-0.81	-1.54	-0.30	yes
A:C:D:alpha:DegreeCorr	0.03	-0.58	0.80	
B:C:D:alpha:DegreeCorr	0.75	0.15	1.57	yes
A:B:MeanPath:alpha:DegreeCorr	0.19	-0.07	0.72	
A:C:MeanPath:alpha:DegreeCorr	-0.58	-0.94	0.03	
B:C:MeanPath:alpha:DegreeCorr	-0.48	-0.84	0.04	
A:D:MeanPath:alpha:DegreeCorr	-0.22	-0.49	0.16	
B:D:MeanPath:alpha:DegreeCorr	0.01	-0.22	0.41	
C:D:MeanPath:alpha:DegreeCorr	0.41	0.10	0.74	yes
A:B:C:D:MeanPath:alpha	0.43	-0.07	1.12	
A:B:C:D:MeanPath:DegreeCorr	0.78	0.30	1.06	yes
A:B:C:D:alpha:DegreeCorr	-0.12	-0.78	0.52	
A:B:C:MeanPath:alpha:DegreeCorr	-0.61	-1.23	-0.21	yes

Regressor	Coefficient	Lower CI	Upper CI	Significant (α = .05)
A:B:D:MeanPath:alpha:DegreeCorr	-0.12	-0.66	0.13	
A:C:D:MeanPath:alpha:DegreeCorr	0.50	-0.01	1.11	
B:C:D:MeanPath:alpha:DegreeCorr	-0.15	-0.70	0.25	
A:B:C:D:MeanPath:alpha:DegreeCorr	0.43	0.05	1.02	yes

Appendix C Multiple Quantile Regression Models

Table C-1 Summary of Effects Across Multiple Quantile Regression Models

Regressor	Effect	Significant (α = .05)
(Intercept)	Positive	yes
A	Negative	yes
B	Negative	yes
C	Negative	yes
D	Positive	yes
MeanPath	Negative	yes
alpha	Positive	yes
DegreeCorr	Negative	yes
A:B	Positive	
A:C	Positive	yes
B:C	Mixed	
A:D	Mixed	
B:D	Negative	
C:D	Negative	yes
A:MeanPath	Positive	
B:MeanPath	Negative	yes
C:MeanPath	Negative	
D:MeanPath	Mixed	
A:alpha	Mixed	
B:alpha	Mixed	
C:alpha	Positive	
D:alpha	Mixed	
MeanPath:alpha	Mixed	
A:DegreeCorr	Positive	yes
B:DegreeCorr	Mixed	
C:DegreeCorr	Negative	
D:DegreeCorr	Mixed	
MeanPath:DegreeCorr	Negative	
alpha:DegreeCorr	Mixed	
A:B:C	Mixed	
A:B:D	Mixed	
A:C:D	Mixed	
B:C:D	Mixed	
A:B:MeanPath	Negative	
A:C:MeanPath	Positive	
B:C:MeanPath	Negative	yes
A:D:MeanPath	Positive	
B:D:MeanPath	Positive	yes

Regressor	Effect	Significant (α = .05)
C:D:MeanPath	Mixed	
A:B:alpha	Positive	
A:C:alpha	Negative	yes
B:C:alpha	Negative	
A:D:alpha	Positive	yes
B:D:alpha	Mixed	
C:D:alpha	Mixed	
A:MeanPath:alpha	Positive	
B:MeanPath:alpha	Mixed	
C:MeanPath:alpha	Negative	
D:MeanPath:alpha	Mixed	
A:B:DegreeCorr	Mixed	
A:C:DegreeCorr	Positive	yes
B:C:DegreeCorr	Negative	
A:D:DegreeCorr	Positive	yes
B:D:DegreeCorr	Positive	yes
C:D:DegreeCorr	Negative	yes
A:MeanPath:DegreeCorr	Mixed	
B:MeanPath:DegreeCorr	Mixed	
C:MeanPath:DegreeCorr	Mixed	
D:MeanPath:DegreeCorr	Mixed	
A:alpha:DegreeCorr	Mixed	
B:alpha:DegreeCorr	Positive	
C:alpha:DegreeCorr	Negative	
D:alpha:DegreeCorr	Mixed	
MeanPath:alpha:DegreeCorr	Mixed	
A:B:C:D	Mixed	yes
A:B:C:MeanPath	Mixed	
A:B:D:MeanPath	Mixed	
A:C:D:MeanPath	Mixed	
B:C:D:MeanPath	Negative	yes
A:B:C:alpha	Mixed	
A:B:D:alpha	Mixed	
A:C:D:alpha	Negative	yes
B:C:D:alpha	Mixed	
A:B:MeanPath:alpha	Mixed	
A:C:MeanPath:alpha	Mixed	
B:C:MeanPath:alpha	Positive	
A:D:MeanPath:alpha	Mixed	
B:D:MeanPath:alpha	Mixed	
C:D:MeanPath:alpha	Negative	
A:B:C:DegreeCorr	Negative	
A:B:D:DegreeCorr	Negative	

Regressor	Effect	Significant (α = .05)
A:C:D:DegreeCorr	Positive	yes
B:C:D:DegreeCorr	Mixed	
A:B:MeanPath:DegreeCorr	Mixed	
A:C:MeanPath:DegreeCorr	Negative	
B:C:MeanPath:DegreeCorr	Mixed	
A:D:MeanPath:DegreeCorr	Negative	
B:D:MeanPath:DegreeCorr	Mixed	
C:D:MeanPath:DegreeCorr	Mixed	
A:B:alpha:DegreeCorr	Negative	
A:C:alpha:DegreeCorr	Mixed	
B:C:alpha:DegreeCorr	Mixed	
A:D:alpha:DegreeCorr	Mixed	
B:D:alpha:DegreeCorr	Mixed	
C:D:alpha:DegreeCorr	Mixed	
A:MeanPath:alpha:DegreeCorr	Positive	yes
B:MeanPath:alpha:DegreeCorr	Mixed	
C:MeanPath:alpha:DegreeCorr	Positive	
D:MeanPath:alpha:DegreeCorr	Mixed	
A:B:C:D:MeanPath	Mixed	
A:B:C:D:alpha	Mixed	
A:B:C:MeanPath:alpha	Positive	
A:B:D:MeanPath:alpha	Negative	
A:C:D:MeanPath:alpha	Mixed	
B:C:D:MeanPath:alpha	Mixed	
A:B:C:D:DegreeCorr	Negative	
A:B:C:MeanPath:DegreeCorr	Mixed	
A:B:D:MeanPath:DegreeCorr	Positive	
A:C:D:MeanPath:DegreeCorr	Positive	
B:C:D:MeanPath:DegreeCorr	Negative	yes
A:B:C:alpha:DegreeCorr	Positive	
A:B:D:alpha:DegreeCorr	Negative	yes
A:C:D:alpha:DegreeCorr	Mixed	yes
B:C:D:alpha:DegreeCorr	Mixed	
A:B:MeanPath:alpha:DegreeCorr	Positive	
A:C:MeanPath:alpha:DegreeCorr	Negative	yes
B:C:MeanPath:alpha:DegreeCorr	Negative	yes
A:D:MeanPath:alpha:DegreeCorr	Mixed	yes
B:D:MeanPath:alpha:DegreeCorr	Mixed	
C:D:MeanPath:alpha:DegreeCorr	Mixed	yes
A:B:C:D:MeanPath:alpha	Positive	
A:B:C:D:MeanPath:DegreeCorr	Positive	yes
A:B:C:D:alpha:DegreeCorr	Mixed	
A:B:C:MeanPath:alpha:DegreeCorr	Negative	yes

Regressor	Effect	Significant (α = .05)
A:B:D:MeanPath:alpha:DegreeCorr	Mixed	yes
A:C:D:MeanPath:alpha:DegreeCorr	Positive	yes
B:C:D:MeanPath:alpha:DegreeCorr	Mixed	
A:B:C:D:MeanPath:alpha:DegreeCorr	Positive	yes

Appendix D Quantile Regression Coefficient Plots

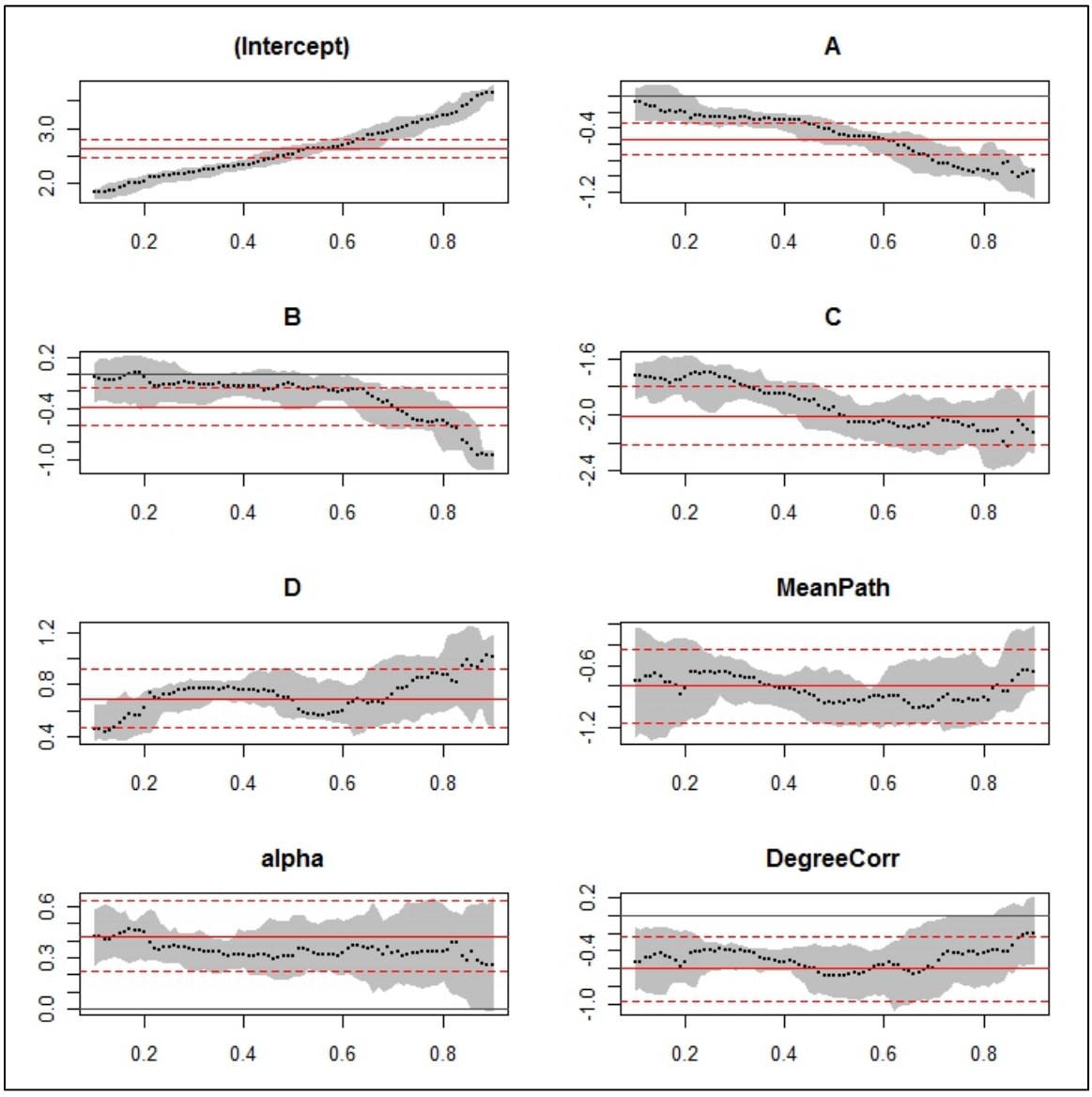

Figure D-1 Main Effects' Coefficients

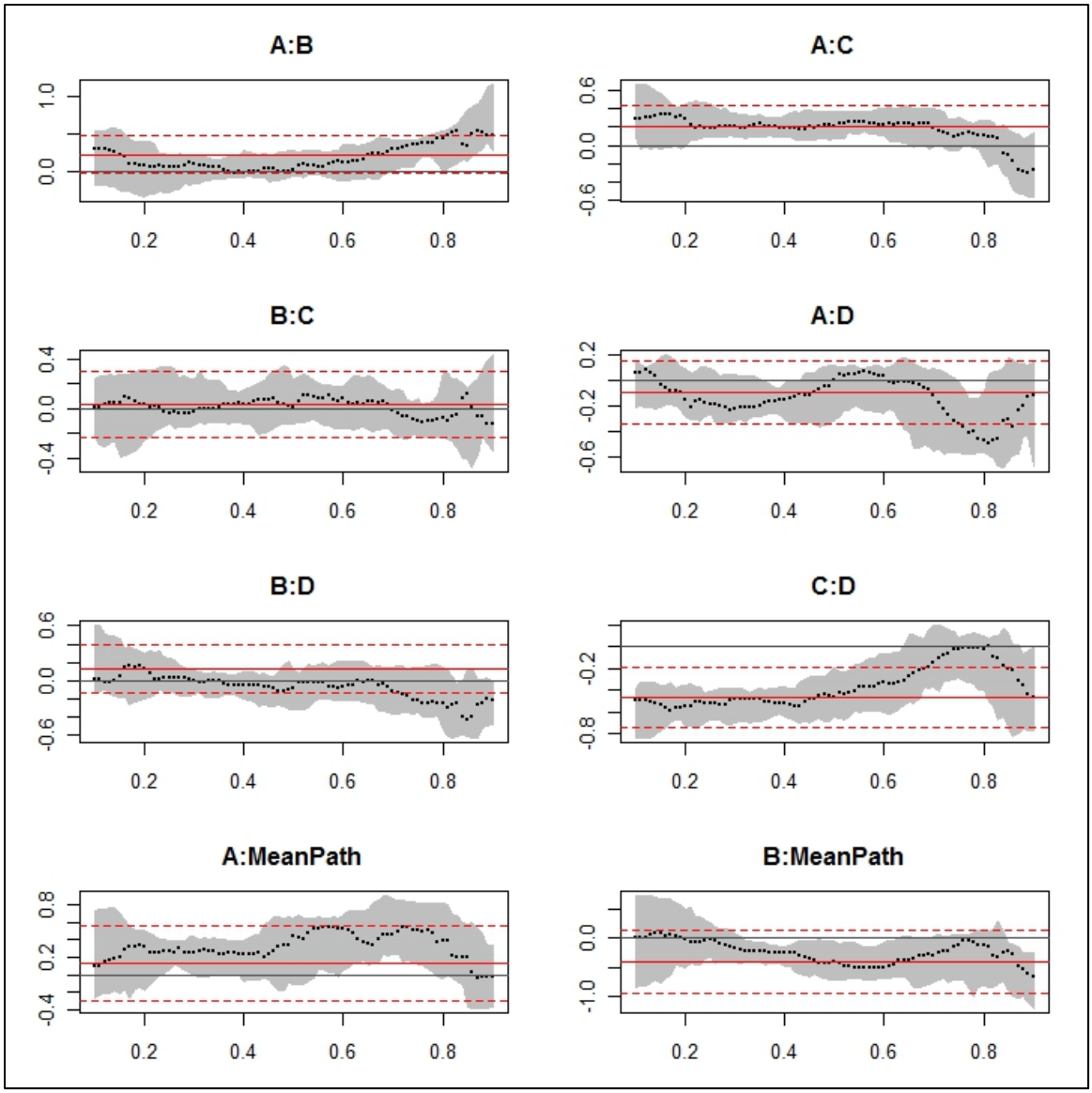

Figure D-2 Two Factor Interactions' Coefficients

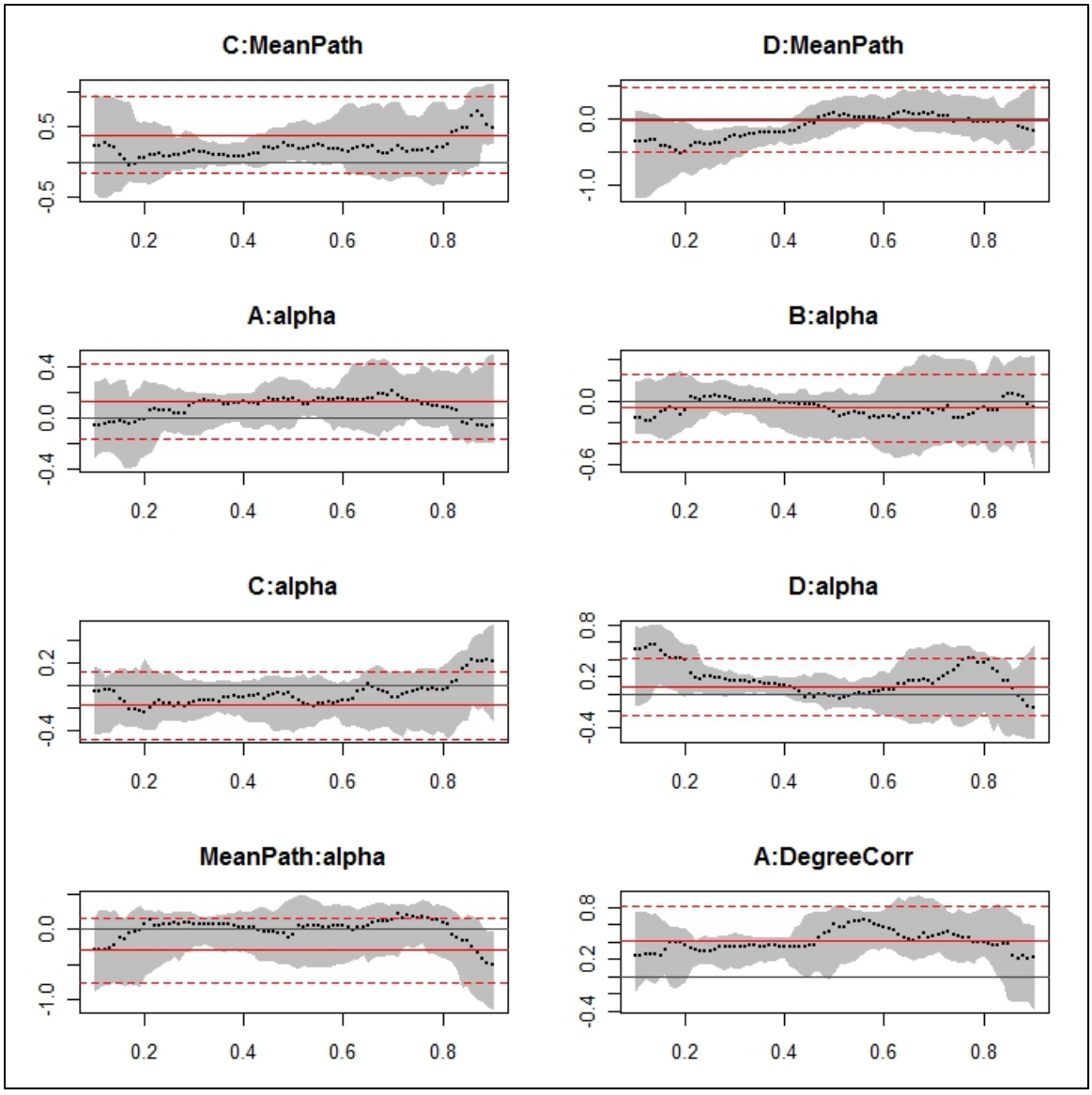

Figure D-3 Two Factor Interactions' Coefficients (cont.)

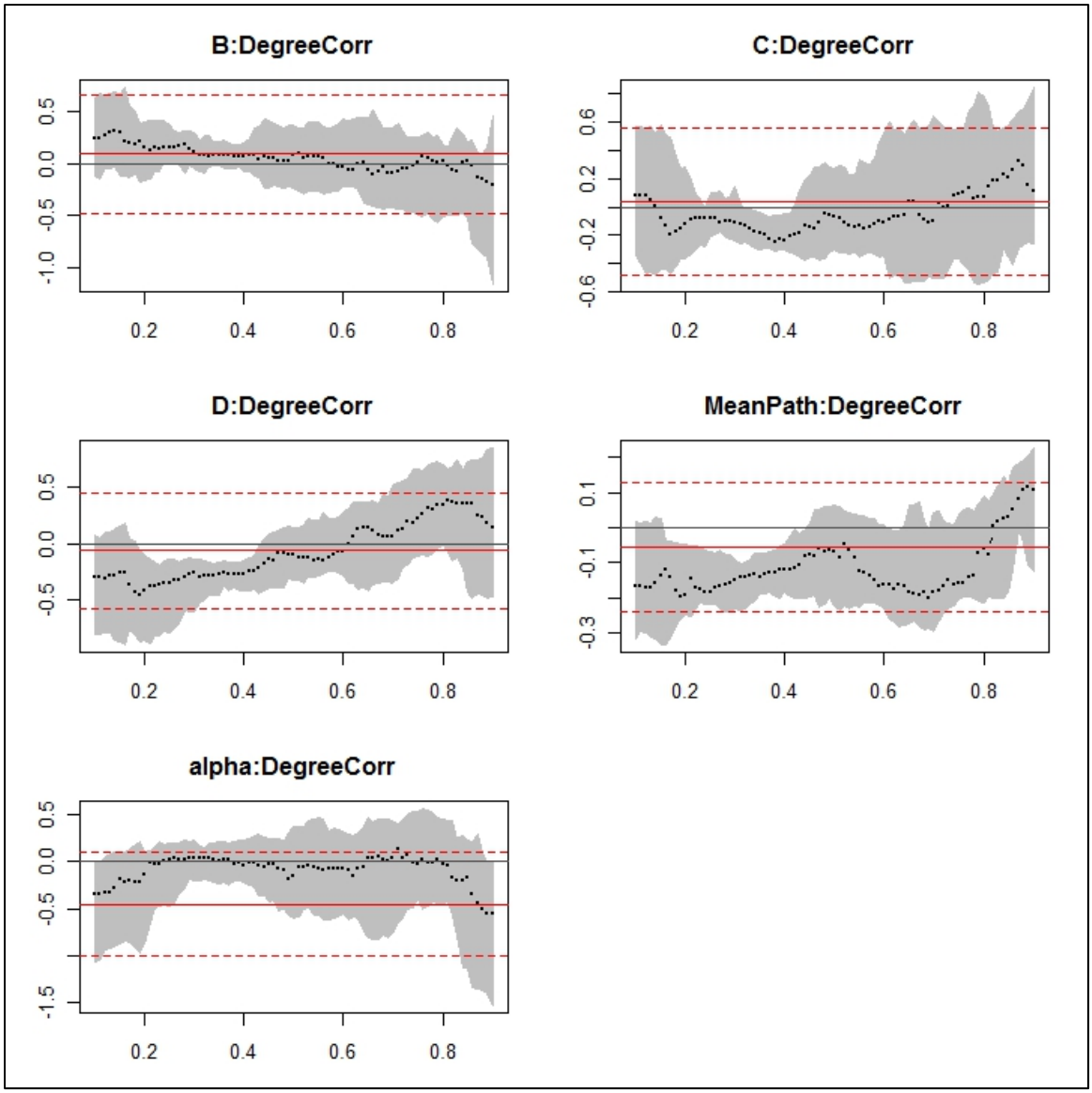

Figure D-4 Two Factor Interactions' Coefficients (cont.)

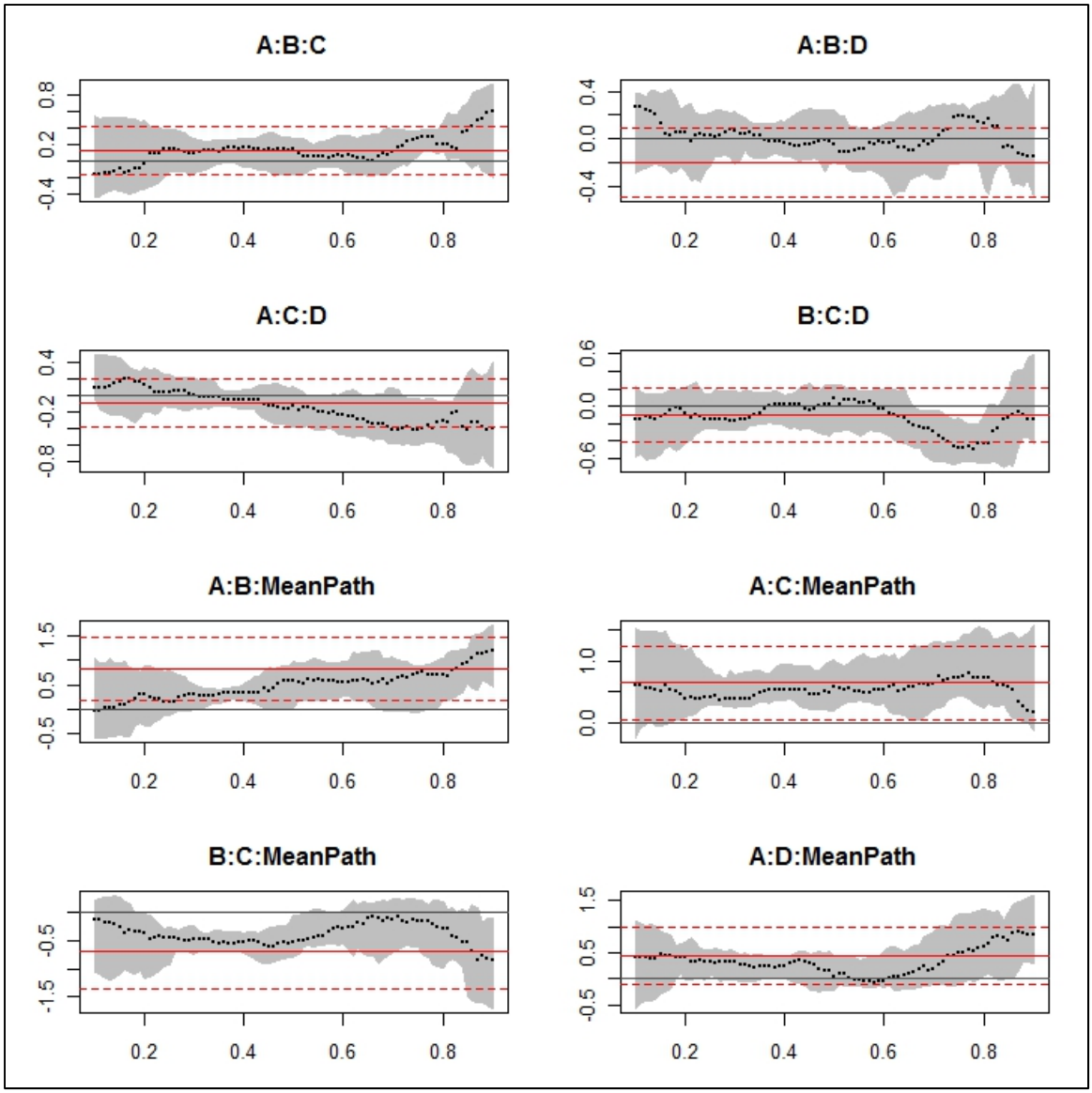

Figure D-5 Three Factor Interactions' Coefficients

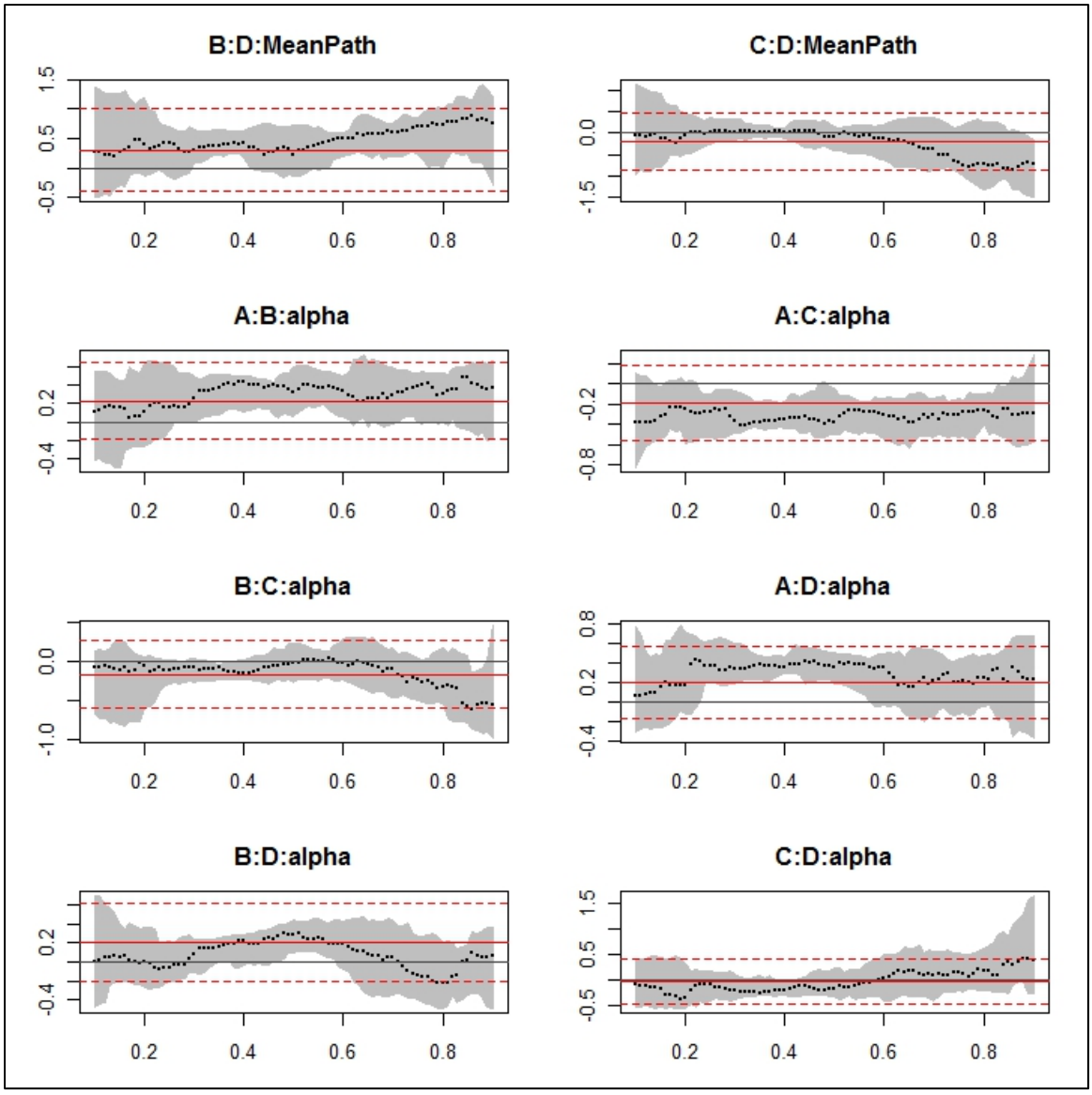

Figure D-6 Three Factor Interactions' Coefficients (cont.)

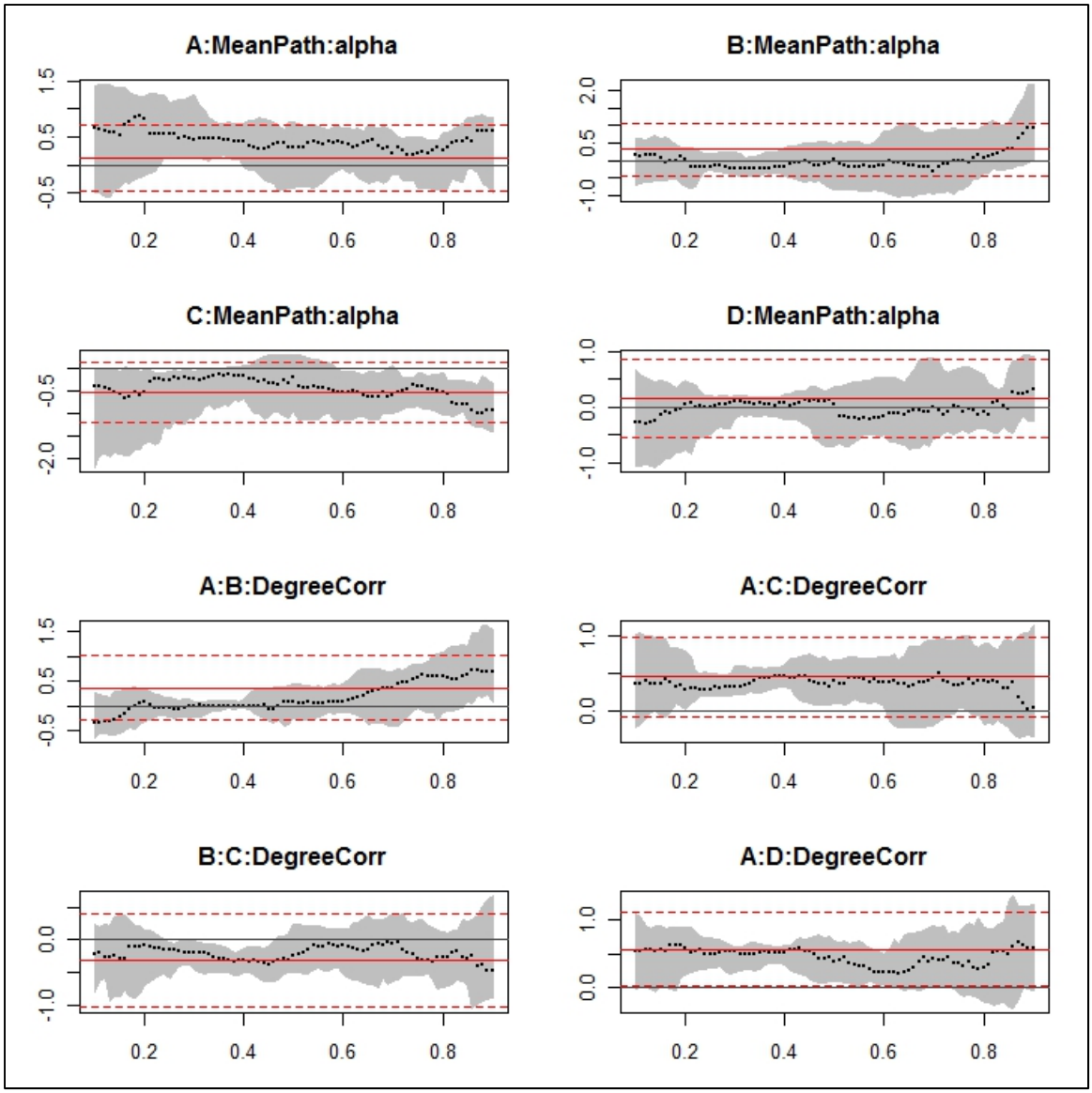

Figure D-7 Three Factor Interactions' Coefficients (cont.)

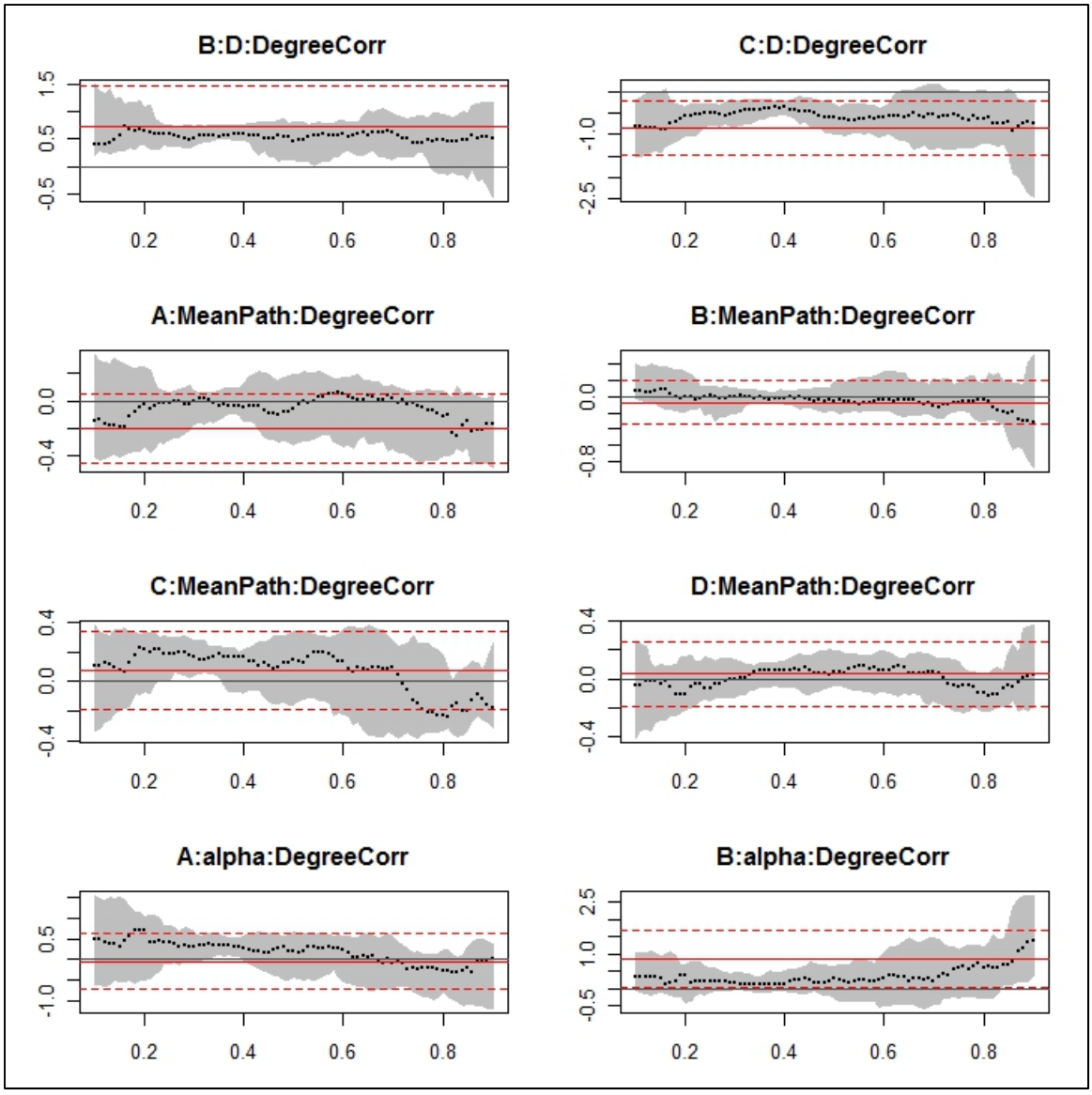

Figure D-8 Three Factor Interactions' Coefficients (cont.)

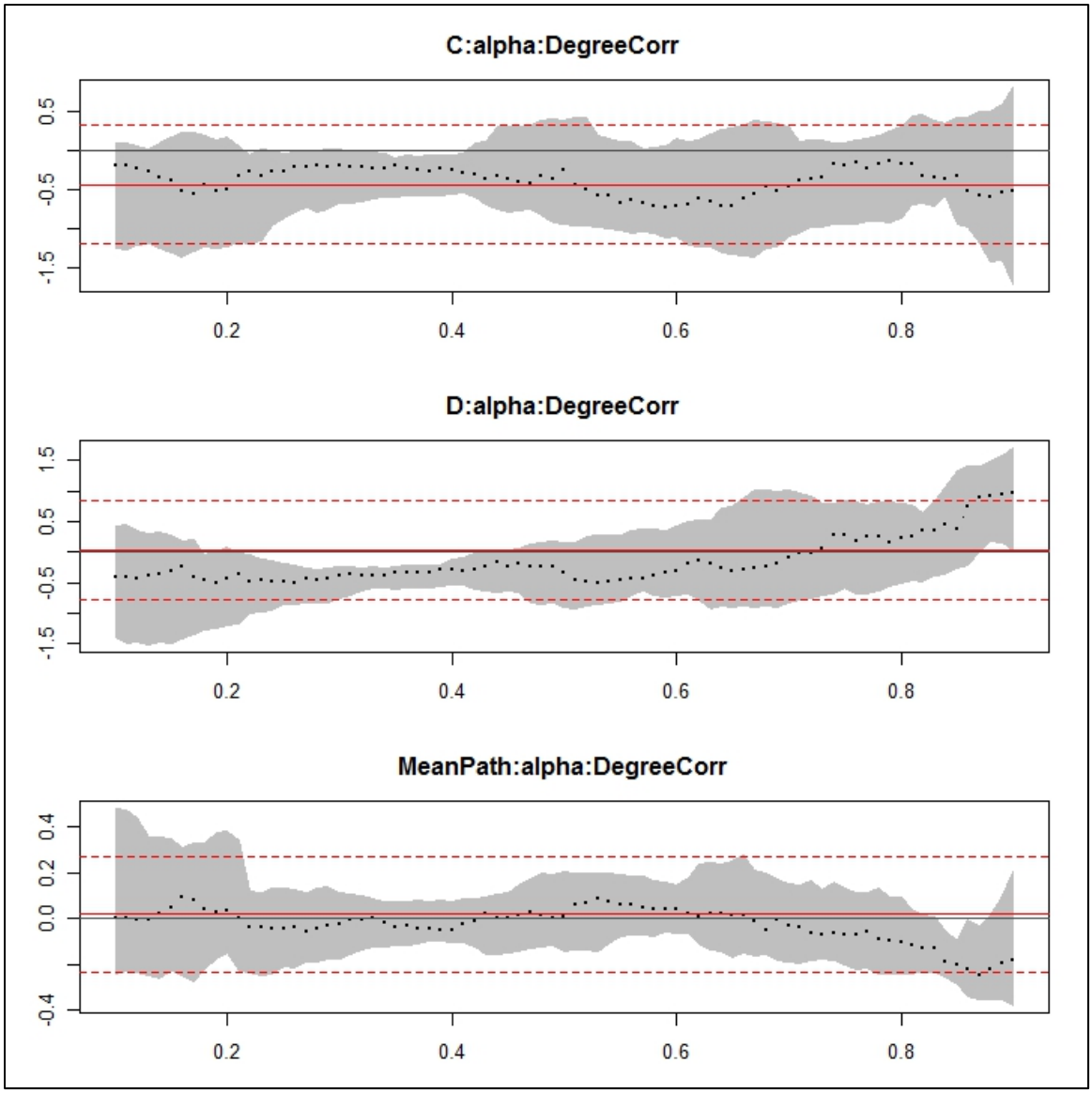

Figure D-9 Three Factor Interactions' Coefficients (cont.)

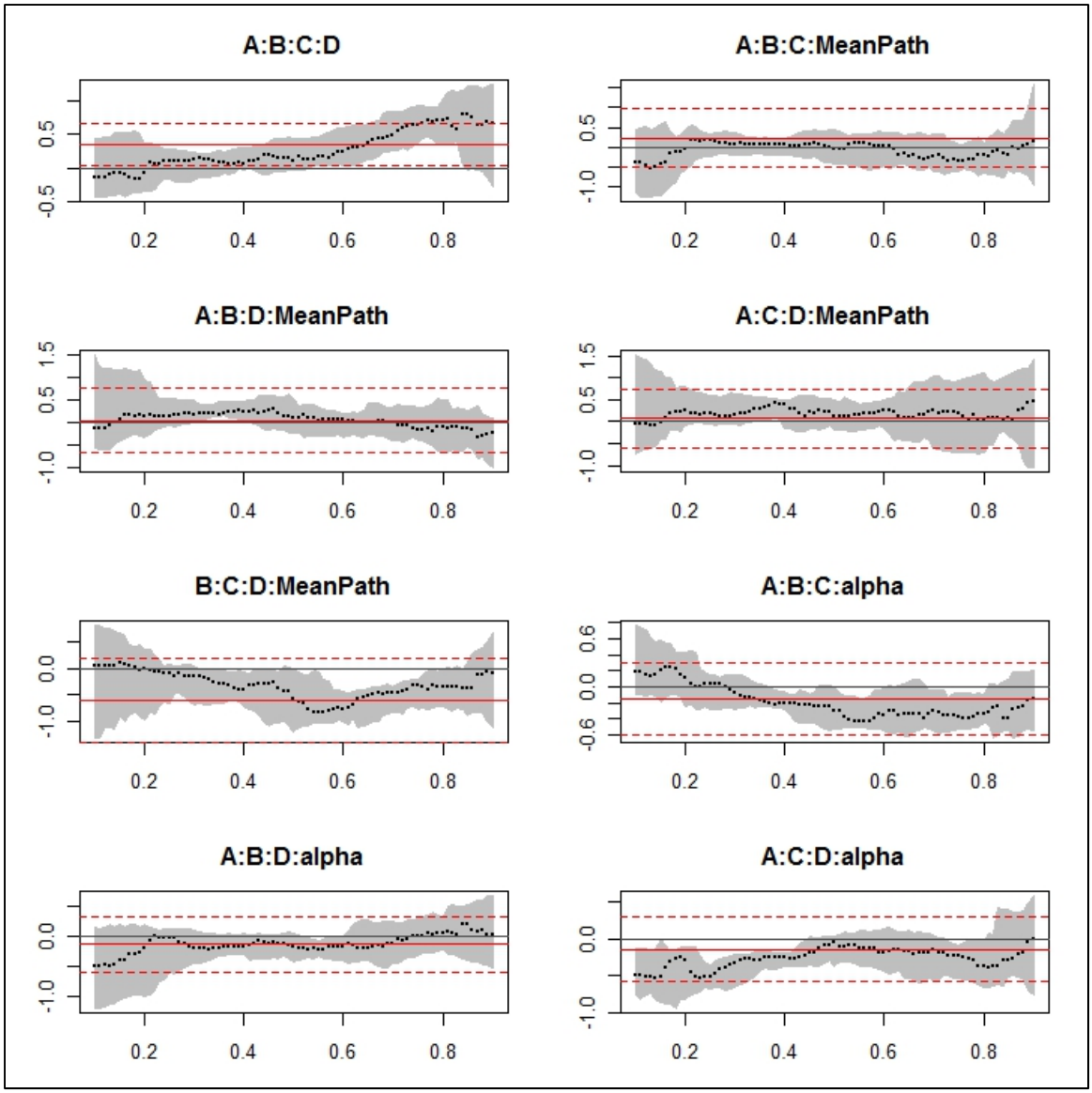

Figure D-10 Four Factor Interactions' Coefficients

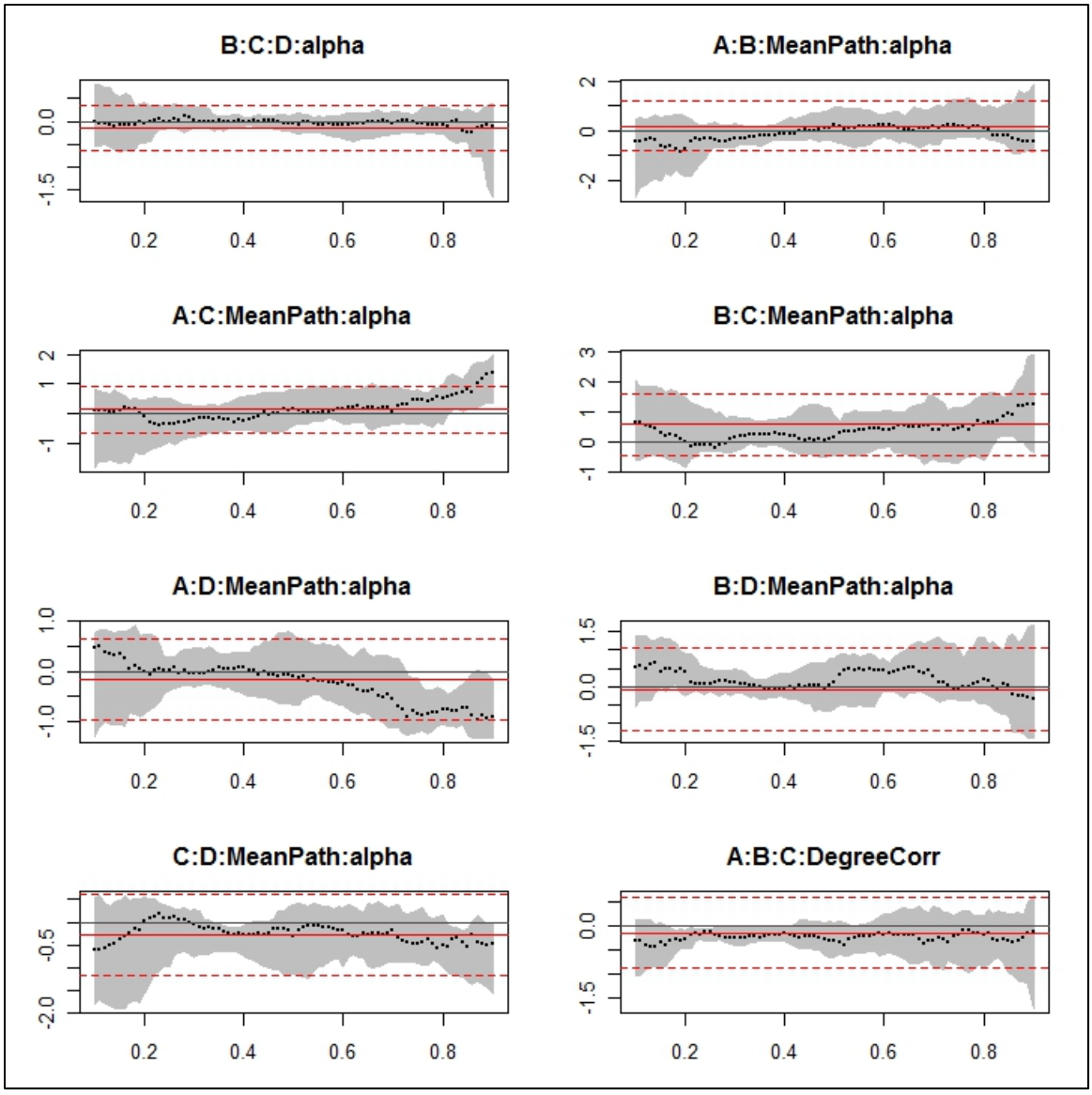

Figure D-11 Four Factor Interactions' Coefficients (cont.)

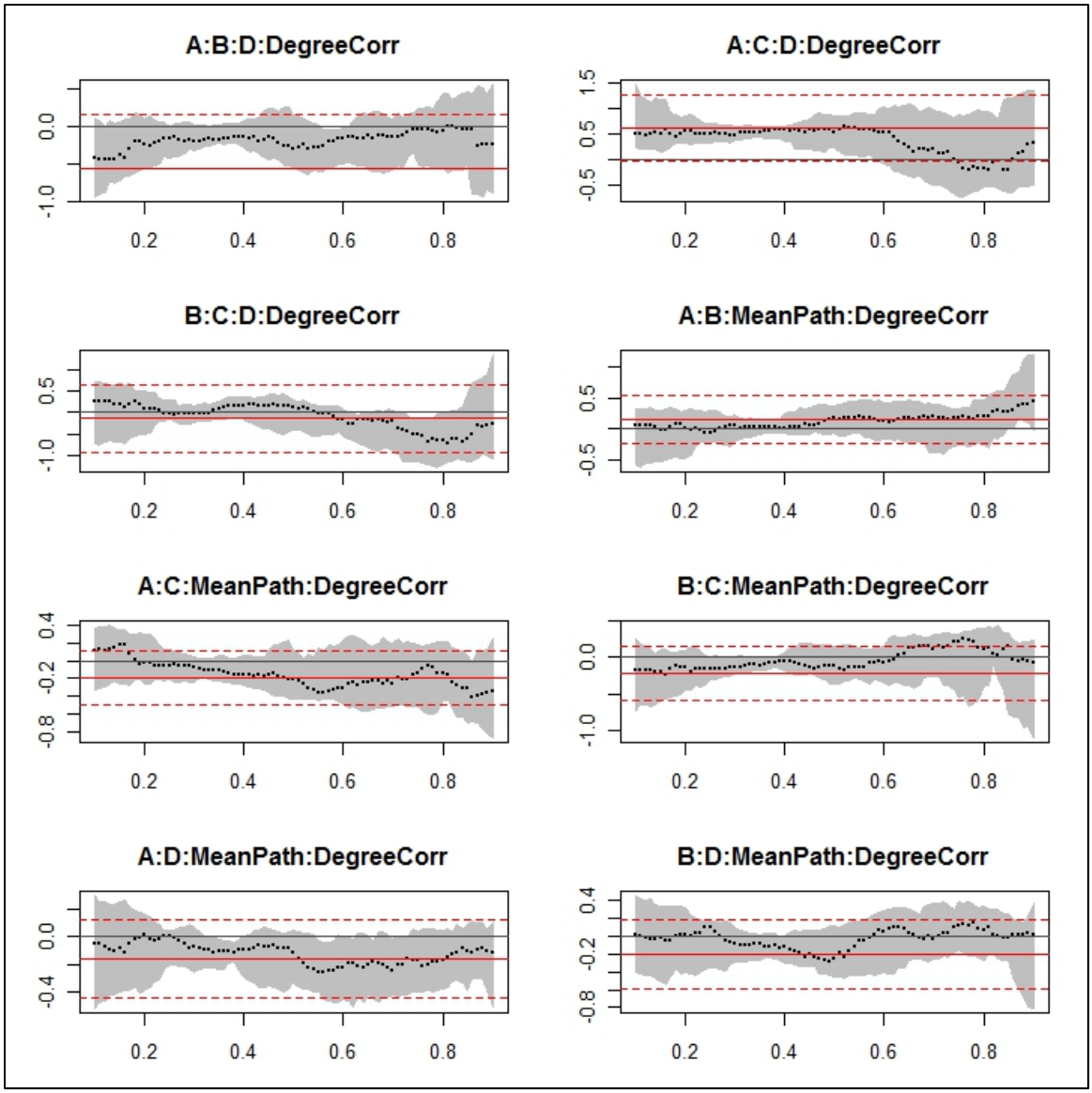

Figure D-12 Four Factor Interactions' Coefficients (cont.)

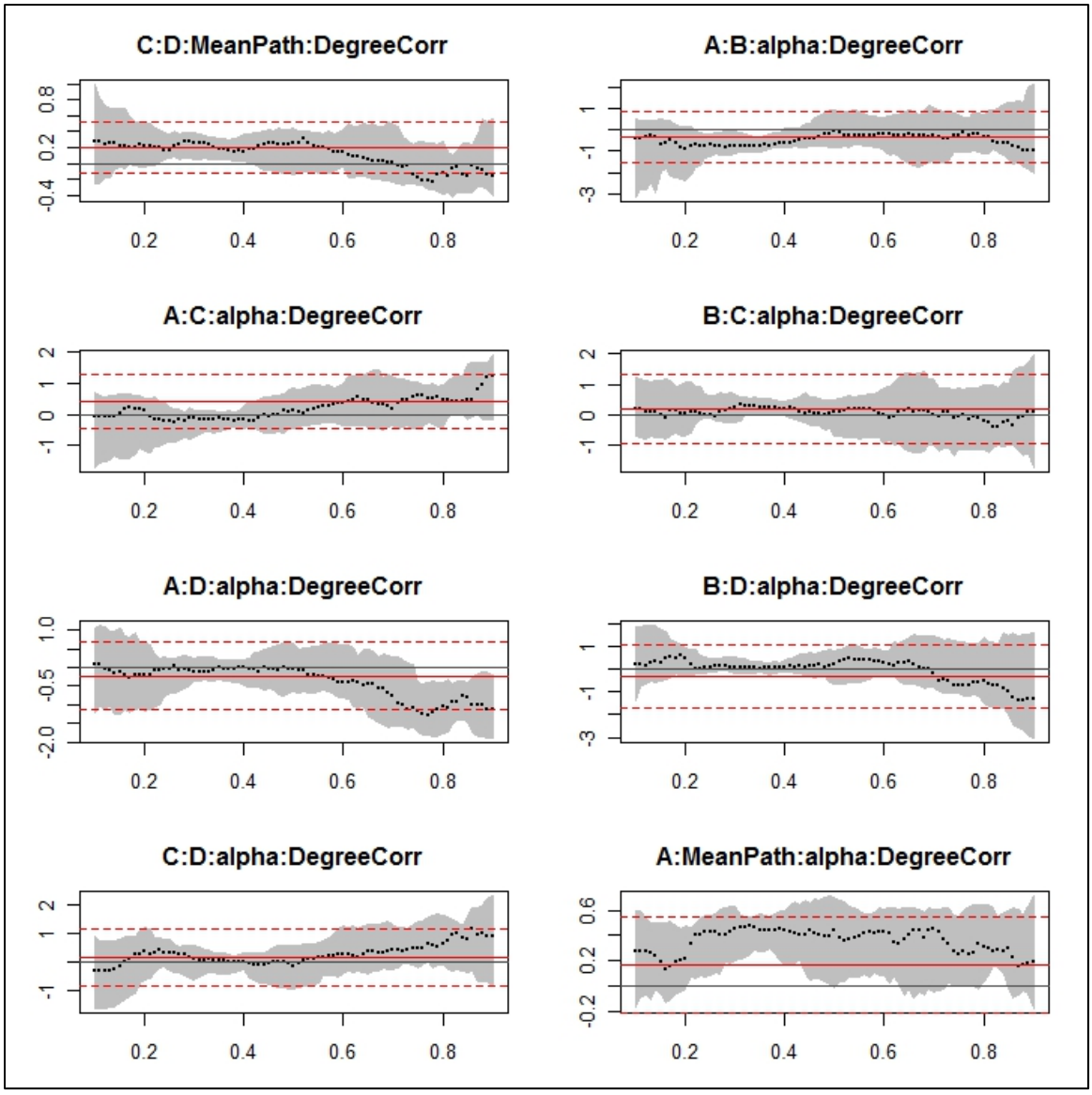

Figure D-13 Four Factor Interactions' Coefficients (cont.)

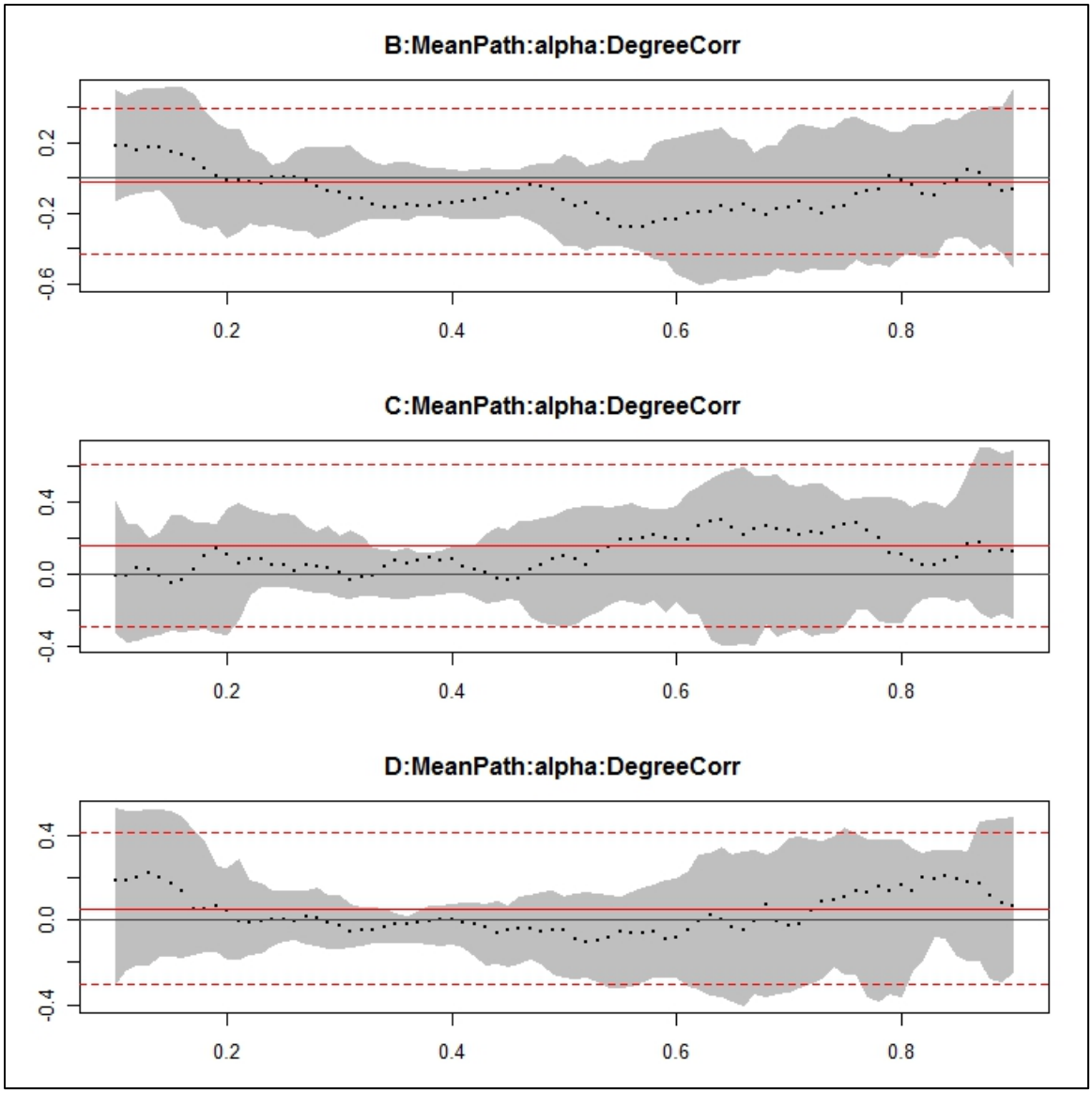

Figure D-14 Four Factor Interactions' Coefficients (cont.)

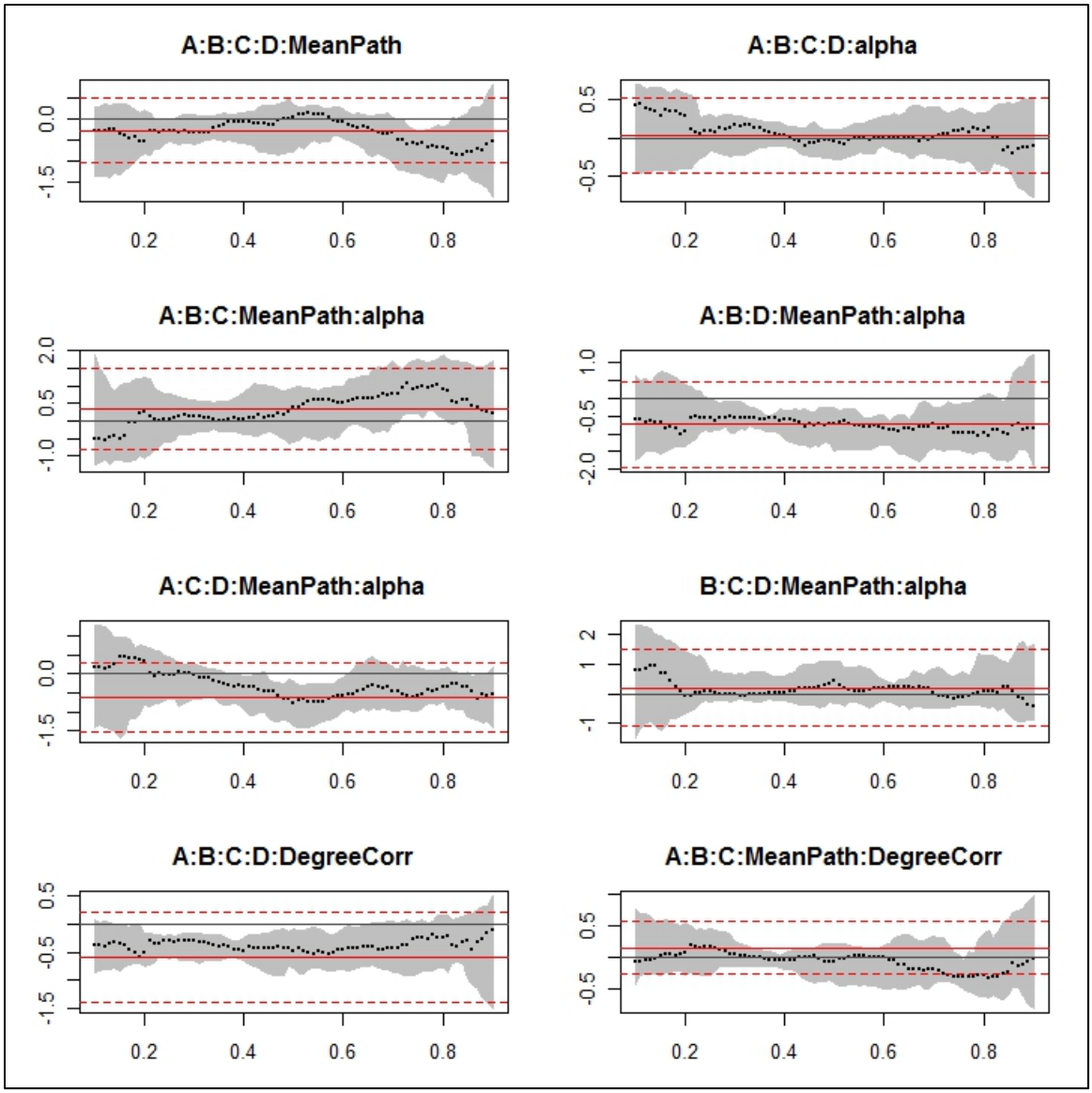

Figure D-15 Five Factor Interactions' Coefficients

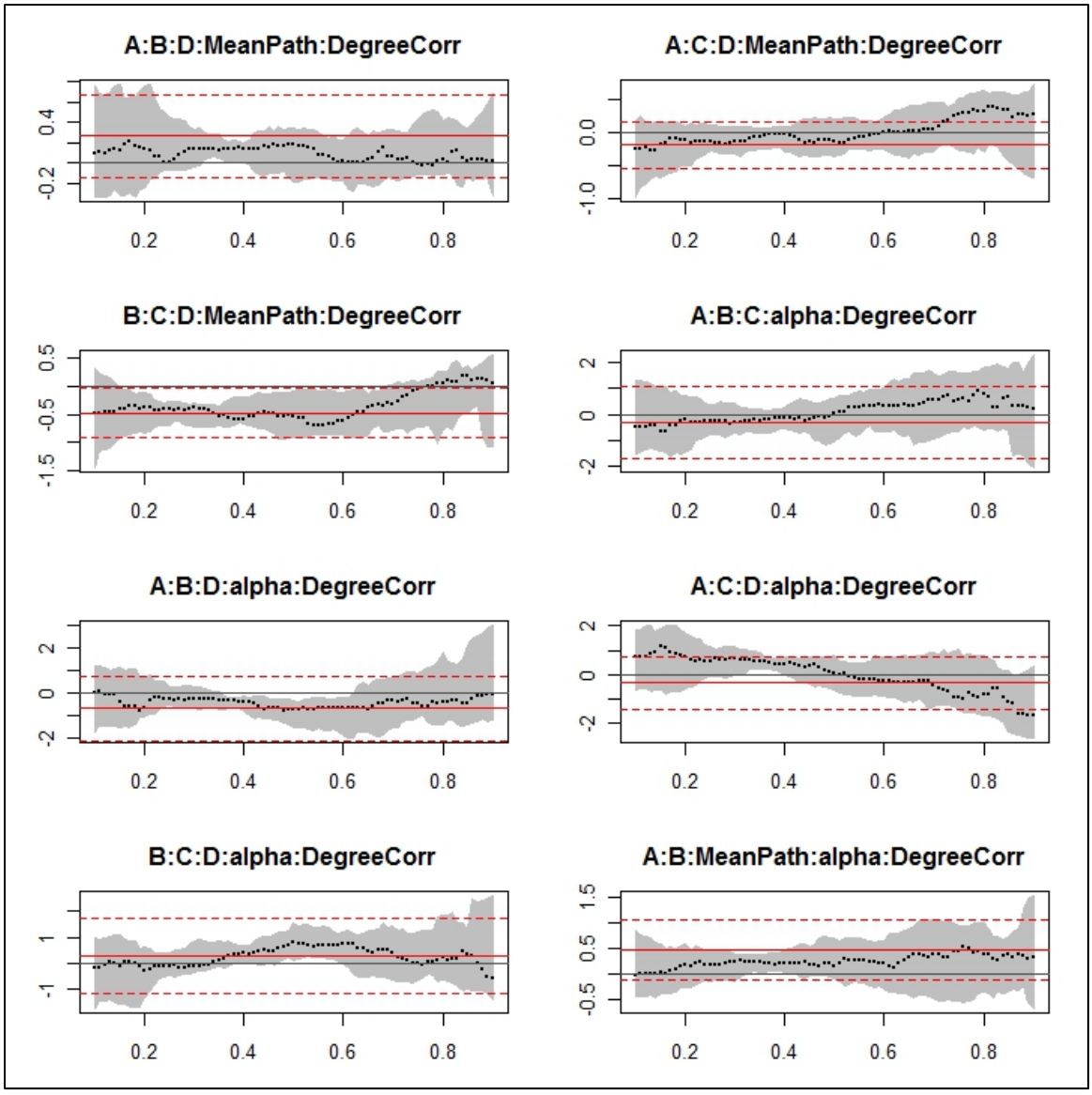

Figure D-16 Five Factor Interactions' Coefficients (cont.)

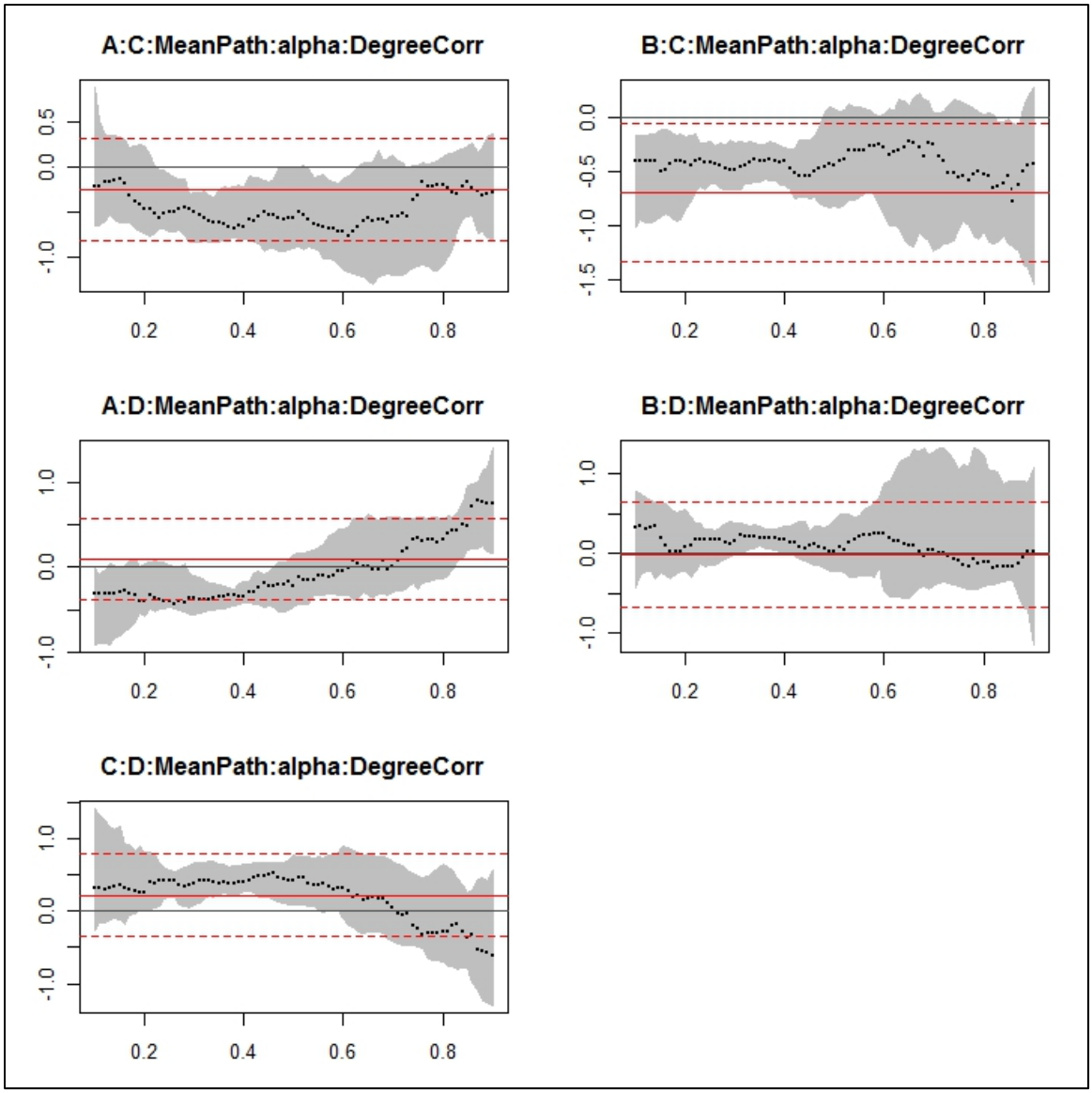

Figure D-17 Five Factor Interactions' Coefficients (cont.)

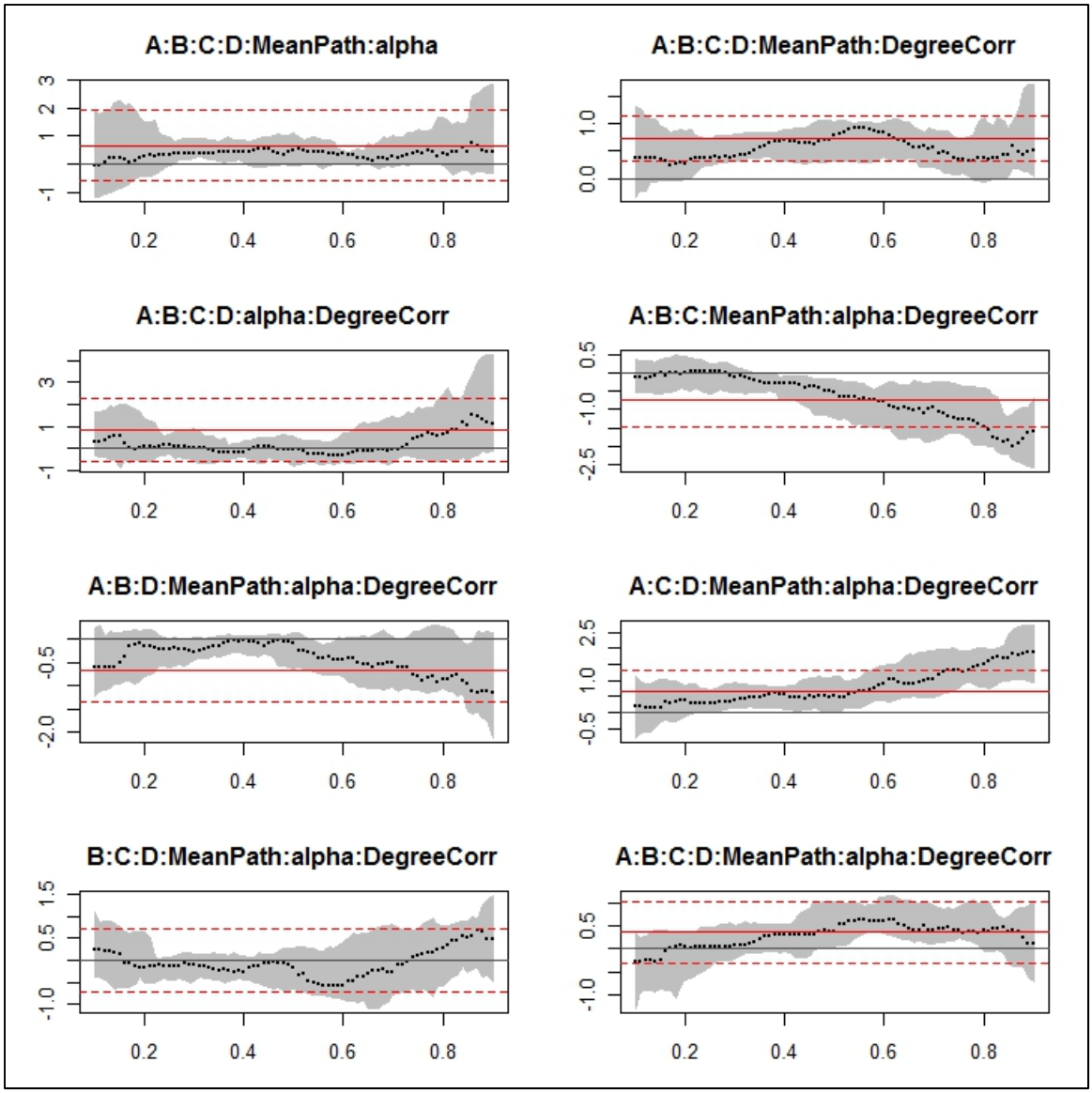

Figure D-18 Six and Seven Factor Interactions' Coefficients

Appendix E R Code

SourceScoring.R inputs the source pairings dissimilarity matrix and the source weightings matrix and conducts the weighted MDS, fuzzy clustering, and AUC computations, outputting the results to a file.

E.1 SourceScoring.R

```
library(smacof)
library(cluster)
library(ROCR)

# DIRECTORIES
dirs <- c("1", "a", "b", "c", "d", "ab", "ac", "ad", "bc", "bd",
"cd", "abc", "abd", "acd", "bcd", "abcd", "SF1", "SF2", "SF3",
"SF4", "SF5", "SF6", "SF7", "SF8", "SF9", "SF10", "SF11", "SF12",
"SF13", "SF14", "SF15", "SF16", "SF17", "centerpt")

for(run in 1:length(dirs)) {

# OUTER LOOP FOR NUMBER OF REPLICATES
for(rep in 1:10) {

# CREATE THE FILE NAME STRINGS FOR DATA INPUT
CohenFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Cohen",rep,
".txt", sep = "")
CohenWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Cohen","Wei
ghts",rep,".txt", sep= "")
GiniFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Gini",rep,"
.txt", sep = "")
GiniWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Gini","Weig
hts",rep,".txt", sep= "")
DispersionFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Dispersion"
,rep,".txt", sep = "")
DispersionWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Dispersion"
,"Weights",rep,".txt", sep= "")
AnderbergFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Anderberg",
rep,".txt", sep = "")
```

```
AnderbergWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Anderberg",
"Weights",rep,".txt", sep= "")
HamannFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Hamann",rep
,".txt", sep = "")
HamannWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Hamann","We
ights",rep,".txt", sep= "")
GKMaxFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","GKMax",rep,
".txt", sep = "")
GKMaxWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","GKMax","Wei
ghts",rep,".txt", sep= "")
PeirceFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Pierce",rep
,".txt", sep = "")
PeirceWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","Pierce","We
ights",rep,".txt", sep= "")
SSIIIFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","SSIII",rep,
".txt", sep = "")
SSIIIWeightsFile <- paste("I:\\My
Documents\\NetBeansProjects\\Research\\",dirs[run],"\\","SSIII","Wei
ghts",rep,".txt", sep= "")

# READ IN THE SOURCES' DISSIMILARITY MATRICES
tryCatch({Error <- FALSE; Cohen<-read.table(CohenFile,header=T)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Cohen <- NA}})
tryCatch({Error <- FALSE; Cohenweights<-read.table(CohenWeightsFile,
header=T)}, error = function(ex) { cat("An error was detected in
run: ", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { Cohenweights <- NA}})
tryCatch({Error <- FALSE; Gini<-read.table(GiniFile,header=T)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Gini <- NA}})
tryCatch({Error <- FALSE; Giniweights<-read.table(GiniWeightsFile,
header=T)}, error = function(ex) { cat("An error was detected in
run: ", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { Giniweights <- NA}})
tryCatch({Error <- FALSE; Dispersion<-
read.table(DispersionFile,header=T)}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { Dispersion <- NA}})
tryCatch({Error <- FALSE; Dispersionweights<-
read.table(DispersionWeightsFile, header=T)}, error = function(ex) {
cat("An error was detected in run: ", dirs[run], ", rep ", rep,":
```

```
"); print(ex); Error <- TRUE}, finally= {if(Error) {
Dispersionweights <- NA}})
tryCatch({Error <- FALSE; Anderberg<-
read.table(AnderbergFile,header=T)}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { Anderberg <- NA}})
tryCatch({Error <- FALSE; Anderbergweights<-
read.table(AnderbergWeightsFile, header=T)}, error = function(ex) {
cat("An error was detected in run: ", dirs[run], ", rep ", rep,":
"); print(ex); Error <- TRUE}, finally= {if(Error) {
Anderbergweights <- NA}})
tryCatch({Error <- FALSE; Hamann<-read.table(HamannFile,header=T)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Hamann <- NA}})
tryCatch({Error <- FALSE; Hamannweights<-
read.table(HamannWeightsFile, header=T)}, error = function(ex) {
cat("An error was detected in run: ", dirs[run], ", rep ", rep,":
"); print(ex); Error <- TRUE}, finally= {if(Error) { Hamannweights
<- NA}})
tryCatch({Error <- FALSE; GKMax<-read.table(GKMaxFile,header=T)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { GKMax <- NA}})
tryCatch({Error <- FALSE; GKMaxweights<-read.table(GKMaxWeightsFile,
header=T)}, error = function(ex) { cat("An error was detected in
run: ", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { GKMaxweights <- NA}})
tryCatch({Error <- FALSE; Peirce<-read.table(PeirceFile,header=T)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Peirce <- NA}})
tryCatch({Error <- FALSE; Peirceweights<-
read.table(PeirceWeightsFile, header=T)}, error = function(ex) {
cat("An error was detected in run: ", dirs[run], ", rep ", rep,":
"); print(ex); Error <- TRUE}, finally= {if(Error) { Peirceweights
<- NA}})

# PERFORM WEIGHTED MDS ON THE SOURCES' DISSIMILARITY MATRICES
tryCatch({Error <- FALSE; CohenMDS <- smacofSym(Cohen, ndim = 2,
Cohenweights, init = NULL, metric = TRUE, ties = "primary", verbose
= FALSE, relax = FALSE, modulus = 1, itmax = 10000, eps = 1e-06)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { CohenMDS <- NA}})
tryCatch({Error <- FALSE; GiniMDS <- smacofSym(Gini, ndim = 2,
Giniweights, init = NULL, metric = TRUE, ties = "primary", verbose =
FALSE, relax = FALSE, modulus = 1, itmax = 10000, eps = 1e-06)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { GiniMDS <- NA}})
```

```
tryCatch({Error <- FALSE; DispersionMDS <- smacofSym(Dispersion,
ndim = 2, Dispersionweights, init = NULL, metric = TRUE, ties =
"primary", verbose = FALSE, relax = FALSE, modulus = 1, itmax =
10000, eps = 1e-06)}, error = function(ex) { cat("An error was
detected in run: ", dirs[run], ", rep ", rep,": "); print(ex); Error
<- TRUE}, finally= {if(Error) { DispersionMDS <- NA}})
tryCatch({Error <- FALSE; AnderbergMDS <- smacofSym(Anderberg, ndim
= 2, Anderbergweights, init = NULL, metric = TRUE, ties = "primary",
verbose = FALSE, relax = FALSE, modulus = 1, itmax = 10000, eps =
1e-06)}, error = function(ex) { cat("An error was detected in run:
", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { AnderbergMDS <- NA}})
tryCatch({Error <- FALSE; HamannMDS <- smacofSym(Hamann, ndim = 2,
Hamannweights, init = NULL, metric = TRUE, ties = "primary", verbose
= FALSE, relax = FALSE, modulus = 1, itmax = 10000, eps = 1e-06)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { HamannMDS <- NA}})
tryCatch({Error <- FALSE; GKMaxMDS <- smacofSym(GKMax, ndim = 2,
GKMaxweights, init = NULL, metric = TRUE, ties = "primary", verbose
= FALSE, relax = FALSE, modulus = 1, itmax = 10000, eps = 1e-06)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { GKMaxMDS <- NA}})
tryCatch({Error <- FALSE; PeirceMDS <- smacofSym(Peirce, ndim = 2,
Peirceweights, init = NULL, metric = TRUE, ties = "primary", verbose
= FALSE, relax = FALSE, modulus = 1, itmax = 10000, eps = 1e-06)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { PeirceMDS <- NA}})

# PERFORM FUZZY CLUSTERING OF THE SOURCES' MDS DISTANCES
tryCatch({Error <- FALSE; Cohenfanny <- fanny(CohenMDS$confdiss,
2)}, error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Cohenfanny <- NA}})
tryCatch({Error <- FALSE; Ginifanny <- fanny(GiniMDS$confdiss, 2)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Ginifanny <- NA}})
tryCatch({Error <- FALSE; Dispersionfanny <-
fanny(DispersionMDS$confdiss, 2)}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { Dispersionfanny <-
NA}})
tryCatch({Error <- FALSE; Anderbergfanny <-
fanny(AnderbergMDS$confdiss, 2)}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { Anderbergfanny <-
NA}})
tryCatch({Error <- FALSE; Hamannfanny <- fanny(HamannMDS$confdiss,
2)}, error = function(ex) { cat("An error was detected in run: ",
```

```
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Hamannfanny <- NA}})
tryCatch({Error <- FALSE; GKMaxfanny <- fanny(GKMaxMDS$confdiss,
2)}, error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { GKMaxfanny <- NA}})
tryCatch({Error <- FALSE; Peircefanny <- fanny(PeirceMDS$confdiss,
2)}, error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { Peircefanny <- NA}})

# DETERMINE THE REAL STATUS OF EACH SOURCE (RELIABLE OR UNRELIABLE)
solution <- substr(names(Cohen),1,1)
if(length(solution)>0) {
for(i in 1:length(solution)) if (solution[i] == "R") solution[i] = 1
else solution[i] = 0
Cohensolution <- as.numeric(solution)
} else {Cohensolution <- NA}
solution <- substr(names(Gini),1,1)
if(length(solution)>0) {
for(i in 1:length(solution)) if (solution[i] == "R") solution[i] = 1
else solution[i] = 0
Ginisolution <- as.numeric(solution)
} else {Ginisolution <- NA}
solution <- substr(names(Dispersion),1,1)
if(length(solution)>0) {
for(i in 1:length(solution)) if (solution[i] == "R") solution[i] = 1
else solution[i] = 0
Dispersionsolution <- as.numeric(solution)
} else {Dispersionsolution <- NA}
solution <- substr(names(Anderberg),1,1)
if(length(solution)>0) {
for(i in 1:length(solution)) if (solution[i] == "R") solution[i] = 1
else solution[i] = 0
Anderbergsolution <- as.numeric(solution)
} else {Anderbergsolution <- NA}
solution <- substr(names(Hamann),1,1)
if(length(solution)>0) {
for(i in 1:length(solution)) if (solution[i] == "R") solution[i] = 1
else solution[i] = 0
Hamannsolution <- as.numeric(solution)
} else {Hamannsolution <- NA}
solution <- substr(names(GKMax),1,1)
if(length(solution)>0) {
for(i in 1:length(solution)) if (solution[i] == "R") solution[i] = 1
else solution[i] = 0
GKMaxsolution <- as.numeric(solution)
} else {GKMaxsolution <- NA}
solution <- substr(names(Peirce),1,1)
if(length(solution)>0) {
for(i in 1:length(solution)) if (solution[i] == "R") solution[i] = 1
else solution[i] = 0
```

```
Peircesolution <- as.numeric(solution)
} else {Peircesolution <- NA}

# COMPUTE THE ROC CURVE FOR EACH DISSIMILARITY STATISTIC
tryCatch({Error <- FALSE; CohenPred <-
prediction(Cohenfanny$membership[,1], Cohensolution)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {if(Error) {
CohenPred <- NA}})
tryCatch({Error <- FALSE; CohenPerf <- performance(CohenPred,
measure = "tpr", x.measure = "fpr")}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { CohenPerf <- NA}})
tryCatch({Error <- FALSE; GiniPred <-
prediction(Ginifanny$membership[,1], Ginisolution)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {if(Error) {
GiniPred <- NA}})
tryCatch({Error <- FALSE; GiniPerf <- performance(GiniPred, measure
= "tpr", x.measure = "fpr")}, error = function(ex) { cat("An error
was detected in run: ", dirs[run], ", rep ", rep,": "); print(ex);
Error <- TRUE}, finally= {if(Error) { GiniPerf <- NA}})
tryCatch({Error <- FALSE; DispersionPred <-
prediction(Dispersionfanny$membership[,1], Dispersionsolution)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { DispersionPred <- NA}})
tryCatch({Error <- FALSE; DispersionPerf <-
performance(DispersionPred, measure = "tpr", x.measure = "fpr")},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { DispersionPerf <- NA}})
tryCatch({Error <- FALSE; AnderbergPred <-
prediction(Anderbergfanny$membership[,1], Anderbergsolution)}, error
= function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {if(Error) {
AnderbergPred <- NA}})
tryCatch({Error <- FALSE; AnderbergPerf <-
performance(AnderbergPred, measure = "tpr", x.measure = "fpr")},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { AnderbergPerf <- NA}})
tryCatch({Error <- FALSE; HamannPred <-
prediction(Hamannfanny$membership[,1], Hamannsolution)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {if(Error) {
HamannPred <- NA}})
tryCatch({Error <- FALSE; HamannPerf <- performance(HamannPred,
measure = "tpr", x.measure = "fpr")}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { HamannPerf <- NA}})
```

```
tryCatch({Error <- FALSE; GKMaxPred <-
prediction(GKMaxfanny$membership[,1], GKMaxsolution)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {if(Error) {
GKMaxPred <- NA}})
tryCatch({Error <- FALSE; GKMaxPerf <- performance(GKMaxPred,
measure = "tpr", x.measure = "fpr")}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { GKMaxPerf <- NA}})
tryCatch({Error <- FALSE; PeircePred <-
prediction(Peircefanny$membership[,1], Peircesolution)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {if(Error) {
PeircePred <- NA}})
tryCatch({Error <- FALSE; PeircePerf <- performance(PeircePred,
measure = "tpr", x.measure = "fpr")}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { PeircePerf <- NA}})

# COMPUTE THE AUC FOR EACH DISSIMILARITY STATISTIC
tryCatch({Error <- FALSE; CohenAUC <- performance(CohenPred,
'auc')}, error = function(ex) { cat("An error was detected in run:
", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { CohenAUC <- NA}})
tryCatch({Error <- FALSE; GiniAUC <- performance(GiniPred, 'auc')},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{if(Error) { GiniAUC <- NA}})
tryCatch({Error <- FALSE; DispersionAUC <-
performance(DispersionPred, 'auc')}, error = function(ex) { cat("An
error was detected in run: ", dirs[run], ", rep ", rep,": ");
print(ex); Error <- TRUE}, finally= {if(Error) { DispersionAUC <-
NA}})
tryCatch({Error <- FALSE; AnderbergAUC <- performance(AnderbergPred,
'auc')}, error = function(ex) { cat("An error was detected in run:
", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { AnderbergAUC <- NA}})
tryCatch({Error <- FALSE; HamannAUC <- performance(HamannPred,
'auc')}, error = function(ex) { cat("An error was detected in run:
", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { HamannAUC <- NA}})
tryCatch({Error <- FALSE; GKMaxAUC <- performance(GKMaxPred,
'auc')}, error = function(ex) { cat("An error was detected in run:
", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { GKMaxAUC <- NA}})
tryCatch({Error <- FALSE; PeirceAUC <- performance(PeircePred,
'auc')}, error = function(ex) { cat("An error was detected in run:
", dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE},
finally= {if(Error) { PeirceAUC <- NA}})

# TEST TO SEE WHICH NUMBERS ARE AVAILABLE FOR OUTPUTTING
output <- dirs[run]
```

```
output <- t(c(output, rep))

# MDS STRESS
tryCatch({Error <- FALSE; num <- CohenMDS$stress.m}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- CohenMDS$stress.m)})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- GiniMDS$stress.m}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- GiniMDS$stress.m)})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- DispersionMDS$stress.m}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- DispersionMDS$stress.m)})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- AnderbergMDS$stress.m}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- AnderbergMDS$stress.m)})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- HamannMDS$stress.m}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- HamannMDS$stress.m)})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- GKMaxMDS$stress.m}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- GKMaxMDS$stress.m)})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- PeirceMDS$stress.m}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- PeirceMDS$stress.m)})
output <- c(output, num)

# DUNN'S COEFFICIENT FOR FUZZY CLUSTERING
tryCatch({Error <- FALSE; num <- Cohenfanny$coeff[2]}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- Cohenfanny$coeff[2])})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- Ginifanny$coeff[2]}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- Ginifanny$coeff[2])})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- Dispersionfanny$coeff[2]}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
```

```
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- Dispersionfanny$coeff[2])})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- Anderbergfanny$coeff[2]}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- Anderbergfanny$coeff[2])})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- Hamannfanny$coeff[2]}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- Hamannfanny$coeff[2])})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- GKMaxfanny$coeff[2]}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- GKMaxfanny$coeff[2])})
output <- c(output, num)
tryCatch({Error <- FALSE; num <- Peircefanny$coeff[2]}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- Peircefanny$coeff[2])})
output <- c(output, num)

# AUC VALUES
tryCatch({Error <- FALSE; num <- unlist(CohenAUC@y.values)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- unlist(CohenAUC@y.values))})
output <- c(output, max(num, 1 - num))
tryCatch({Error <- FALSE; num <- unlist(GiniAUC@y.values)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- unlist(GiniAUC@y.values))})
output <- c(output, max(num, 1 - num))
tryCatch({Error <- FALSE; num <- unlist(DispersionAUC@y.values)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{ifelse(Error, num <- NA, num <- unlist(DispersionAUC@y.values))})
output <- c(output, max(num, 1 - num))
tryCatch({Error <- FALSE; num <- unlist(AnderbergAUC@y.values)},
error = function(ex) { cat("An error was detected in run: ",
dirs[run], ", rep ", rep,": "); print(ex); Error <- TRUE}, finally=
{ifelse(Error, num <- NA, num <- unlist(AnderbergAUC@y.values))})
output <- c(output, max(num, 1 - num))
tryCatch({Error <- FALSE; num <- unlist(HamannAUC@y.values)}, error
= function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- unlist(HamannAUC@y.values))})
output <- c(output, max(num, 1 - num))
tryCatch({Error <- FALSE; num <- unlist(GKMaxAUC@y.values)}, error =
function(ex) { cat("An error was detected in run: ", dirs[run], ",
```

```
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- unlist(GKMaxAUC@y.values))})
output <- c(output, max(num, 1 - num))
tryCatch({Error <- FALSE; num <- unlist(PeirceAUC@y.values)}, error
= function(ex) { cat("An error was detected in run: ", dirs[run], ",
rep ", rep,": "); print(ex); Error <- TRUE}, finally= {ifelse(Error,
num <- NA, num <- unlist(PeirceAUC@y.values))})
output <- c(output, max(num, 1 - num))

# FIRST TIME - CREATE RESULTS, AFTER THAT - ADD TO RESULTS
ifelse(run == 1 & rep == 1, results <- t(output), results <-
rbind(results, t(output)))
}
}

# OUTPUT THE DATA TO A FILE
colnames(results) <- c("Run", "Rep", "CohenStress", "GiniStress",
"DispersionStress", "AnderbergStress", "HamannStress",
"GKMaxStress", "PeirceStress", "CohenDunn", "GiniDunn",
"DispersionDunn", "AnderbergDunn", "HamannDunn", "GKMaxDunn",
"PeirceDunn", "CohenAUC", "GiniAUC", "DispersionAUC",
"AnderbergAUC", "HamannAUC", "GKMaxAUC", "PeirceAUC")
# rownames(results) <- 1:rep
write.table(results, file = "I:\\My
Documents\\NetBeansProjects\\Research\\AUCresults.txt", sep = "\t",
row.names = FALSE, quote = FALSE)
```

Appendix F Java Code for Social Network Source Generation

GenerateReliable.java and GenerateUnreliable.java compose the main java code components used to generate reliable and unreliable social network information sources from input graphs as described in Section 3.2.2. Both files utilize the JUNG java library (O'Madadhain, Fisher, Nelson, White, & Boey, 2010) and additional java files to execute.

F.1 GenerateReliable.java

```
package morris.james.sna.sources;

import morris.james.sna.io.userinterface.ConsoleInput;

import morris.james.sna.io.MODFile;
import morris.james.sna.io.MODFileWriter;
import morris.james.sna.sampling.UAR;

import edu.uci.ics.jung.graph.Graph;
import edu.uci.ics.jung.algorithms.filters.EdgePredicateFilter;
import edu.uci.ics.jung.algorithms.filters.VertexPredicateFilter;
import edu.uci.ics.jung.algorithms.filters.KNeighborhoodFilter;
import edu.uci.ics.jung.graph.util.EdgeType;

import java.util.HashSet;

/**
 * Generates a user specified number of reliable sources drawing
from a given
 * input graph.
 *
 * @author James Morris
 * December 14, 2011
 */
public class GenerateReliable {

    public static String filenameprefix = "R";
    public static String inputfileprefix = "output";
    public static String directory = "";

    public static void main(String[] arguments) {
        int numsources = 0;

        System.out.println();
        System.out.print("Please enter the number of reliable
sources to be generated: ");
```

```
        String convert = ConsoleInput.readline();
        try {
            numsources = Integer.parseInt(convert.substring(0,
convert.length() - 1));
        } catch (NumberFormatException nan) {
            return;
        } catch (NullPointerException npe) {
            return;
        }

        int numReps = 0;
        System.out.println();
        System.out.print("Please enter the number of repetitions:
");
        String RepConvert = ConsoleInput.readline();
        try {
            numReps = Integer.parseInt(RepConvert.substring(0,
RepConvert.length() - 1));
        } catch (NumberFormatException nan) {
            return;
        } catch (NullPointerException npe) {
            return;
        }

        generateReliable(numsources, .10f, numReps);
    }

    /**
     * Generates a user specified number of reliable sources drawing
UAR
     * from a given input graph.
     *
     */
    public static void generateReliable(int numsources, float
samplePct, int numReps) {

        Graph inputgraph;
        boolean directed = false;

        if (!directory.isEmpty()) {
            directory += "/";
        }

        for (int Reps = 1; Reps <= numReps; Reps++) {

            try {
                MODFile input = new MODFile(directory +
inputfileprefix + Reps + ".txt");
                inputgraph = input.load();
                System.out.println("Input Graph Nodes: " +
inputgraph.getVertexCount() + " Edges: " +
inputgraph.getEdgeCount());
```

```
                if
(inputgraph.getDefaultEdgeType().equals(EdgeType.DIRECTED)) {
                    directed = true;
                }

                for (int n = 1; n <= numsources; n++) {

//            System.out.println("Generating Reliable Source #" +
Integer.toString(n));
                    String outputfile = directory + filenameprefix +
Integer.toString(n + (Reps - 1) * numsources) + ".txt";

                    // create source by selecting edges UAR
                    EdgePredicateFilter edges = new
EdgePredicateFilter(new UAR(samplePct));
                    Graph outputgraph = edges.transform(inputgraph);

                    System.out.println("Repetition #" + Reps + ":
Reliable Source #" + n + " contains " + outputgraph.getVertexCount()
+ " nodes and " + outputgraph.getEdgeCount() + " edges.");

                    for (Object v : inputgraph.getVertices()) {
                        if (!outputgraph.containsVertex(v)) {
                            outputgraph.addVertex(v);
                        }
                    }

                    // Check to see if output graph is empty, if so,
redo iteration
                    if (outputgraph.getEdgeCount() > 0) {
                        try {
                            MODFileWriter output = new
MODFileWriter();
                            output.save(outputgraph, outputfile);
                        } catch (RuntimeException ioe) {
                            System.out.println("OUTPUT ERROR: " +
ioe.getMessage());
                        }
                    } else {
                        n--;
                    }
                }
            } catch (RuntimeException ioe) {
                System.out.println("INPUT ERROR: " +
ioe.getMessage());
            }
        }
    }

    /**
     * Generates a user specified number of reliable sources drawing
```

```
     * from a given input graph by selecting nodes UAR and randomly
sampling
     * their specified neighborhood.
     *
     */
    public static void generateReliable(int numsources, int
neighborhood, float seedPct, float samplePct, int numReps) {

        Graph inputgraph;
        boolean directed = false;

        if (!directory.isEmpty()) {
            directory += "/";
        }

        for (int Reps = 1; Reps <= numReps; Reps++) {

            try {
                MODFile input = new MODFile(directory +
inputfileprefix + Reps + ".txt");
                inputgraph = input.load();
                System.out.println("Input Graph: Nodes: " +
inputgraph.getVertexCount() + " Edges: " +
inputgraph.getEdgeCount());

                if
(inputgraph.getDefaultEdgeType().equals(EdgeType.DIRECTED)) {
                    directed = true;
                }

                for (int n = 1; n <= numsources; n++) {

//            System.out.println("Generating Reliable Source #" +
Integer.toString(n));
                    String outputfile = directory + filenameprefix +
Integer.toString(n + (Reps - 1) * numsources) + ".txt";

                    // k-neighborhood for a collection of random
seeds from the input graph
                    VertexPredicateFilter nodes = new
VertexPredicateFilter(new UAR(seedPct));
                    HashSet rootnodes = new
HashSet(nodes.transform(inputgraph).getVertices());
//            System.out.println("Root Nodes: " + rootnodes.size());
                    KNeighborhoodFilter subgraph = new
KNeighborhoodFilter(rootnodes, neighborhood,
KNeighborhoodFilter.EdgeType.IN_OUT);

                    Graph outputgraph =
subgraph.transform(inputgraph);

//                    sample neighborhood by selecting edges UAR
```

```
                    if (samplePct < 1) {
                        EdgePredicateFilter edges = new
EdgePredicateFilter(new UAR(samplePct));
                        outputgraph = edges.transform(outputgraph);
                    }

                    System.out.println("Repetition #" + Reps + ":
Reliable Source #" + n + " contains " + outputgraph.getVertexCount()
+ " nodes and " + outputgraph.getEdgeCount() + " edges.");

                    for (Object v : inputgraph.getVertices()) {
                        if (!outputgraph.containsVertex(v)) {
                            outputgraph.addVertex(v);
                        }
                    }

                    // Check to see if output graph is empty, if so,
redo iteration
                    if (outputgraph.getEdgeCount() > 0) {
                        try {
                            MODFileWriter output = new
MODFileWriter();
                            output.save(outputgraph, outputfile);
                        } catch (RuntimeException ioe) {
                            System.out.println("OUTPUT ERROR: " +
ioe.getMessage());
                        }
                    } else {
                        n--;
                    }
                }

            } catch (RuntimeException ioe) {
                System.out.println("INPUT ERROR: " +
ioe.getMessage());
            }
        }
    }
}
```

F.2 GenerateUnrelaible.java

```
package morris.james.sna.sources;

import morris.james.sna.io.userinterface.ConsoleInput;
import morris.james.sna.io.MODFile;
import morris.james.sna.io.MODFileWriter;
import morris.james.sna.sampling.UAR;

import edu.uci.ics.jung.graph.Graph;
import edu.uci.ics.jung.algorithms.filters.EdgePredicateFilter;
import edu.uci.ics.jung.algorithms.filters.KNeighborhoodFilter;
import edu.uci.ics.jung.algorithms.filters.VertexPredicateFilter;
import edu.uci.ics.jung.graph.util.EdgeType;

import java.util.HashSet;

/**
 * Generates a user specified number of unreliable sources drawing
from a set
 * of given input graphs.
 *
 * @author James Morris
 * December 14, 2011
 */
public class GenerateUnreliable {

    public static String outputfilenameprefix = "U";
    public static String inputfilenameprefix = "output";
    public static String directory = "";

    public static void main(String[] arguments) {
        int numsources = 0;

        System.out.println();
        System.out.print("Please enter the number of unreliable
sources to be generated: ");
        String convert = ConsoleInput.readline();
        try {
            numsources = Integer.parseInt(convert.substring(0,
convert.length() - 1));
        } catch (NumberFormatException nan) {
            return;
        } catch (NullPointerException npe) {
            return;
        }

//        generateUnreliable(numsources, 1, .05f, .10f, 1);

    }

    /**
```

```
     * Generates a user specified number of unreliable sources
drawing from a set
     * of given input graphs.
     * inputstart indicates number of input graphs to start creating
unreliable
     * sources, generally input start is the number of repetitions
of generating
     * reliable sources.
     *
     */
//    public static void generateUnreliable(int numsources, int
neighborhood, float seedPct, float samplePct, int inputstart) {
    public static void generateUnreliable(int numsources, float
samplePct, int inputstart) {

        Graph inputgraph = null;

        if (!directory.isEmpty()) {
            directory += "/";
        }

        for (int n = 1; n <= numsources; n++) {
            try {
                System.out.println("Loading Unreliable Source Input
Graph #" + Integer.toString(n));
                String inputfile = directory + inputfilenameprefix +
Integer.toString(n + inputstart) + ".txt";
                MODFile input = new MODFile(inputfile);
                inputgraph = input.load();
                System.out.println("Input Graph: Nodes: " +
inputgraph.getVertexCount() + " Edges: " +
inputgraph.getEdgeCount());

                boolean directed = false;
                if
(inputgraph.getDefaultEdgeType().equals(EdgeType.DIRECTED)) {
                    directed = true;
                }
            } catch (RuntimeException ioe) {
                System.out.println("INPUT ERROR: " +
ioe.getMessage());
            }

            System.out.println("Generating Unreliable Source #" +
Integer.toString(n));
            String outputfile = directory + outputfilenameprefix +
Integer.toString(n) + ".txt";

            EdgePredicateFilter subgraph = new
EdgePredicateFilter(new UAR(samplePct));

            Graph outputgraph = subgraph.transform(inputgraph);
```

```
//                    System.out.println("Repetition #" + Reps + ":
Unreliable Source #" + n + " contains " +
outputgraph.getVertexCount() + " nodes and " +
outputgraph.getEdgeCount() + " edges.");
            System.out.println("Unreliable Source #" + n + "
contains " + outputgraph.getVertexCount() + " nodes and " +
outputgraph.getEdgeCount() + " edges.");

            for (Object v : inputgraph.getVertices()) {
                if (!outputgraph.containsVertex(v)) {
                    outputgraph.addVertex(v);
                }
            }

            // Check to see if output graph is empty, if so, redo
iteration
            if (outputgraph.getEdgeCount() > 0) {
                try {
                    MODFileWriter output = new MODFileWriter();
                    output.save(outputgraph, outputfile);
                } catch (RuntimeException ioe) {
                    System.out.println("OUTPUT ERROR: " +
ioe.getMessage());
                }
            } else {
                n--;
            }
        }
    }

    public static void generateUnreliable(int numsources, int
neighborhood, float seedPct, float samplePct, int inputstart) {

        Graph inputgraph = null;

        if (!directory.isEmpty()) {
            directory += "/";
        }

        for (int n = 1; n <= numsources; n++) {
            try {
                System.out.println("Loading Unreliable Source Input
Graph #" + Integer.toString(n));
                String inputfile = directory + inputfilenameprefix +
Integer.toString(n + inputstart) + ".txt";
                MODFile input = new MODFile(inputfile);
                inputgraph = input.load();
                System.out.println("Input Graph: Nodes: " +
inputgraph.getVertexCount() + " Edges: " +
inputgraph.getEdgeCount());
```

```
                boolean directed = false;
                if
(inputgraph.getDefaultEdgeType().equals(EdgeType.DIRECTED)) {
                    directed = true;
                }
            } catch (RuntimeException ioe) {
                System.out.println("INPUT ERROR: " +
ioe.getMessage());
            }

            System.out.println("Generating Unreliable Source #" +
Integer.toString(n));
            String outputfile = directory + outputfilenameprefix +
Integer.toString(n) + ".txt";

            // k-neighborhood for a collection of random seeds from
the input graph
            VertexPredicateFilter nodes = new
VertexPredicateFilter(new UAR(seedPct));
            HashSet rootnodes = new
HashSet(nodes.transform(inputgraph).getVertices());
//            System.out.println("Root Nodes: " + rootnodes.size());
            KNeighborhoodFilter subgraph = new
KNeighborhoodFilter(rootnodes, neighborhood,
KNeighborhoodFilter.EdgeType.IN_OUT);

            Graph outputgraph = subgraph.transform(inputgraph);

//                    sample neighborhood by selecting edges UAR
            if (samplePct < 1) {
                EdgePredicateFilter edges = new
EdgePredicateFilter(new UAR(samplePct));
                outputgraph = edges.transform(outputgraph);
            }

//                  System.out.println("Repetition #" + Reps + ":
Unreliable Source #" + n + " contains " +
outputgraph.getVertexCount() + " nodes and " +
outputgraph.getEdgeCount() + " edges.");
            System.out.println("Unreliable Source #" + n + "
contains " + outputgraph.getVertexCount() + " nodes and " +
outputgraph.getEdgeCount() + " edges.");

            for (Object v : inputgraph.getVertices()) {
                if (!outputgraph.containsVertex(v)) {
                    outputgraph.addVertex(v);
                }
            }

            // Check to see if output graph is empty, if so, redo
iteration
            if (outputgraph.getEdgeCount() > 0) {
```

```
                try {
                    MODFileWriter output = new MODFileWriter();
                    output.save(outputgraph, outputfile);
                } catch (RuntimeException ioe) {
                    System.out.println("OUTPUT ERROR: " +
ioe.getMessage());
                }
            } else {
                n--;
            }
        }
    }
}
```

Appendix G Java Code for Source Pairwise Comparisons

SevenMeasureBatchComparison.java is the main java file to conduct batch processing of the information sources' pairwise comparisons to compute the source dissimilarity and weightings matrices for the seven selected binary similarity measures used in Chapter V. SourceCompare.java conducts the pairwise source comparison. Both files utilize the JUNG java library (O'Madadhain, Fisher, Nelson, White, & Boey, 2010) and additional java files to execute.

G.1 SevenMeasureBatchComparison.java

```
package morris.james.sna.experiments.sourcecomparison;

import morris.james.sna.sources.*;
import morris.james.sna.io.userinterface.ConsoleInput;
import morris.james.sna.io.MODFile;
import morris.james.sna.morphisms.GraphMorphisms;
import morris.james.sna.morphisms.MatrixManipulation;

import edu.uci.ics.jung.graph.Graph;

import java.io.BufferedWriter;
import java.io.FileWriter;

import morris.james.statistics.nonparametric.BinarySimilarity;

/**
 * Batch loads source reporting files, trims isolate nodes and
conducts all
 * pairwise source comparisons using seven binary comparison
measures.
 *
 * @author James Morris
 * 3 January 2012
 */
public class SevenMeasureBatchComparison {

    public static String RSprefix = "R";
    public static String USprefix = "U";
    public static String unassessedSourcesFile =
"UnassessedSources.txt";
    public static String directory = "";
// Selected comparison measures output files
```

```
    public static String WeightsPrefix = "Weights";
    static String CohenOutputfilePrefix = "Cohen";
    static String GiniOutputfilePrefix = "Gini";
    static String DispersionOutputfilePrefix = "Dispersion";
    static String AnderbergOutputfilePrefix = "Anderberg";
    static String HamannOutputfilePrefix = "Hamann";
    static String GKMaxOutputfilePrefix = "GKMax";
    static String PierceOutputfilePrefix = "Pierce";

    public static void main(String[] arguments) {

        int numReliable = 0;
        System.out.println();
        System.out.print("Please enter the number of RELIABLE
sources to be compared: ");
        String RSconvert = ConsoleInput.readline();
        try {
            numReliable = Integer.parseInt(RSconvert.substring(0,
RSconvert.length() - 1));
        } catch (NumberFormatException nan) {
            return;
        } catch (NullPointerException npe) {
            return;
        }

        int numUnreliable = 0;
        System.out.println();
        System.out.print("Please enter the number of UNRELIABLE
sources to be compared: ");
        String USconvert = ConsoleInput.readline();
        try {
            numUnreliable = Integer.parseInt(USconvert.substring(0,
USconvert.length() - 1));
        } catch (NumberFormatException nan) {
            return;
        } catch (NullPointerException npe) {
            return;
        }

        int numReps = 0;
        System.out.println();
        System.out.print("Please enter the number of repetitions:
");
        String RepConvert = ConsoleInput.readline();
        try {
            numReps = Integer.parseInt(RepConvert.substring(0,
RepConvert.length() - 1));
        } catch (NumberFormatException nan) {
            return;
        } catch (NullPointerException npe) {
            return;
        }
```

```
    }

    /**
     * Batch loads source reporting files, trims isolate nodes and
conducts all
     * pairwise source comparisons.
     *
     * @author James Morris
     * 3 January 2012
     */
    public static void BatchCompare(int numReps, int numReliable,
int numUnreliable) {

        Graph inputgraph1, inputgraph2;

        if (!directory.isEmpty()) {
            directory += "/";
        }

        for (int Rep = 1; Rep <= numReps; Rep++) {

            // Comparison Results Storage  (STORED AS
DISSIMILARITIES)
// Cohen's Kappa
            float[][] CohenKappa = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
            float[][] CohenWeights = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
// Gini
            float[][] Gini = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
            float[][] GiniWeights = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
// Dispersion
            float[][] Dispersion = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
            float[][] DispersionWeights = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
// AnderbergD
            float[][] Anderberg = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
            float[][] AnderbergWeights = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
// Hamann
            float[][] Hamann = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
            float[][] HamannWeights = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
// Goodman& Kruskal's Maximum Formula
            float[][] GKMax = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
```

```
            float[][] GKMaxWeights = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
// Peirce
            float[][] Pierce = new float[numReliable +
numUnreliable][numReliable + numUnreliable];
            float[][] PierceWeights = new float[numReliable +
numUnreliable][numReliable + numUnreliable];

            String inputfile1 = new String();
            String inputfile2 = new String();

            for (int i = 1; i < (numReliable + numUnreliable); i++)
{

                if (i <= numReliable) {
                    inputfile1 = directory + RSprefix +
Integer.toString(i + (Rep - 1) * numReliable) + ".txt";
                } else {
                    inputfile1 = directory + USprefix +
Integer.toString(i - numReliable + (Rep - 1) * numUnreliable) +
".txt";
                }
                try {
                    MODFile input1 = new MODFile(inputfile1);
                    inputgraph1 = input1.load();

                    // Remove Isolates from graphs
                    Graph graph1 =
GraphMorphisms.removeIsolates(inputgraph1);

                    for (int j = i + 1; j <= (numReliable +
numUnreliable); j++) {

                        if (j <= numReliable) {
                            inputfile2 = directory + RSprefix +
Integer.toString(j + (Rep - 1) * numReliable) + ".txt";
                        } else {
                            inputfile2 = directory + USprefix +
Integer.toString(j - numReliable + (Rep - 1) * numUnreliable) +
".txt";
                        }

                        try {
                            MODFile input2 = new
MODFile(inputfile2);
                            inputgraph2 = input2.load();
//                            System.out.println("\nComparing " +
inputfile1 + " and " + inputfile2);

                            // Remove Isolates from graphs
                            Graph graph2 =
GraphMorphisms.removeIsolates(inputgraph2);
```

```
SourceCompare.computeConfusionMatrix(graph1, graph2);

                            int[][] confusion =
SourceCompare.getConfusionMatrix();

                            SourceCompare.computePctOverlap(graph1,
graph2);

                            CohenKappa[i - 1][j - 1] = 1 -
BinarySimilarity.CohenKappa(confusion);
                            if (Float.isNaN(CohenKappa[i - 1][j -
1])) {
                                CohenKappa[i - 1][j - 1] = 1.5f;
                                CohenWeights[i - 1][j - 1] = 0;
                            } else {
                                CohenWeights[i - 1][j - 1] =
SourceCompare.getPctOverlap();
                            }

                            Gini[i - 1][j - 1] = -
BinarySimilarity.Gini(confusion);
                            if (Float.isNaN(Gini[i - 1][j - 1])) {
                                Gini[i - 1][j - 1] = 4f / 3;
                                GiniWeights[i - 1][j - 1] = 0;
                            } else {
                                GiniWeights[i - 1][j - 1] =
SourceCompare.getPctOverlap();
                            }

                            Dispersion[i - 1][j - 1] = 1f / 3 -
BinarySimilarity.Dispersion(confusion);
                            if (Float.isNaN(Dispersion[i - 1][j -
1])) {
                                Dispersion[i - 1][j - 1] = 2f / 3;
                                DispersionWeights[i - 1][j - 1] = 0;
                            } else {
                                DispersionWeights[i - 1][j - 1] =
SourceCompare.getPctOverlap();
                            }

                            Anderberg[i - 1][j - 1] = .5f -
BinarySimilarity.AnderbergD(confusion);
                            if (Float.isNaN(Anderberg[i - 1][j -
1])) {
                                Anderberg[i - 1][j - 1] = .5f;
                                AnderbergWeights[i - 1][j - 1] = 0;
                            } else {
                                AnderbergWeights[i - 1][j - 1] =
SourceCompare.getPctOverlap();
                            }
```

```
                            Hamann[i - 1][j - 1] = 1 -
BinarySimilarity.Hamann(confusion);
                            if (Float.isNaN(Hamann[i - 1][j - 1])) {
                                Hamann[i - 1][j - 1] = 2;
                                HamannWeights[i - 1][j - 1] = 0;
                            } else {
                                HamannWeights[i - 1][j - 1] =
SourceCompare.getPctOverlap();
                            }

                            GKMax[i - 1][j - 1] = 1 -
BinarySimilarity.GoodmanKruskalMax(confusion);
                            if (Float.isNaN(GKMax[i - 1][j - 1])) {
                                GKMax[i - 1][j - 1] = 2;
                                GKMaxWeights[i - 1][j - 1] = 0;
                            } else {
                                GKMaxWeights[i - 1][j - 1] =
SourceCompare.getPctOverlap();
                            }

                            Pierce[i - 1][j - 1] = 1 -
BinarySimilarity.PeirceIII(confusion);
                            if (Float.isNaN(Pierce[i - 1][j - 1])) {
                                Pierce[i - 1][j - 1] = 1;
                                PierceWeights[i - 1][j - 1] = 0;
                            } else {
                                PierceWeights[i - 1][j - 1] =
SourceCompare.getPctOverlap();
                            }
                        } catch (RuntimeException ioe) {
                            System.out.println("INPUT ERROR loading
Graph 2: " + ioe.getMessage());
                        }
                    }
                } catch (RuntimeException ioe) {
                    System.out.println("INPUT ERROR loading Graph 1:
" + ioe.getMessage());
                }
            }

            MatrixManipulation.constructSymmMatrix(CohenKappa);
            MatrixManipulation.constructSymmMatrix(CohenWeights);
            MatrixManipulation.constructSymmMatrix(Gini);
            MatrixManipulation.constructSymmMatrix(GiniWeights);
            MatrixManipulation.constructSymmMatrix(Dispersion);

MatrixManipulation.constructSymmMatrix(DispersionWeights);
            MatrixManipulation.constructSymmMatrix(Anderberg);

MatrixManipulation.constructSymmMatrix(AnderbergWeights);
            MatrixManipulation.constructSymmMatrix(Hamann);
```

```
            MatrixManipulation.constructSymmMatrix(HamannWeights);
            MatrixManipulation.constructSymmMatrix(GKMax);
            MatrixManipulation.constructSymmMatrix(GKMaxWeights);
            MatrixManipulation.constructSymmMatrix(Pierce);
            MatrixManipulation.constructSymmMatrix(PierceWeights);

            int[] CohenKeptSources =
MatrixManipulation.findNonZeroRows(CohenWeights);
            int[] GiniKeptSources =
MatrixManipulation.findNonZeroRows(GiniWeights);
            int[] DispersionKeptSources =
MatrixManipulation.findNonZeroRows(DispersionWeights);
            int[] AnderbergKeptSources =
MatrixManipulation.findNonZeroRows(AnderbergWeights);
            int[] HamannKeptSources =
MatrixManipulation.findNonZeroRows(HamannWeights);
            int[] GKMaxKeptSources =
MatrixManipulation.findNonZeroRows(GKMaxWeights);
            int[] PierceKeptSources =
MatrixManipulation.findNonZeroRows(PierceWeights);


MatrixManipulation.eliminateZeroRowsSymmetric(CohenWeights);

MatrixManipulation.eliminateZeroRowsSymmetric(GiniWeights);

MatrixManipulation.eliminateZeroRowsSymmetric(DispersionWeights);

MatrixManipulation.eliminateZeroRowsSymmetric(AnderbergWeights);

MatrixManipulation.eliminateZeroRowsSymmetric(HamannWeights);

MatrixManipulation.eliminateZeroRowsSymmetric(GKMaxWeights);

MatrixManipulation.eliminateZeroRowsSymmetric(PierceWeights);


MatrixManipulation.keepIdentifiedRowsSymmetric(CohenKappa,
CohenKeptSources);
            MatrixManipulation.keepIdentifiedRowsSymmetric(Gini,
GiniKeptSources);

MatrixManipulation.keepIdentifiedRowsSymmetric(Dispersion,
DispersionKeptSources);

MatrixManipulation.keepIdentifiedRowsSymmetric(Anderberg,
AnderbergKeptSources);
            MatrixManipulation.keepIdentifiedRowsSymmetric(Hamann,
HamannKeptSources);
            MatrixManipulation.keepIdentifiedRowsSymmetric(GKMax,
GKMaxKeptSources);
```

```
            MatrixManipulation.keepIdentifiedRowsSymmetric(Pierce,
PierceKeptSources);

            // Write results to output files
            try {

saveResults(MatrixManipulation.keepIdentifiedRowsSymmetric(CohenKapp
a, CohenKeptSources), numReliable, numUnreliable, CohenKeptSources,
CohenOutputfilePrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Cohen's Kappa OUTPUT RESULTS
ERROR: " + ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.eliminateZeroRowsSymmetric(CohenWeigh
ts), numReliable, numUnreliable, CohenKeptSources,
CohenOutputfilePrefix + WeightsPrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Cohen's Kappa Weightings OUTPUT
RESULTS ERROR: " + ioe.getMessage());
            }
            try {
                saveUnassessedSources(Rep, "CohenKappa",
numReliable, numUnreliable, CohenKeptSources);
            } catch (RuntimeException ioe) {
                System.out.println("Cohen's Kappa Kept Sources
OUTPUT RESULTS ERROR: " + ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.keepIdentifiedRowsSymmetric(Gini,
GiniKeptSources), numReliable, numUnreliable, GiniKeptSources,
GiniOutputfilePrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Gini OUTPUT RESULTS ERROR: " +
ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.eliminateZeroRowsSymmetric(GiniWeight
s), numReliable, numUnreliable, GiniKeptSources,
GiniOutputfilePrefix + WeightsPrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Gini Weightings OUTPUT RESULTS
ERROR: " + ioe.getMessage());
            }
            try {
                saveUnassessedSources(Rep, "Gini", numReliable,
numUnreliable, GiniKeptSources);
            } catch (RuntimeException ioe) {
```

```
                System.out.println("Gini Kept Sources OUTPUT RESULTS
ERROR: " + ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.keepIdentifiedRowsSymmetric(Dispersio
n, DispersionKeptSources), numReliable, numUnreliable,
DispersionKeptSources, DispersionOutputfilePrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Dispersion OUTPUT RESULTS ERROR:
" + ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.eliminateZeroRowsSymmetric(Dispersion
Weights), numReliable, numUnreliable, DispersionKeptSources,
DispersionOutputfilePrefix + WeightsPrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Dispersion Weightings OUTPUT
RESULTS ERROR: " + ioe.getMessage());
            }
            try {
                saveUnassessedSources(Rep, "Dispersion",
numReliable, numUnreliable, DispersionKeptSources);
            } catch (RuntimeException ioe) {
                System.out.println("Dispersion Kept Sources OUTPUT
RESULTS ERROR: " + ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.keepIdentifiedRowsSymmetric(Anderberg
, AnderbergKeptSources), numReliable, numUnreliable,
AnderbergKeptSources, AnderbergOutputfilePrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Anderberg OUTPUT RESULTS ERROR:
" + ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.eliminateZeroRowsSymmetric(AnderbergW
eights), numReliable, numUnreliable, AnderbergKeptSources,
AnderbergOutputfilePrefix + WeightsPrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Anderberg Weightings OUTPUT
RESULTS ERROR: " + ioe.getMessage());
            }
            try {
                saveUnassessedSources(Rep, "Anderberg", numReliable,
numUnreliable, AnderbergKeptSources);
            } catch (RuntimeException ioe) {
                System.out.println("Anderberg Kept Sources OUTPUT
RESULTS ERROR: " + ioe.getMessage());
```

```
            }
            try {

saveResults(MatrixManipulation.keepIdentifiedRowsSymmetric(Hamann,
HamannKeptSources), numReliable, numUnreliable, HamannKeptSources,
HamannOutputfilePrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Hamann OUTPUT RESULTS ERROR: " +
ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.eliminateZeroRowsSymmetric(HamannWeig
hts), numReliable, numUnreliable, HamannKeptSources,
HamannOutputfilePrefix + WeightsPrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Hamann Weightings OUTPUT RESULTS
ERROR: " + ioe.getMessage());
            }
            try {
                saveUnassessedSources(Rep, "Hamann", numReliable,
numUnreliable, HamannKeptSources);
            } catch (RuntimeException ioe) {
                System.out.println("Hamann Kept Sources OUTPUT
RESULTS ERROR: " + ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.keepIdentifiedRowsSymmetric(GKMax,
GKMaxKeptSources), numReliable, numUnreliable, GKMaxKeptSources,
GKMaxOutputfilePrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("GKMax OUTPUT RESULTS ERROR: " +
ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.eliminateZeroRowsSymmetric(GKMaxWeigh
ts), numReliable, numUnreliable, GKMaxKeptSources,
GKMaxOutputfilePrefix + WeightsPrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("GKMax Weightings OUTPUT RESULTS
ERROR: " + ioe.getMessage());
            }
            try {
                saveUnassessedSources(Rep, "GKMax", numReliable,
numUnreliable, GKMaxKeptSources);
            } catch (RuntimeException ioe) {
                System.out.println("GKMax Kept Sources OUTPUT
RESULTS ERROR: " + ioe.getMessage());
            }
            try {
```

```
saveResults(MatrixManipulation.keepIdentifiedRowsSymmetric(Pierce,
PierceKeptSources), numReliable, numUnreliable, PierceKeptSources,
PierceOutputfilePrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Pierce OUTPUT RESULTS ERROR: " +
ioe.getMessage());
            }
            try {

saveResults(MatrixManipulation.eliminateZeroRowsSymmetric(PierceWeig
hts), numReliable, numUnreliable, PierceKeptSources,
PierceOutputfilePrefix + WeightsPrefix + Rep + ".txt");
            } catch (RuntimeException ioe) {
                System.out.println("Pierce Weightings OUTPUT RESULTS
ERROR: " + ioe.getMessage());
            }
            try {
                saveUnassessedSources(Rep, "Pierce", numReliable,
numUnreliable, PierceKeptSources);
            } catch (RuntimeException ioe) {
                System.out.println("Pierce Kept Sources OUTPUT
RESULTS ERROR: " + ioe.getMessage());
            }
        }
    }

    private static void saveResults(float[][] matrix, int
numReliable, int numUnreliable, int[] keptsources, String filename)
throws RuntimeException {
        try {
            BufferedWriter writer = new BufferedWriter(new
FileWriter(directory + filename));

            for (int index = 1; index <= (numReliable +
numUnreliable); index++) {
                if (keptsources[index-1] > 0) {
                    if (index <= numReliable) {
                        writer.write(RSprefix +
Integer.toString(index) + "\t");
                    } else {
                        writer.write(USprefix +
Integer.toString(index - numReliable) + "\t");
                    }
                }
            }
            writer.write("\r\n");

            for (int i = 0; i < matrix.length; i++) {
                for (int j = 0; j < matrix[i].length; j++) {
                    writer.write(Float.toString(matrix[i][j]) +
"\t");
```

```
                }
                writer.write("\r\n");
            }
            writer.close();
        } catch (Exception e) {
            throw new RuntimeException("Error saving file: " +
directory + filename, e);
        }
    }

        private static void saveUnassessedSources(int Rep, String
measure, int numReliable, int numUnreliable, int[] keptsources)
throws RuntimeException {
        try {
            BufferedWriter writer = new BufferedWriter(new
FileWriter(directory + measure + Rep + unassessedSourcesFile));

            writer.write(directory + "\t");
            writer.write(Rep + "\t");
            writer.write(measure + "\t");
            writer.write(numReliable + "\t");
            writer.write(numUnreliable + "\t");

            for (int index = 1; index <= (numReliable +
numUnreliable); index++) {
                if (keptsources[index-1] == 0) {
                    if (index <= numReliable) {
                        writer.write(RSprefix +
Integer.toString(index) + "\t");
                    } else {
                        writer.write(USprefix +
Integer.toString(index - numReliable) + "\t");
                    }
                }
            }
            writer.write("\r\n");
            writer.close();
        } catch (Exception e) {
            throw new RuntimeException("Error saving file: " +
directory + measure + Rep + unassessedSourcesFile, e);
        }
    }
}
```

G.2 SourceCompare.java

```
package morris.james.sna.sources;

import edu.uci.ics.jung.graph.Graph;
import edu.uci.ics.jung.graph.util.EdgeType;
import edu.uci.ics.jung.graph.util.Pair;

import java.util.Collection;
import org.apache.commons.collections15.CollectionUtils;

/**
 * Compare two sources' social network models.
 *
 * @author James Morris
 * December 15, 2011
 */
public class SourceCompare {

    private static int[][] ConfusionMatrix = {{0, 0}, {0, 0}};
    private static int intersectionsize = 0;
    private static float PCToverlap = 0;

    /*
     * Computes the Confusion Matrix for the two inputted graphs
     */
    public static void computeConfusionMatrix(Graph graph1, Graph
graph2) {

        initializeConfusionMatrix();

        // Undirected Graphs
        if ((graph1.getDefaultEdgeType() == EdgeType.UNDIRECTED) &&
(graph1.getDefaultEdgeType() == EdgeType.UNDIRECTED)) {
            for (Object edge1 : graph1.getEdges()) {
                Pair nodes = graph1.getEndpoints(edge1);
                Object node1 = nodes.getFirst();
                Object node2 = nodes.getSecond();

                if (graph2.containsVertex(node1) &&
graph2.containsVertex(node2)) {
                    if (graph2.isNeighbor(node1, node2)) {
                        ConfusionMatrix[0][0]++;
                    } else {
                        ConfusionMatrix[0][1]++;
                    }
                }
            }

            for (Object edge2 : graph2.getEdges()) {
                Pair nodes = graph2.getEndpoints(edge2);
                Object node1 = nodes.getFirst();
```

```
                Object node2 = nodes.getSecond();
                if (graph1.containsVertex(node1) &&
graph1.containsVertex(node2)) {
                    if (!graph1.isNeighbor(node1, node2)) {
                        ConfusionMatrix[1][0]++;
                    }
                }
            }

            int intersectsize =
CollectionUtils.intersection(graph1.getVertices(),
graph2.getVertices()).size();
            ConfusionMatrix[1][1] = (intersectsize * (intersectsize
- 1) / 2) - ConfusionMatrix[0][0] - ConfusionMatrix[0][1] -
ConfusionMatrix[1][0];
        }

        // Directed Graphs
        if ((graph1.getDefaultEdgeType() == EdgeType.DIRECTED) &&
(graph1.getDefaultEdgeType() == EdgeType.DIRECTED)) {
            for (Object node1 : graph1.getVertices()) {
                for (Object node2 : graph1.getVertices()) {
                    if (!node1.equals(node2) &&
graph2.containsVertex(node1) && graph2.containsVertex(node2)) {
                        if (graph1.isPredecessor(node1, node2)) {
                            if (graph2.isPredecessor(node1, node2))
{
                                ConfusionMatrix[0][0]++;
                            } else {
                                ConfusionMatrix[0][1]++;
                            }
                        } else if (graph2.isPredecessor(node1,
node2)) {
                            ConfusionMatrix[1][0]++;
                        } else {
                            ConfusionMatrix[1][1]++;
                        }
                    }
                }
            }
        }
//        System.out.println("Confusion Matrix");
//        System.out.print(new Integer(confusion[0][0]).toString() +
"\t");
//        System.out.println(new
Integer(confusion[0][1]).toString());
//        System.out.print(new Integer(confusion[1][0]).toString() +
"\t");
//        System.out.println(new
Integer(confusion[1][1]).toString());
    }
```

```
    /*
     * Returns the Confusion Matrix
     */
    public static int[][] getConfusionMatrix() {
        return ConfusionMatrix;
    }

    /*
     * Iniatilizes the Confusion Matrix
     */
    private static void initializeConfusionMatrix() {
        ConfusionMatrix[0][0] = 0;
        ConfusionMatrix[0][1] = 0;
        ConfusionMatrix[1][0] = 0;
        ConfusionMatrix[1][1] = 0;
    }


    /*
     * Gets the Jaccard weightings for the two inputted graphs
     * Confirmed nominations + disputed nominations
     */
    public static float getJaccardWeightings() {
        return (ConfusionMatrix[0][0] + ConfusionMatrix[0][1] +
ConfusionMatrix[1][0]);
    }

    /*
     * Computes Percent Nodes Overlap.
     * Percent Overlap is |N1 intersection N2|/|N1 union N2|
     */
    public static void computePctOverlap(Graph graph1, Graph graph2)
{
        Collection graph1nodes = graph1.getVertices();
        Collection graph2nodes = graph2.getVertices();
        int unionsize = CollectionUtils.union(graph1.getVertices(),
graph2.getVertices()).size();
        int intersectsize =
CollectionUtils.intersection(graph1.getVertices(),
graph2.getVertices()).size();
        PCToverlap = (float) intersectsize / unionsize;
//        System.out.println("Union Size: " + unionsize);
//        System.out.println("Intersection Size: " + intersectsize);
    }

 /*
     * Gets Percent Nodes Overlap.
     * Percent Overlap is |N1 intersection N2|/|N1 union N2|
     */
    public static float getPctOverlap() {
        return PCToverlap;
//        System.out.println("Union Size: " + unionsize);
```

```
//        System.out.println("Intersection Size: " + intersectsize);
    }

    /*
     * Computes Size of Intersection of Nodes.  |N1 intersection N2|
     */
    public static void computeNodeIntersectSize(Graph graph1, Graph
graph2) {
        Collection graph1nodes = graph1.getVertices();
        Collection graph2nodes = graph2.getVertices();
        intersectionsize =
CollectionUtils.intersection(graph1.getVertices(),
graph2.getVertices()).size();
//        System.out.println("Intersection Size: " +
intersectionsize);
    }

    /*
     * Returns Size of Intersection of Nodes.  |N1 intersection N2|
     */
    public static int getNodeIntersectSize() {
        return intersectionsize;
//        System.out.println("Intersection Size: " +
intersectionsize);
    }
}
```

Bibliography

adams, j., & Moody, J. (2007). To tell the truth: Measuring concordance in multiply reported network data. *Social Networks , 29*, 44-58.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics , 23* (2), 193-212.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science , 286*, 509-512.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science , 286*, 509-512.

Barabási, A.-L., & Bonabeau, E. (2003). Scale-Free Networks. *Scientific American , 288* (5), 60-69.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology , 62*, 569-582.

Bernard, H. R., & Killworth, P. D. (1977). Informant Accuracy in Social Network Data II. *Human Communication Research , 4* (1), 3-18.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1979/1980). Informant Accuracy in Social Network Data IV: A Comparison of Clique-Level Structure in Behavioral and Cognitive Network Data. *Social Networks , 2*, 191-218.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1982). Informant Accuracy in Social Network Data V: An Experimental Attempt to Predict Actual Communication from Recall Data. *Social Science Research , 11*, 30-66.

Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The Problem of Informant Accuracy: The Validity of Retrospective Data. *Annual Review of Anthropology , 13*, 495-517.

Bollobás, B., & Riordan, O. M. (2003). Mathematical results on scale-free random graphs. In S. Bornholdt, & H. G. Schuster, *Handbook of Graphs and Networks: From the Genome to the Internet* (pp. 1-34). Darmstadt: Wiley-VCH.

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

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology , 92* (5), 1170-1182.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks , 29*, 555-564.

Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks , 23*, 191-201.

Bondonio, D. (1998). Predictors of accuracy in perceiving informal social networks. *Social Networks , 20*, 301-330.

Bonneau, J., Anderson, J., Anderson, R., & Stajano, F. (2009). Eight Friends are Enough: Social Graph Approximation via Public Listings. *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, (pp. 13-18).

Borg, I., & Groenen, P. J. (2005). *Modern Multidimensional Scaling: Theory and Applications* (2nd edition ed.). New York: Springer.

Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks , 28*, 124-136.

Bottai, M., Cai, B., & McKeown, R. E. (2010). Logistic quantile regression for bounded outcomes. *Statistics in Medicine , 29*, 309-317.

Brands, H. (2010). *Crime, Violence, and the Crisis in Guatemala: A Case Study in the Erosion of the State.* Monograph, U.S. Army War College, Strategic Studies Institute, Carlisle, PA.

Brass, D. J. (1995). A Social Network Perspective on Human Resources Management. *Research in Personnel and Human Resources Management , 13*, 39-79.

Buneman, P., Khanna, S., & Tan, W.-C. (2000). Data Provenance: Some Basic Issues. *Proceedings of the 20th Conference of Foundation of Software Technology and Theoretical Computer Science* (pp. 87-93). New Delhi, India: Springer.

Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: a Bayesian approach. *Social Networks , 25*, 103-140.

Cade, B. S., & Noon, B. R. (2003). A Gentle Introduction to Quantile Regression for Ecologists. *Frontiers in Ecology and the Environment , 1* (8), 412-420.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association , 103* (484), 1438-1456.

Choi, S.-S. (2008). *Correlation Analysis of Binary Similarity and Dissimilarity Measures.* PhD Dissertation, Pace University, Seidenberg School of Computer Science and Information Systems, New York City.

Choi, S.-S., Cha, S.-H., & Tappert, C. C. (2010). A Survey of Binary Similarity and Distance Measures. *Journal of Systemics, Cybernetics and Informatics , 8* (1), 43-48.

Cioppa, T. M., & Lucas, T. W. (2007). Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes. *Technometrics , 49* (1), 45-55.

Clark, C. R. (2005). *Modeling and Analysis of Clandestine Networks.* MS Thesis, AFIT/GOR/ENS/05-04, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB OH.

Clauset, A., Shalizi, C. R., & Newman, M. (2009). Power-Law Distributions in Empirical Data. *SIAM Review , 51* (4), 661-703.

Coles, N. (2001). It's not what you know-it's who you know that counts: Analysing Serious Crime Groups as Social Networks. *British Journal of Criminology , 41*, 580-594.

Conover, W. J. (1971). *Practical Nonparametric Statistics.* New York: John Wiley & Sons, Inc.

Costenbader, E., & Valente, T. (2003). The stability of centrality measures when networks are sampled. *Social Networks , 25*, 283-307.

Costenbader, E., & Valente, T. W. (2004). Corrigendum to "The stability of centrality measures when networks are sampled" [Social Networks 25 (2003) 283-307]. *Social Networks , 26*, 351.

Crawley, M. J. (2007). *The R Book.* West Sussex, England: John Wiley & Sons, Ltd.

Curtis, E. W. (1976). *Factor Regression Analysis: A New Method for Weighting Predictors.* Defense Technical Information Center.

de Leeuw, J., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software , 31* (3), 1-30.

Department of Defense. (2001, April 12). Department of Defense Dictionary of Military and Associated Terms. *Joint Publication 1-02* . Washington, D.C.: Government Printing Office.

Department of Defense. (2004, October 7). Joint and National Intelligence Support to Military Operations. *Joint Publication 2-01* . Washington, D.C.: Government Printing Office.

Department of Defense. (2006, July 13). Military Deception. *Joint Publication 3-13.4* . Washington, D.C.: Government Printing Office.

Department of Defense. (2006, June 29). Operations Security. *Joint Publication 3-13.3* . Washington, D.C.: Government Printing Office.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications.* New York: John Wiley & Sons.

Eidenberger, H. (2011). *Fundamental Media Understanding.* Norderstedt: atpress.

Erdös, P., & Renyi, A. (1959). On random graphs. *Publicationes Mathematicae , 6*, 290-297.

Erdös, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae , 6*, 290-297.

Erickson, B. H. (1981). Secret Societies and Social Structure. *Social Forces , 60* (1), 188-210.

Farley, J. D. (2007). Toward a Mathematical Theory of Counterterrorism. *The Proteus Monograph Series , 1* (2), 1-72.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters , 27*, 861-874.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions* (3rd edition ed.). Hoboken: John Wiley & Sons, Inc.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry , 40* (1), 35-41.

Freeman, L. C. (1978/1979). Centrality in Social Networks Conceptual Clarification. *Social Networks , 1*, 215-239.

Gastwirth, J. L. (1972). The Estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics , 54*, 306-316.

Geffre, J. L. (2007). *A Layered Social and Operational Network Analysis.* MS Thesis, AFIT/GOR/ENS/07-07, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric Statistical Inference, 4th edition.* New York: Marcel Dekker, Inc.

Gladwell, M. (2002). *The Tipping Point: How Little Things Can Make a Big Difference.* New York: Little Brown and Company.

Goodman, L. A. (1961). Snowball Sampling. *The Annals of Mathematical Statistics , 32* (1), 148-170.

Goos, P., & Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach.* West Sussex: John Wiley & Sons Ltd.

Granovetter, M. (1976). Network Sampling: Some First Steps. *The American Journal of Sociology , 81* (6), 1287-1303.

Grant, E. L., & Leavenworth, R. S. (1996). *Statistical Quality Control* (Seventh ed.). Boston: WCB McGraw-Hill.

Gross, J. (2006). nortest: Tests for Normality. *R package version 1.0* .

Guzman, J. D. (2012). *Analysis of Social Network Measures with Respect to the Structural Properties of Networks.* MS Thesis, AFIT/GOR/ENS/12-12, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Hamill, J. T. (2006). *Analysis of Layered Social Networks.* PhD Dissertation, AFIT/DS/ENS/06-03, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *The Bell System Technical Journal , 29* (2), 147-160.

Hao, L., & Naiman, D. Q. (2007). *Quantile Regression.* Thousand Oaks, CA: Sage Publications, Inc.

Herbranson, T. J. (2007). *Isolating Key Players in Clandestine Networks.* MS Thesis, AFIT/GOR/ENS/07-11, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Holland, P. W., & Leinhardt, S. (1973). The Structural Implications of Measurement Error in Sociometry. *Journal of Mathematical Sociology , 3*, 85-111.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression, Second Edition.* New York: John Wiley & Sons, Inc.

Hubbell, C. H. (1965). An Input-Output Approach to Clique Identification. *Sociometry , 28* (4), 377-399.

Illenberger, J., Flotterod, G., & Nagel, K. (2008, August 1). *An Approach to Correct Biases Induced by Snowball Sampling.* Retrieved January 26, 2010, from Technical University of Berlin: https://svn.vsp.tu-berlin.de/repos/public-svn/publications/vspwp/2008/08-16/snowball.pdf

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis , 6*, 429-449.

Katz, L. (1953). A New Status Index Derived from Sociometric Analysis. *Psychometrika , 18* (1), 39-43.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley & Sons, Inc.

Keegan, B., Ahmed, M. A., Williams, D., Srivastava, J., & Contractor, N. (2010). Dark Gold: Statistical Properties of Clandestine Networks in Massively Multiplayer Online Games. *IEEE International Conference on Social Computing*, (pp. 201-208). Minneapolis, MN.

Kennedy, K. T. (2009). *Synthesis, Interdiction and Protection of Layered Networks.* PhD Dissertation, AFIT/DS/ENS/09-01, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Killworth, P. D., & Bernard, H. R. (1976). Informant Accuracy in Social Network Data. *Human Organization , 35* (3), 269-286.

Killworth, P. D., & Bernard, H. R. (1979/1980). Informant Accuracy in Social Network Data III: A Comparison of Triadic Structure in Behavioral and Cognitive Data. *Social Networks , 2*, 19-46.

Kim, P.-J., & Jeong, H. (2007). Reliability of rank order in sampled networks. *The European Physical Journal B , 55*, 109-114.

Kmenta, J. (1971). *Elements of Econometrics.* New York: The Macmillan Company.

Koenker, R. (2005). *Quantile Regression.* Cambridge: Cambridge University Press.

Koenker, R. (2011). quantreg: Quantile Regression. *R package version 4.76* .

Koenker, R., & Hallock, K. F. (2001). Quantile Regression. *Journal of Economic Perspectives , 15* (4), 143-156.

Koenker, R., & Machado, J. A. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association , 94* (448), 1296-1310.

Kossinets, G. (2006). Effects of Missing Data in Social Networks. *Social Networks , 28*, 247-268.

Kossinets, G. (2008, Feb. 2). Effects of Missing Data in Social Networks. *E-Print, arXiv:cond-mat/0306335v2* , 1-31.

Krackhardt, D. (1987). Cognitive Social Structures. *Social Networks , 9*, 109-134.

Laumann, E. O., Marsden, P. V., & Prensky, D. (1983). The Boundary Specification Problem in Network Analysis. In R. a. Burt, *Applied Network Analysis: A Methodological Introduction* (pp. 18-34). London: Sage Publications.

Leinart, J. A. (2008). *Characterizing and Detecting Unrevealed Elements of Network Systems.* PhD Dissertation, AFIT/DS/ENS/08-01W, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Leskovec, J., & Faloutsos, C. (2006). Sampling from Large Graphs. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Philadelphia, PA: ACM.

Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd edition.* Hoboken: John Wiley & Sons, Inc.

Little, R. J., & Rubin, D. B. (1989/1990). The Analysis of Social Science Data with Missing Values. *Sociological Methods and Research , 18* (2 & 3), 292-326.

Maddaloni, J.-P. N. (2009). *An Analysis of the FARC in Colombia: Breaking the Frame of FM 3-24.* Master's Monograph, United States Army Command and General Staff College, School of Advanced Military Studies, Fort Leavenworth, KS.

Maechler, M., Rousseeuw, P., & Struy, A. (2005). Cluster Analysis Basics and Extensions. *unpublished* .

Marsden, P. V. (1990). Network Data and Measurement. *Annual Review of Sociology* , 435-463.

Milward, H. B., & Raab, J. (2006). Dark Networks as Organizational Problems: Elements of a Theory. *International Public Management Journal , 9* (3), 333-360.

Mitra, A. (1993). *Fundamentals of Quality Control and Improvement.* New York: Macmillan Publishing Company.

Mizruchi, M. S., Mariolis, P., Schwartz, M., & Mintz, B. (1986). Techniques for Disaggregating Centrality Scores in Social Networks. In N. B. Tuma (Ed.), *Sociological Methodology* (pp. 26-48). San Francisco, CA: Jossey-Bass.

Montgomery, D. C. (2005). *Design and Analysis of Experiments* (6th ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to Linear Regression Analysis* (4th ed.). Hoboken, New Jersey: Wiley-Interscience.

Morris, J. F., O'Neal, J. W., & Deckro, R. F. (Forthcoming). A Random Graph Generation Algorithm for the Analysis of Social Networks. *Journal of Defense Modeling and Simulation* .

Morselli, C., & Petit, K. (2007). Law-Enforcement Disruption of a Drug Importation Network. *Global Crime , 8* (2), 109-130.

Murphy, A. H. (1996). The Finley Affair: A Signal Event in the History of Forecast Verification. *Weather and Forecasting* , 3-20.

Natarajan, M. (2006). Understanding the Structure of a Large Heroin Distribution Network: A Quantitative Analysis of Qualititative Data. *Journal of Quantitative Criminology , 22* (2), 171-192.

Neal, J. W. (2008). "Kracking" the Missing Data Problem: Applying Krackhardt's Cognitive Social Structures to School-Based Social Networks. *Sociology of Education , 81*, 140-162.

Neter, J., Wasserman, W., & Kutner, M. (1985). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs* (2nd ed.). Homewood, IL: Richard D. Irwin, Inc.

Newman, M. E. (2002). Assortative Mixing in Networks. *Physical Review Letters , 89* (20), 208701-1 - 208701-4.

Newman, M. E., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E , 68*, 036122-1 : 036122-8.

Newman, M. E., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E , 68*, 036122.

Newman, M. E., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E , 64* (2), 026118.

O'Madadhain, J., Fisher, D., Nelson, T., White, S., & Boey, Y.-B. (2010, January 24). Java Universial Network/Graph Framework (JUNG).

Oracle Corporation. (2010, June). Netbeans IDE 6.9.1.

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria.

Raab, J., & Milward, H. B. (2003). Dark Networks as Problems. *Journal of Public Administration Research and Theory , 13* (4), 413-439.

Reed, B. J. (2006). *Formalizing the Informal: A Network Analysis of an Insurgency.* PhD Dissertation, University of Maryland, Department of Sociology, College Park, MD.

Renfro, I. R. (2001). *Modeling and Analysis of Social Networks.* PhD Dissertation, AFIT/GOR/ENS/07-11, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Ressler, S. (2006). Social Network Analysis as an Approach to Combat Terrorism: Past, Present, and Future Research. *Homeland Security Affairs , II* (2), 1-10.

Romney, A. K., & Faust, K. (1982). Predicting the Structure of a Communications Network From Recalled Data. *Social Networks , 4*, 285-304.

Romney, A. K., & Weller, S. C. (1984). Predicting Informant Accuracy From Patterns of Recall Among Individuals. *Social Networks , 6*, 59-77.

Sabidussi, G. (1966). The Centrality Index of a Graph. *Psychometrika , 31* (4), 581-603.

Sanchez, S. M. (2005). *Software Downloads.* Retrieved November 7, 2011, from SEED Lab: http://diana.cs.nps.navy.mil/SeedLab/

Sarkadi, K., & Vincze, I. (1974). *Mathematical Methods of Statistical Quality Control.* New York: Academic Press.

Schum, D. A., & Kelly, C. W. (1973). A Problem in Cascaded Inference: Determining the Inferential Impact of Confirming and Conflicting Reports from Several Unreliable Sources. *Organizational Behavior and Human Performance , 10*, 404-423.

Seder, J. S. (2007). *Examining Clandestine Social Networks for the Presence of Non-random Structure.* MS Thesis, AFIT/GOR/ENS/07-24, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika , 52* (3/4), 591-611.

Siegel, J. S. (2002). *Applied Demography: Applications to Business, Government, Law, and Public Policy.* San Diego: Academic Press.

Sing, T., Sander, O., Beerenwinkel, N. B., & Lengauer, T. (2009). ROCR: Visualizing the performance of scoring classifiers. *R package version 1.0-4* .

Smithson, M., & Verkuilen, J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. *Psychological Methods , 11* (1), 54-71.

Soffer, S. N., & Vázquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E , 71*, 057101.

Sparrow, M. K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks , 13*, 251-274.

Sterling, S. E. (2004). *Aggregation Techniques to Characterize Social Networks.* MS Thesis, AFIT/GOR/ENS/04-12, Air Force Institute of Technology, Department of Operational Sciences, Wright-Patterson AFB, OH.

Stork, D., & Richards, W. D. (1992). Nonrespondents in Communication Network Studies: Problems and Possibilities. *Group & Organization Management , 17* (2), 193-209.

Tsvetovat, M., & Carley, K. M. (2007). On Effectiveness of Wiretap Programs in Mapping Social Networks. *Computational & Mathematical Organization Theory* , 63-87.

Valente, T. W., & Foreman, R. K. (1998). Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network. *Social Networks , 20*, 89-105.

van der Hulst, R. C. (2009). Introduction to Social Network Analysis (SNA) as an investigative tool. *Trends in Organized Crime , 12* (2), 101-121.

Vardeman, S. B., & Jobe, M. J. (1999). *Statistical Quality Assurance Methods for Engineers.* New York: John Wiley & Sons, Inc.

Warrens, M. J. (2008). *Similarity Coefficients for Binary Data.* PhD Dissertation, Leiden University, Leiden, Netherlands.

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

Watts, D. J. (1999). *Small Worlds: The Dynamics of Networks between Order and Randomness.* Princeton: Princeton University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature , 393*, 440-442.

Wei, T. (2011). corrplot: Visualization of a correlation matrix. *R package version 0.60* .

Xu, J., & Chen, H. (2008). The topology of dark networks. *Communications of the ACM , 51* (10), 58-65.

Zilli, A., Grippa, F., Gloor, P., & Laubacher, R. (2006). One in Four is Enough - Strategies for Selecting Ego Mailboxes for a Group Network View. *Proceedings of the European Conference on Complex Systems 2006*, (pp. 1-14). Oxford, UK.

Vita

Mr. James F. Morris graduated from New Albany High School in New Albany, Indiana. He entered undergraduate studies at Purdue University in West Lafayette, Indiana where he graduated with a Bachelor of Science degree in Industrial Engineering in December 1999. He continued on to graduate school at Purdue University in West Lafayette, Indiana completing a Master's of Science degree in Industrial Engineering in December 2001. In August of 2001, he became a federal civil servant at the National Air & Space Intelligence Center (NASIC) serving as an intelligence analyst. He conducted his doctoral studies at the Air Force Institute of Technology (AFIT) under the auspices of: a one year long-term full-time study sponsored by NASIC, part-time research support scholarship from the Dayton Area Graduate Studies Institute, the Bonder Scholarship for Applied Operations Research in Military Applications, and a Department of Defense SMART scholarship sponsored by NASIC enabling full-time study at AFIT to complete his doctoral research and dissertation. Upon graduation, Mr. Morris will return to NASIC to fulfill his service commitment as a condition of his SMART scholarship.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
06-14-2012	**Doctoral Dissertation**	Oct 2010 – Jun 2012

4. TITLE AND SUBTITLE

A Quantitative Methodology for Vetting "Dark Network" Intelligence Sources for Social Network Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

James F. Morris

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/DS/ENS/12-05

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NASIC/GTRB
Attn: Mr. Robert K. Mussen
4150 Watson Way DSN: 787-2895
WPAFB OH 45433 e-mail:
Robert.Mussen@wpafb.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Social network analysis (SNA) is used by the DoD to describe and analyze social networks, leading to recommendations for operational decisions. However, social network models are constructed from various information sources of indeterminate reliability. Inclusion of unreliable information can lead to incorrect models resulting in flawed analysis and decisions. This research develops a methodology to assist the analyst by quantitatively identifying and categorizing information sources so that determinations on including or excluding provided data can be made.

This research pursued three main thrusts. It consolidated binary similarity measures to determine social network information sources' concordance and developed a methodology to select suitable measures dependent upon application considerations. A methodology was developed to assess the validity of individual sources of social network data. This methodology utilized source pairwise comparisons to measure information sources' concordance and a weighting schema to account for sources' unique perspectives of the underlying social network. Finally, the developed methodology was tested over a variety of generated networks with varying parameters in a design of experiments paradigm (DOE). Various factors relevant to conditions faced by SNA analysts potentially employing this methodology were examined. The DOE was comprised of a 2^4 full factorial design augmented with a nearly orthogonal Latin hypercube. A linear model was constructed using quantile regression to mitigate the non-normality of the error terms.

15. SUBJECT TERMS

Social Network Analysis (SNA), dark networks, information source assessment, social network information sources, binary similarity measures, quantile regression, Design of Experiments (DOE), random social network generation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Richard F. Deckro (ENS)
U	U	U	UU	416	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-6565, ext 4325; e-mail: Richard.Deckro@afit.edu