3-26-2015

A Risk Based Approach to Node Insertion within Social Networks

Chancellor A. J. Johnstone

A Risk Based Approach to Node Insertion Within Social Networks

THESIS

MARCH 2015

Chancellor A. J. Johnstone, Second Lieutenant, USAF

AFIT-ENS-MS-15-M-136

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, United States Department of Defense or the United States Government. This is an academic work and should not be used to imply or infer actual mission capability or limitations.

AFIT-ENS-MS-15-M-136

A RISK BASED APPROACH TO NODE INSERTION WITHIN SOCIAL NETWORKS

THESIS

Presented to the Faculty
Department of Operational Sciences
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the Requirements for the
Degree of Master of Science (Operations Research)

Chancellor A. J. Johnstone, BS
Second Lieutenant, USAF

March 2015

AFIT-ENS-MS-15-M-136

A RISK BASED APPROACH TO NODE INSERTION WITHIN SOCIAL NETWORKS

THESIS

Chancellor A. J. Johnstone, BS
Second Lieutenant, USAF

Committee Membership:

Maj Jennifer L. Geffre, Ph.D.

Dr. James F. Morris, Ph.D.

AFIT-ENS-MS-15-M-136

Abstract

Social Network Analysis (SNA) is a primary tool for counter-terrorism operations, ranging from resiliency and influence to interdiction on threats stemming from illicit overt and clandestine network operations. In an ideal world, SNA would provide a perfect course of action to eliminate dangerous situations that terrorist organizations bring. Unfortunately, the covert nature of terrorist networks makes the effects of these techniques unknown and possibly detrimental. To avoid potentially harmful changes to enemy networks, tactical involvement must evolve, beginning with the intelligent use of network infiltration through the application of the node insertion problem. The framework for the node insertion problem includes a risk-benefit model to assess the utility of various node insertion scenarios. This model incorporates local, intermediate and global SNA measures, such as Laplacian centrality and assortative mixing, to account for the benefit and risk. Application of the model to the Zachary Karate Club produces a set of recommended insertion scenarios. A designed experiment validates the robustness of the methodology against network structure and characteristics. Ultimately, the research provides an SNA method to identify optimal and near-optimal node insertion strategies and extend past node utility models into a general form with the inclusion of benefit, risk, and bias functions.

For the fellows of Miami Valley Philosophical and Lifting Society

Acknowledgements

I would like to especially thank Maj Jennifer L. Geffre. Your advising, not only throughout this endeavor but also AFIT as a whole, was the lynch-pin for the document's completion. The interest you showed this research and in my officer development was motivating throughout this process. Additional appreciation goes to Dr. James L. Morris for coming on to aid in both the formulation of this ill-defined problem and providing direction when none seemed to be had.

Chancellor A. J. Johnstone

Table of Contents

Page

Page

List of Figures

List of Tables

Table Page

A RISK BASED APPROACH TO NODE INSERTION WITHIN SOCIAL NETWORKS

I. Introduction

1.1 Background

While social network analysis (SNA) is not a new discipline, first gaining notoriety following three independent research movements occurring in the 1930s, its importance has exploded in recent years due to the rise of terrorism [7]. A terrorist organization is one that takes part in actions including, but not limited to, "kidnapping, assassination, hijacking, nuclear, biological, or chemical agents, the use of firearms or other dangerous devices" [8] or endorses such actions.

The United States' primary goal is to achieve and protect "national interests through diplomacy, economic development, cooperation and engagement, and through the power of [its] ideas", but in a world where possible nuclear proliferation and terrorism go hand-in-hand "the willingness and ability to resort to force in defense of our national interests and the common good" becomes a necessity [9]. The mitigation and elimination of terrorism is crucial, and as the *National Strategy for Counterterrorism* states, "the American people and interests will not be secure from attacks until this threat is eliminated–its primary individuals and groups rendered powerless, and its message relegated to irrelevance" [10]. The Department of Homeland Security echoes these focuses with their strategic priorities, shown in the following quote [11]:

> “The evolution of the terrorist threat demands a well-informed, highly agile, and well-networked group of partners and stakeholders to anticipate, detect, target, and disrupt threats that challenge national security, economic prosperity, and public safety. To improve overall unity of effort, we will work with our partners to identify, investigate, and interdict legitimate threats as early as possible; expand risk-based security; focus on countering violent extremism and helping to prevent complex mass casualty attacks; reduce vulnerabilities by denying resources and targets; and uncover patterns and faint signals through enhanced data integration and analysis.”

In an ideal world, SNA would provide perfect insight as to the correct course of action (COA) to eliminate the dangerous situations that terrorism and its associated organizations bring. Unfortunately, the covert nature, and the unforeseen secondary and tertiary effects of dealing with theses networks, makes determining the best COA difficult.

Network research in the private sector emphasizes primarily resiliency, whereas defense research additionally emphasizes network interdiction and influence. Joint Publication 3-03 specifically defines this interdiction as “actions to divert, disrupt, degrade, or destroy” a network [12]. Benjamin in 2008 [13] identifies these actions as “tactical counterterrorism” and compares them specifically to “the catching and killing of terrorists and disruption of their operations”. Benjamin also states the need to continue these methods, while calling for a “significant departure from the current policy” in strategic counterterrorism (CT) [13].

While the need for the continuation of tactical counterterrorism is apparent, methods for this tactical involvement must evolve. Currently, the selection and removal of key network actors and relationships is the primary way by which the United States attempts to combat terrorist threats. Kathleen Carley in 2003 supported this by asserting that “node changes can be more devastating on system performance than relationship changes" [14].

While Carley emphasizes node removal and isolation as the most effective node changes due to their practicality, the National Strategy for Counterterrorism in 2011 [10] identifies the need to diminish the strength of "local and regional affiliates...monitor communications...drive fissures between these groups and their bases of support", showing the need for multifaceted CT methods [14]. While the current tactics seem to be effective in removing individual threats, the integration of other methods could provide additional insight to the counterterrorism decision landscape.

1.2 Undercover Operations and Network Infiltration

The evolution of tactical counterterrorism begins with the intelligent use of network infiltration, or the covert insertion of assets into a network, otherwise known as node insertion. The Federal Bureau of Intelligence (FBI) defines an undercover operation as "an investigation involving a series of related undercover activities over a period of time by an undercover employee" [15]. These operations aid in the "detection, prevention and prosecution of white collar crimes, public corruption, terrorism, organized crime, offenses involving controlled substances, and other priority areas of investigation" [15]. The systematic use of this could provide opportunities to increase the overall effectiveness of US CT efforts.

Fijnaut and Marx [16] trace formal undercover operations to 16th century Europe, used in order to secure "political, military, and economic interests" later highlighting their role in the formation of policing agencies in the United States. A more infamous instance of undercover operations occurred with Operation Black Biscuit. Staged in response to the growth of the Hell's Angels Motorcycle Club (HAMC), Operation

Black Biscuit resulted in the indictment of sixteen Hell's Angels' members, raids all thorughout the western United States, and the seizure of over 1,600 pieces of evidence [17]. However, this operation is seen as a classic example of "the misuse of informers" both because of illicit actions some of these informers took part in and the handling of these agents following the culmination of the operation [18].

In terms of network infiltration, information collection, and overall disruption, the operations was a success, but Droban in 2007 [19] adds a clarification by saying "the operatives may have crippled the Hell's Angels enterprise, but like a true crime family, the club was self-perpetuating and there would always be replacements." This statement is true but it is also important to note that the network never regained the level of power and influence it had prior to the operation. The clarification again highlights the dynamic nature of clandestine organizations.

The purpose behind node insertion is two-fold: information collection and future network disruption. For the collection purpose, it allows for the possible gathering of intelligence directly from the network, instead of through reconnaissance or informants, which could prove unreliable. By applying node insertion, information regarding the individuals within the network, their involvement with other organizations, possible past and future network activities or operations, and even the means by which the network operates becomes obtainable. Even the determination of group ideology and motivations becomes possible prior to the network's execution of some major event.

For future network disruption, an asset inserted into a network prior to a proposed key actor's removal, or any other node changes, might allow for easier degradation of relationships. While this second purpose requires more intelligence on the network to determine which relationships should be degraded, assumptions could be made in order to apply the node insertion methodology. This could allow for the elimination of multiple actors, or lines of communication, at a single instant, possibly creating a disconnect large enough to completely dissolve the network, or disrupt it to the point where it no longer poses a legitimate threat.

1.3 Problem Statement

According to FBI doctrine, “any official considering approval or authorization of a proposed undercover application shall weigh the risks and benefits of the operation” [15]. While protocol exists for undercover operations, there does not seem to be objective or quantitative methodologies to determine these risks and benefits, only subjective definitions.

This research aims to provide a structured methodology for covert network infiltration through the application of node insertion. This includes the formulation of quantitative risk and benefit measures from the perspective of the inserted node. The use of these two independent measure sets allows for the creation of a trade off space, and ultimately a pareto frontier for possible recommendations. From the analysis we hope to gain recommendations for the most effective node insertion scenarios, specifically to which actors in the network the inserted node should establish relationships.

The ultimate goal is to provide the ability to make smarter decisions regarding undercover operations and node insertion when compared to current qualitative methodologies. While this analysis will be performed on randomly generated networks, the methodology remains applicable to real world networks. The direct application of this research is towards counterterrorism, but node insertion applies to the infiltration of any clandestine network, which could include human trafficking cells, gangs, or drug distribution networks, among others.

1.4 Research Scope

The scope of this research extends to only social networks. While node insertion could apply to computer network interdiction or disruption, the methodology does not address the translation to a cyber-based network. Within a social network sense, the scope is limited to the identification of relationships for a covert operative to make, and does not provide information on the strength of the intended relationships. Neither does it involve the selection of a specific network that would be most susceptible to node insertion, nor lend any insight to the actual covert action of the insertion.

1.5 Assumptions

Within the analysis, there exists several different stages. Governing assumptions are made initially and then dropped as the analysis progresses.

The first is that the network model is 100% certain, meaning all of the nodes and edges that are currently in the network model account for 100% of the true nodes and edges. The second is that the the network is unweighted and undirected, meaning that only the relationships existing between between actors matter, not the strength of the relationship, or where the influence in the relationship comes from. The third is that the network is not dynamic, and will not change as a result of the recommended insertion course of action. With this it is also assumed that the communities within the network are also formed with certainty. Networks following these assumption are under the category of Overt Networks.

The next assumption set are under the category of Clandestine Networks. As we encounter clandestine networks, the uncertainty within a network increases, allowing for the possibility of both missing nodes and missing edges. By allowing for this uncertainty, the fidelity of this system increases in that it more accurately models the current intelligence gathering and network disruption situations. However, it also increases the difficulty in recommending the truly best node insertion strategy.

The inclusion of dynamic networks in this analysis continues to increase both the fidelity and difficulty of the analysis, making the selection of a node insertion strategy a dynamic decision.

An assumption exists on the actual insertion of nodes. Operationally it would be unrealistic to attempt to create relationships with every node within a large network. The risk involved in creating this large number of relationships and the risk involved with remaining inconspicuous while maintaining them is extremely large. Knowing this, there is is a realistic limit on the number of relationships that can be created for an inserted node. The assumption then is that the maximum number of relationships that can be added is is a function of the total number of node within the network.

1.6 Organization

The structure of the research falls into four remaining chapters. The second chapter involves a literature review of the applicable techniques and measures involved in the quantification of the risk and benefit within node insertion. The third chapter outlines the methodology of research, and creates a framework for how the analysis will be performed, in addition to justification on why the framework is both correct and appropriate for the research. The fourth chapter applies the methodology to multiple experimental runs and case studies, allowing for a real world application of the proposed technique and theoretical results. In addition, it will provide results on the differences between node insertion techniques on differing networks. The fifth chapter discusses the results of the analysis, also analyzing the overall effect of the technique and identifying areas for future research.

II. Literature Review

Before attempting to explain node insertion analysis techniques among different networks it is important to deliver a review of the past Social Network Analysis (SNA) techniques, the understanding of which is crucial to the following research. Leveraging past research allows for the research at hand to complement previous works and provide a basis for the methodology used in the application of node insertion to real world networks. To do this, we highlight work with network structures, centrality measures, structural holes, network functionality, and network disruption.

Networks in general can range from simple adjacency matrices where only the presence of a connection is important, to weighted, and directed networks which show not only the strength of a connection, but also the directions of each connection, an example of which is shown in Figure 1. Overt networks are characterized as being known with certainty; they are often the focus of SNA, historically.

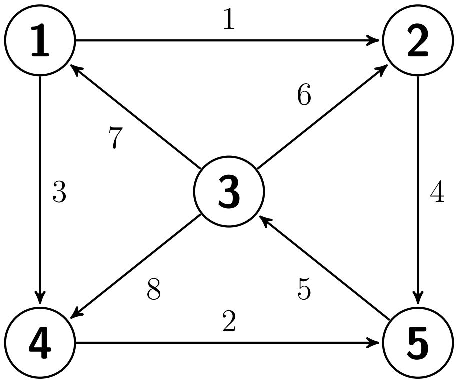


Figure 1. Weighted Directed Network Example

As fewer assumptions are made within the network of interest, SNA becomes more difficult. With Clandestine networks, where networks are uncertain, the stability of analysis becomes an issue [20]. Dynamic networks, where a network changes in response to stimuli, is no longer out of the question and precautions need to be made in order to ensure actions do not affect the network in a way detrimental to the stakeholders of the analysis.

In order to explain the measures identified in the following sections, an example graph will be used. This graph is shown in Figure 2.

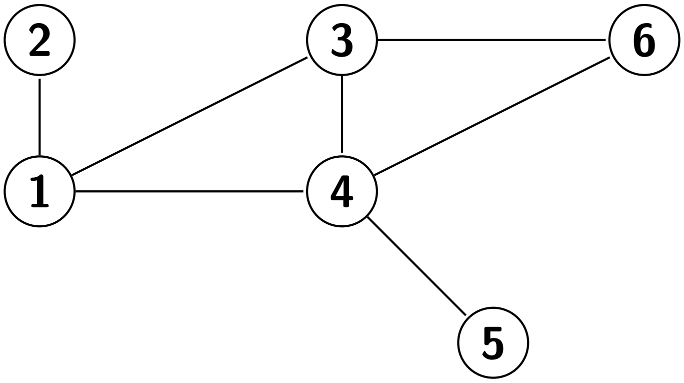


Figure 2. Example Network

2.1 Network Structure

In the study of networks, the importance of the overall structure has been made known by authors such as ös and Réyni, Watts and Strogatz, and Barabási and Albert. The structure itself of the network gives insight to the actors with high social capital, purely based on the presence or absence of lines of communication. While looking solely at structure may lack the insight that could be gained from an analysis of the characteristics of actors and the purpose behind the links, studies of the structure allow for an initial profile of the network to be created.

Scale Free.

The first network structure of interest is the scale free network, introduced by Barabási and Albert in 1999 [21]. This network type is based on the degree distribution of the network, which allows for the elicitation of certain parameters that describe the extent to which a network is scale free. Given a graph $G = (V, E)$, with V as the set of vertices in G and E as the set of edges between each vertices in V. The degree of a node i, or d_i, is the number of edges that come from that node and extend out to other nodes. The mathematical interpretation of this is shown in Equations 1a and 1b.

$$d_i = \sum_{j \in E} A_{ij} \quad \forall \quad i \tag{1a}$$

$$d_j = \sum_{i \in E} A_{ij} \quad \forall \quad j \tag{1b}$$

The ith, jth position in A_{ij} represents the existence of a connection or the weight of a relationship between the ith and jth nodes of the adjacency matrix $\mathbf{A}$. For an unweighted, undirected network is binary. Because of this, the indegree of node i, the number of relationships coming into node i, is equivalent to the outdegree of node i, the number of relationships coming out of node i. This is not necessarily the case with a directed network where A_{ij} is not always equivalent to A_{ji}. When this occurs, the indegree and outdegree become nontrivial.

In a scale free network, higher degree nodes are less common than low degree nodes. The degree distribution of a network is based on the overall frequency that a degree of k occurs. Examples of scale free networks include citation networks, and airline travel networks [22].

Networks of this structure are defined using the idea of power laws, and more specifically with the exponent of the power law, α, which is a constant for each network [23]. Equations 2 and 3 below define the relationships between p_k, k, and α. p_k is this frequency of the appearance of a node with a degree equal to k. According to

Newman, the exponent of the power law, α,typically has values between 2 and 3 for scale free networks.

$$\ln p_k = -\alpha \ln k + c \tag{2}$$

$$p_k = Ck^{-\alpha} \tag{3}$$

C is defined as the constant, e^c. Newman also explains that bias occurs when using these straight line fits to determine the α for a particular network. To remedy this problem, he identifies Equation 4 as a more dependable way to calculate the exponent for a network [24].

$$\alpha = 1 + N\left[\sum_i \ln \frac{d_i}{d_{min} - \frac{1}{2}}\right]^{-1} \tag{4}$$

d_i is the degree of node i and k_{min} is the lowest degree for which the power law holds. This d_{min} value is such that the distribution is monotonically decreasing and for Equation 4 works well for $d_{min} > 6$ [23].

An example of a scale free network, specifically the 9/11 terrorist network collected by Krebs, is shown in Figure 3.

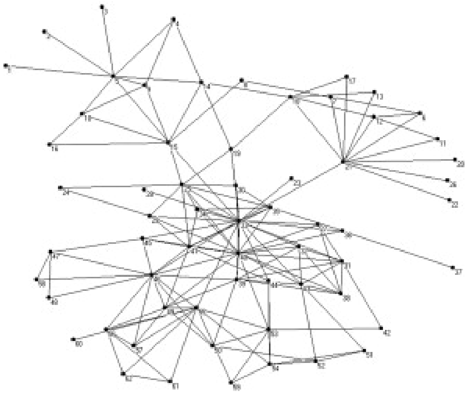


Figure 3. Krebs' 9/11 Terrorist Network Scale Free [1]

Random Graphs.

The second network structure is that of a random graph, which is generated from a uniform distribution with exactly n nodes and m edges. Newman identifies that a random graph is not defined in terms of a single network but with an "ensemble of networks", where there is a distribution for all graphs where $P(G) = \frac{1}{\Omega}$ where $P(G)$ is probability of a certain random graph appearing. [23]. Ω defines the total number of possible graphs that can occur in the ensemble [23]. The seminal paper on this structure was written by Erdös and Rényi in 1959 [25]. An example of what is now called an ER random graph is shown in Figure 4.

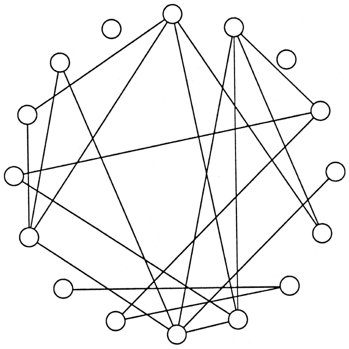


Figure 4. Example ER Random Graph [2]

The generation of random graphs can be governed by necessitating the structure follow certain parameters, such as the number of edges, mean degree, and even global clustering coefficient. True ER Random graphs have a global clustering coefficient of essentially zero. Normally, random graphs tend to have a small diameter, usually around $\frac{\log n}{\log np}$, given the expected degree of a node is at least 1 [26].

These graphs can also be created using the idea of preferential attachment, where nodes with high degree have a higher probability of being a part of a newly created edge, in order to create random scale free networks. The degree distributions of random graphs also seem to follow a more bell-shaped degree distribution, in contrast to the monotonically decreasing degree distribution of true scale free networks.

Small World.

The work of Watts and Strogatz in the late 1990s introduced the small world network model as a mix between Erdős-Rényi random networks and simple lattice networks with boundary conditions [27]. A lattice network is one where no variation exists within the degree and relationship pattern of each node. The result is a constant local clustering coefficient, which is further explained in Section 2.5.

The difference between a circle network and a random network comes from the act of rewiring, or the exchanging or relocation of the edges of a specific actor [27]. For this process, the parameter p defines the probability of the occurrence for a rewiring of a node. The probability remains the same for each node, no matter its degree, which is different from preferential attachment, where higher degree nodes have a higher probability of receiving an additional edge.

A selected node can undergo one of two different rewiring models. The first is where a node is selected for rewiring and one of its current edges is deleted, followed by the addition of a new edge, or "shortcut" [23]. The second model does not delete the initial edge, but solely performs the rewiring. It is important to note that if p is equal to 1, the result of the rewiring process would be a random graph with n nodes and m edges.

Examples of a small world network includes the famed “Six Degrees of Separation”, which states that no one person is more than six contacts way from another person in terms of relationships. Biological networks also seem follow the small world network model [28]. An example of a small world network is shown in Figure 5 .

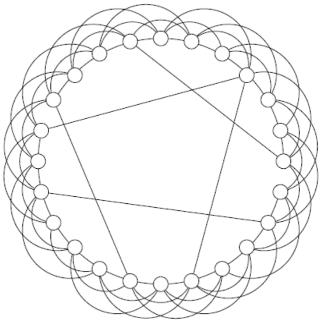


Figure 5. Watts and Strogatz' Small World Model [3]

2.2 Centrality Measures

Throughout the study of networks in the last sixty years, especially social networks, arguably the most influential and well studied measures are those involved with centrality. In general, a node's centrality allows for the measure of its overall importance in the network. The idea of centrality, first introduced by Alex Bavelas in 1948 [29] as a means for explanation of human communication, has evolved into the multiple measures we know today including the centralities of degree, closeness, betweenness, and eigenvector, among others .

While it is generally accepted that the influence of an actor within a network is strongly correlated with its centrality, different centrality measures allow for strongly competing ideas as to which node has the most power, or has the most social capital as defined by Newman [23]. Others like Cook *et al.* [30], and Bonacich [31] have determined the need for a family of centrality measures to define the power that an actor has in a network. The following sections will compare the four centralities mentioned because collectively, these four measures seems to describe nodes adequately enough to allow for sufficient analysis [6].

Types of Centrality Measures.

According to Everett and Borgatti [32], three types of centrality measures exist, which encompass the possible conceivable measures known thus far. The three types include induced, endogenous, and exogenous centralities. Induced centralities are made up of any measures that involve the calculation of a change in some network structure, specifically dealing with graph invariants, or properties which depend on "graph structure and not on a representation or a labeling of a graph" [32]. This induced centrality, $C_f(x)$, is specifically defined in Equation 5.

$$C_f(x) = f(G) - f(G - \{x\}) \tag{5}$$

The variable x is the removed entity, which extends to a vertex or an edge. $G - \{x\}$ then is the subgraph with the entity of interest removed. Because induced centralities are a result of the difference between a graph and a particular subgraph, it follows that any two vertex or edge removals that result in isomorphic subgraphs, or the same subgraphs, should have the same induced centrality.

Everett and Borgatti also show that the induced centrality of a removed entity is made up of both endogenous and exogenous centralities, relating induced centrality to "total centrality" [32]. Endogenous centralities are the centralities associated with solely the entity of interest, whereas exogenous centralities are associated with the entity of interest and its neighbors. Within a directed network for example, degree centrality, previously outlined in Section 2.2, is made up of in-degree, the endogenous centrality, and out-degree, the exogenous centrality [32].

Degree Centrality.

Degree centrality is the simplest of the centrality measures. Introduced by Bavelas and mathematically defined by Freeman, it measures the number of contacts an actor is involved with. In Freeman's words, "the degree of point, p_i, is simply the count of the number of other points, where $i \neq j$, that are adjacent to it and with which it is, therefore, in direct contact" [33]. While other more involved centrality measures exist, degree centrality allows for the initial analysis of a network, relating overall social capital to the number of contact an actor maintains. It acts as the backbone for all other measures explained in the following sections. It is calculated using Equation 1a or 1b, shown in Section 2.1.

The degree centralities for the example network in Figure 2, along with the unity-based normalization values, are shown in Table 1.

Table 1. Example Network Degree Centrality

Node	Degree	Normalized
1	3	0.67
2	1	0.00
3	3	0.67
4	4	1.00
5	1	0.00
6	2	0.33

Node 4 is the most central actor with a degree of four, while nodes two and five seem have the least importance. In terms of degree centrality, node 4 is the most powerful actor.

Closeness Centrality.

Viewed as the independence of an actor, closeness centrality relates power of an actor to "the extent that it can avoid the control potential of others" [33]. Freeman also highlights Beauchamp's definition, which follows more of a behavioral science model. For Beauchamp, closeness is "optimum...efficiency" [34]. In more specific terms, closeness is defined as the average geodesic distance from point p_i to p_j where $i \neq j$ for all combinations of vertices in the network. The geodesic distance more simply is the shortest path from one vertex to another and is also used in measuring network functionality, explained in Section 2.5.

Newman identifies closeness in multiple ways, shown in Equations 6a and 6b [23].

$$C_i = \frac{1}{\ell_i} = \frac{n}{\sum_j d_{ij}} \tag{6a}$$

$$C'_i = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}} \tag{6b}$$

d_{ij} is defined as the length of the geodesic path from node i to j; and ℓ_i is the farness of from node i, or the mean of all geodesic paths from each node to every other node. Geodesic distance is further explained in Section 2.5. According to Newman, Equation 6a is the most widely used due to the fact that any unconnected components in social networks are generally disregarded. Equation 6b is the harmonic mean closeness and allows for the calculation of closeness centrality for networks with multiple components, or ones that are not connected.

Newman also identifies some weaknesses of closeness centrality relating to the range of the values within a network. This makes it difficult to determine differences between actors and allows for large variance in the closeness of each node with minimal structure change. This phenomenon was analyzed by Herland with results which support Newman's claims [35]. The closeness of an actor dropped dramatically with the addition of new nodes and edges, which relates to an increase in the average geodesic distance for the network. The amount of variance within the individual scores also changes across networks, suggesting that different networks respond differently to changes in their structure [35].

In terms of node insertion, the idea of closeness is relative. If an inserted node has high closeness, the node could be involved with a large amount of the information sharing within the network. While the benefit to such a position could be high, from an information collection perspective the risk for this asset could be even higher, coming from possible interactions with highly connected members of a network. Analysis must take into account these risks when determining the best possible scenario for node insertion. Closeness centrality does not allow for the calculation of induced centrality because the removal of a node or edge could result in a disconnected graph [32].

The closeness centralities for the example network in Figure 2 using Equation 6a are shown in Table 2.

Table 2. Example Network Closeness Centralities

Node	Closeness	Normalized
1	0.1429	0.69
2	0.0909	0.00
3	0.1429	0.69
4	0.1667	1.00
5	0.1000	0.12
6	0.1111	0.27

Node four has the highest closeness centrality, while nodes two and five have the lowest. The small range in values addressed by Newman is seen here as well. The range from node four to node two is only 0.0758.

Betweenness Centrality.

Betweenness centrality identifies actors that lie between different paths to connecting other actors, and as suggested by Bavelas, is another measure of centrality [29]. Freeman states that “a person in such a position can influence the group by withholding or distorting information in transmission” [33]. Everett and Borgatti [32] highlight betweeness as an endogenous centrality measure. The equation to calculate the betweenness of a node i, or b_i, is shown in Equation 7 . Freeman also identifies that calculating betweenness becomes increasingly complex when more than one geodesic path connects a set of points.

$$b_i = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}, i \neq j \neq k \tag{7}$$

g_{ijk} is defined as the number of geodesic paths from node i to node j that include node k, whereas g_{ij} is the total number of geodesic paths from node i to j. The betweenness centralities for the example network in Figure 2 are shown in Table 3.

Table 3. Example Network Betweenness Centralities

Node	Betweenness	Normalized
1	8	0.80
2	0	0.00
3	2	0.20
4	10	1.00
5	0	0.00
6	0	0.00

Node four has the highest betweenness centrality, followed by node one. The three nodes that have betweenness of zero do not lie on any shortest path. If the aim of network interdiction was to isolate two sets of nodes, it would be beneficial to remove the nodes of high betweenness.

Eigenvector Centrality.

Eigenvector centrality is a measure of indirect connections and relates the social capital of an actor to the social capital of his connections, meaning an influential actor is connected to other influential members of a network. Emphasized by Bonacich in 1972 [36], this centrality measure takes into account the entire structure of the network, and not just the structure of the ego within a network. An ego, according to Borgatti is "the person whose social capital we are measuring" [37]. It explicitly comes from the eigenvector, $\vec{x}$, associated with the leading eigenvalue, λ_i, in the adjacency matrix, a value that satisfies $\mathbf{A}\vec{x} = \lambda\vec{x}$. Alternatively, this is the eigenvalue that satisfies the inequality $\lambda_i > \lambda_k$ for any $k \neq i$.

Because finding the eigenvalues and eigenvectors of a matrix can be computationally expensive, other methods can be implemented. The power law method is one such way to decrease this computational load and involves solely the calculation of the leading eigenvector and leading eigenvalue of the adjacency matrix. The method involves an initial nonzero vector x_0 and the matrix multiplication of this vector and the adjacency matrix $\mathbf{A}$ over multiple iterations, until the change occurring between x_i and x_{i+1} is less than some predetermined threshold, usually a few orders of magnitude larger than the machine accuracy value of the coding language used. The pseudocode for this algorithm is shown in Figure 6.

```
Algorithm 1 Power Method($A, \epsilon, error$)
  Initialize $x_0, \mu_0$
  while $error > \epsilon$ do
      $C = \mathbf{A} \times x_0$
      if $\max(C) > \min(C)$ then
          $\mu = \max(C)$
      else
          $\mu = \min(C)$
      $x_0 = \frac{1}{\mu} \times C$
      $error = |\mu_0 - \mu|$
      $\mu_0 = \mu$
  end
  return $x_0, \mu$
```

Figure 6. Power Method Pseudocode

x_0 is a random, nonzero $1 \times m$ vector. ϵ is the acceptable tolerance while *error* itself is the difference between iterations of the eigenvalue, μ_0. These values will converge to the eigenvector and eigenvalue of the matrix respectively. For the example network shown in Figure 2 the eigenvector centralities are shown in Table 4.

Table 4. Example Network Eigenvector Centralities

Node	Value
1	0.4491
2	0.1631
3	0.5095
4	0.5641
5	0.2048
6	0.3899

Node four has the highest eigenvector centrality, but it is closely followed by node three. Previously, nodes one and three were tied in importance, but a shift in rank occurs when dealing with eigenvector centrality due to node three's connection to nodes one, four, and six, three highly connected network members. Node one connects to nodes three, four, and two, the latter of which is not well connected.

2.3 Community Structure

At a strategic level, the analysis of the interaction between networks dominates. At the operational level of analysis, interactions between network subgroups become important. At the tactical level, the interactions between individual nodes become crucial. These subnetworks, or communities, bring different information to the foreground following analysis. The allow for the assessment of different network components, and the relationships between smaller groups of actors within the network, instead of a full network analysis, or an individual node analysis.

In graph theory, a clique is defined as a subset of vertices of a graph where each vertex connects to every other vertex in the subset [38]. The formation of communities is far less stringent, but relates to the same concept. Girvan and Newman in 2002 explain that this phenomenon occurs when "subsets of vertices within which vertex-vertex connections are dense, but between which connections are less dense" [28]. They exist because they seem to provide a natural barrier from outside intrusions [4].

The process of determining where a community exists relates to the problem of graph partitioning, one of Karp's original non deterministic polynomial (NP) complete proofs [39]. Due to this, heuristic methods are necessary to approach a near optimal solution, without an optimal solution being guaranteed. Multiple methods exist for this purpose, including hierarchical clustering and edge betweenness. The former attempts to find the central points in a community and the latter attempts to find points not central to any one community, or broker points, which make connections between communities [28]. These brokers also tend to fill structural holes, which are further explained in Section 2.4.

The most common method of determining communities is the method of modularity maximization [40], which describes the extent to which like actors, or actors of the same community, are connected to each other. The equation for modularity, Q, for a network is shown in Equation 8a [23].

$$Q = \frac{1}{2m} \sum_{ij} B_{ij} \delta(c_i, c_j) \tag{8a}$$

$$B_{ij} = A_{ij} - \frac{d_i d_j}{2m} \tag{8b}$$

c_i defines which community node i belongs to; $\delta(c_i, c_j)$ is 1 if c_i and c_j are the same, and zero otherwise; and B_{ij} is known as the modularity matrix, which is calculated using Equation 8b. B_{ij} is the difference between A_{ij} and the ratio of connections between groups that would have occurred at random, identified by $\frac{d_i d_j}{2m}$ where m is the total number of edges in G.

Newman and Girvan extend this technique to a much simpler equation to determine the modularity for unweighted and undirected networks, shown in Equation 9.

$$Q = \sum_{i=1}^{k} \left(\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right) \tag{9}$$

e_i is the number of vertices within community i that do not go between communities, where k is the total number of communities. d_i is the sum of degrees of each node within community i. m is the total number of vertices in the network.

The tendency of communities to exist between nodes of the same demographic has also been studied, and is known as assortative mixing. With this, the communities are not generated based on the modularity, but are created using other qualities like age, race, sex, and even political background [23]. The modularity is then calculated to assess whether or not the communities in the network actually form based on these demographics. The calculation of this *assortivity coefficient* [23] is very similar to Equation 8a, but the $\delta(c_i, c_j)$ is replaced with just $x_i x_j$ where x_i is the demographic value for node i. This is then normalized by dividing over a perfectly mixed network where edges only fall between vertices of the same demographics. The calculation of the assortivity coefficient, r, is shown Equation 10.

$$r = \frac{\sum_{ij}(A_{ij} - d_i d_j / 2m) x_i x_j}{\sum_{ij}(d_i \delta_{ij} - d_i d_j / 2m) x_i x_j} \tag{10}$$

x_i is the value associated with the demographic of node i. δ_{ij} is the ith, jth position in the Kronecker matrix of a perfectly mixed network, where a one exists if node i and node j are of the same demographic, and a zero exists otherwise. Assortative mixing by degree is the tendency of nodes to connect to other nodes of like degree. For this, Equation 10 is altered to include the degrees of nodes i and j, instead of the demographics of i and j. This alteration is shown in Equation 11.

$$r = \frac{\sum_{ij}(A_{ij} - d_i d_j / 2m) d_i d_j}{\sum_{ij}(d_i \delta_{ij} - d_i d_j / 2m) d_i d_j} \tag{11}$$

Figure 7 shows the community detection analysis of Rocco and Marquez applied to the Krebs 9/11 network previously shown in Figure 3 [4].

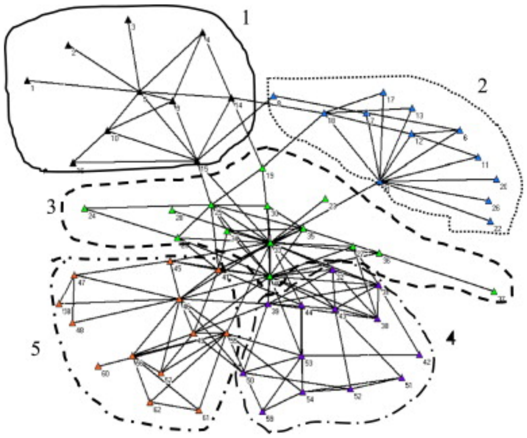


Figure 7. Krebs 9/11 Network Community Structure [4]

Rocco and Marquez [4] use the Fast Modularity technique in an application of community structure to the idea of vulnerability sets, or vertices that could eliminate the connections between communities.

The example network was analyzed to determine its community structure, which is shown in Table 5.

Table 5. Example Network Community Structure

Node	Community
1	1
2	1
3	2
4	2
5	2
6	2

It seems that nodes 1 and 2 are within the same community, with the remaining nodes being a part of the second community.

The use of community analysis could allow for a node to be inserted into a community that is less beneficial, perhaps based on a relative lack of modularity, but also less dangerous in terms of risk than being involved, or associated with, a more densely connected group.

2.4 Structural Holes

First introduced by Burt 1992 as an expansion of previous network structure analysis, structural holes exist as "separation between nonredundant contacts" [41]. Redundancy exists when contacts are connected to the same contacts as their neighbors, similar to the triads that are created when addressing clustering within a network.

Burt identifies two types of redundancy; cohesion and structural equivalence. Cohesion occurs when two contacts are strongly connected to each other, whereas equivalence occurs when two actors "are related in the same ways to the same other points" [42]. An example of cohesion is shown in Figure 8. The more cohesive relationships are shown with thicker black lines, while the less cohesive relationships are shown with thin lines. This might correspond with higher weights associated with the more cohesive pairs in a weighted network.

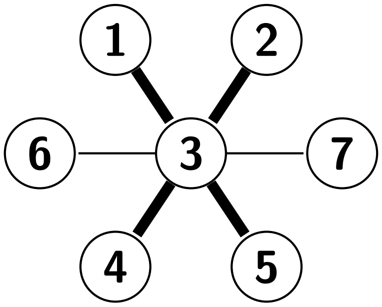


Figure 8. Cohesion Example

Figure 9 shows a network where all relationships are equivalent, which is the case for all unweighted networks.

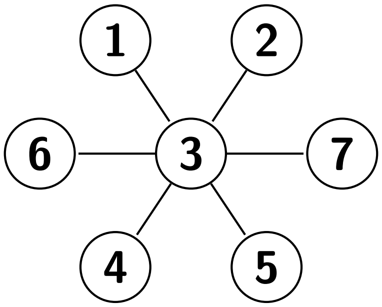


Figure 9. Equivalence Example

Redundancy by structural equivalence comes from similar connections to other actors. Burt identifies that two structurally equivalent actors are privy to the same information. The difference between the two types of redundancy stems from the difference between direct and indirect relationships. A competitive advantage exists for actors that fill these holes [41].

The first measure identified as a part of structural hole analysis is the efficiency of a certain ego, or the actor of interest, within a network. This is a measure of the overall redundancy in which an ego is involved. The higher the efficiency of the ego, the less its neighbors are also each other's neighbors. Higher degree egos with low efficiency provide the same informational benefits as lower degree egos with the same efficiency. Figure 9 shows a network where each node has the same efficiency. No neighbors of any node are connected. The efficiency of a community could affect the amount of risk involved in placing a new actor amongst the current network actors, similar to the extent to which a community functions, as explained in Section 2.3.

Effective Size and Efficiency.

Borgatti expresses Burt's efficiency measure in three parts, redundancy, effective size, and total size [37]. Redundancy, r_i, is shown in Equation 12a and represents the contacts of the ego that are also contacts of each other. Effective size, s_i, is shown in Equation 12b and expresses the difference between the amount of redundancies and total relationships for an ego, which is total size. Total size, d_i is the degree of ego i as explained in section 2.2. It is important to note that the matrix $\mathbf{A}$ still represents the connections in the network and is symmetric and binary based on Borgatti's assumption of a connected, unweighted, undirected network. A_{ij} then represents the connection between actor i and j, when $i \neq j$. When $i = j$ the value is 0. These assumptions result in the equations for redundancy, effective size, and efficiency [37].

$$r_i = \sum_j \sum_k \frac{A_{ik} A_{kj}}{d_i} \qquad \forall \qquad i : (i, j) \in \mathbf{A} \tag{12a}$$

$$s_i = d_i - r_i \tag{12b}$$

$$e_i = \frac{s_i}{d_i} \tag{12c}$$

When Borgatti's initial assumption drops, and $\mathbf{A}$ is no longer symmetric and binary, the calculations for s_i change immensely as shown in Equation 13 [37].

$$s_i = \sum_j [1 - \sum_q p_{iq} m_{jq}] \qquad \forall \qquad q \neq i, j \tag{13a}$$

$$p_{iq} = \frac{A_{iq} + A_{qi}}{\sum_j (A_{ij} + A_{ji})} \qquad \forall \qquad i \neq j \tag{13b}$$

$$m_{jq} = \frac{A_{jq} + A_{qj}}{max_k (A_{jk} + A_{kj})} \qquad \forall \qquad j \neq k \tag{13c}$$

p_{iq} is the ith, qth entry in the row-stochastic matrix $\mathbf{P}$, which has weighted entries based on the ties of each node. The matrix $\mathbf{P}$ for the example network is shown in Figure 10. m_{jq} becomes A_{jq} in an unweighted, undirected network.

$$\begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 & 0 & 0 \\ 1.0 & 0 & 0 & 0 & 0 & 0 \\ 0.33 & 0 & 0 & 0.33 & 0 & 0.33 \\ 0.25 & 0 & 0.25 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0 & 1.0 & 0 & 0 \\ 0 & 0 & 0.50 & 0.50 & 0 & 0 \end{bmatrix}$$

Figure 10. Example Row-Stochastic Matrix

The effective size and efficiencies of the example network in Figure 2 are shown in Table 6.

Table 6. Example Network Structural Holes Measures

Node	Effective Size	Efficiency
1	2.3333	0.7778
2	1.0000	1.0000
3	1.6667	0.5556
4	3.0000	0.7500
5	1.0000	1.0000
6	1.0000	0.5000

The highest efficiency occurs with nodes two and five, which have degree one. This is a trivial find. However, node one has an efficiency of 0.77 and has more than degree one. This means 77% of this nodes neighbors are not connected to each other, resulting in a non-redundant structure, where each relationship means more because node one fills a structural hole.

Structural Hole Constraint Measures.

The two final parts of Burt's structural holes measures include network constraint and indirect constraint. Network constraint describes brokerage opportunities. Specifically, it "measures the extent to which a manager's time and energy are concentrated in a single group of interconnected colleagues" [41]. This constraint value is made up of three parts, which Burt identifies as C-size, C-density, and C-hierarchy. Indirect constraint however is the "average network constraint on ego's direct contacts" [41]. The higher the constraint, the lower the access to structural holes [37].

2.5 Network Functionality

Network functionality and its associated measures attempt to describe the network in its entirety, based on the characteristics of the individual nodes. While networks are classified into three different types: scale free, random, and small world, networks of the same type can have significantly different functionality.

Geodesic Distance.

The first functionality measure involves the idea of geodesic distance, or the shortest path, d_{ij}, from node i to node j . The geodesic distance from one vertex to another directly effects network communication. Network functionality then is the speed of this communication. Natural disaster response efforts benefit from a smaller average geodesic, whereas network interdiction benefits from increasing the average geodesic distance as much as possible. Interdiction techniques might also benefit from an decrease in average geodesic distance given the travel of misinformation across the network.

Holme and Kim [43] advocate the use of the average geodesic distance, ℓ, to determine the overall functionality of a network, shown in Equation 14a. Equation 14b allows for real values to occur when analyzing disconnected graphs, due to the fact that the lack of a path between any two nodes results in an infinite path length when using Equation 14a.

$$\ell = \frac{1}{N(N-1)} \sum_i \sum_j d_{ij} \tag{14a}$$

$$\ell^{-1} = \frac{1}{N(N-1)} \sum_i \sum_j \frac{1}{d_{ij}} \tag{14b}$$

When using Equation 14b, the higher the value of ℓ^{-1}, the higher the functionality of the network. For the network in Figure 2, the average geodesic distance ℓ is 1.6667, while ℓ^{-1} is 0.7111. These values show the relatively high speed of travel for this extremely small network.

Clustering Coefficients.

While the idea of modularity, explained in Section 2.3, is a full network based measure, clustering coefficients can determine the density of connections for an individual node as well. The difference occurs between local and global clustering coefficients. The local clustering coefficient for a node i is a value representing the extent to which a node's neighbors are connected to each other. For node i, this value, c_i is calculated using Equation 15.

$$c_i = \frac{r_i}{d_i - 1} \tag{15}$$

d_i is the degree for node i and r_i is defined as the redundancy of node i as shown in Equation 12a.

With a local clustering coefficient for each node in the network, a global clustering coefficient can be calculated. Watts and Strogatz provide a measure of this value shown in Equation 16a that is different than Newman's definition, shown in Equation

16b.

$$C = \frac{1}{n}\sum_{i}^{n} C_i \tag{16a}$$

$$C = \frac{\text{\# of triangles} \times 3}{\text{\# of connected triples}} \tag{16b}$$

Newman identifies that the Equation 16a is dominated by lower degree nodes, and can give an inaccurate description of the overall network, but is more widely used because of its earlier inception.

Both the local and global clustering coefficients could be very beneficial for node insertion analysis from an individual and complete network standpoint. The higher a local clustering value for a given node, the more redundant the relationships are, and the riskier it might be to successfully implement an insertion at that point. In addition the change in the global clustering value once a node is inserted could also be used as a risk measure. It is important to note that local clustering is negatively correlated with the efficiency measure mentioned in Section 2.4.

The local clustering coefficients of the network in Figure 2 are shown in Table 7. The global clustering coefficient using Equation 16a is 0.3889, which means that on average, 38.89% of a node's neighbors are connected to each other.

Table 7. Example Network Local Clustering Values

Node	Value
1	0.3333
2	0.0
3	0.6667
4	0.3333
5	0.0
6	1.0000

Node 6 has the highest local clustering because each of its neighbors are connected to each other, whereas node 4 has a lower local clustering value because only two of the possible six connections of its neighbors are present in the network. It is important to note that a nodes local clustering value means nothing unless taken into account with its overall degree.

Laplacian Centrality.

Laplacian centrality is a measure introduced in 2013 as an extension of Gutman's work with Laplacian energy [44]. Specifically stated in Qi *et al.* [45], Laplacian energy reflects the internal connectivity of a network and is based off of the graph Laplacian, L, shown in Equation 17. Because the Laplacian centrality can be calculated using the difference in structural characteristics, specifically based on the removal of a node, this measure is an induced measure and provides a total centrality for a removed entity.

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{17}$$

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & \cdots \\ 0 & d_2 & 0 & \cdots \\ \vdots & 0 & \ddots & \ddots \\ \vdots & \vdots & \ddots & d_n \end{bmatrix} \tag{18}$$

D is defined as the degree matrix, shown in Equation 18, where each d_i is the degree of node i, and A is the adjacency matrix for a graph G [46]. The Laplacian energy of a graph is defined in Equation 19a with λ_i representing the ith eigenvalue of the graph Laplacian. The initial derivation of Laplacian energy was made by Lazić and is shown in Equation 19b, followed by the simplification shown in Equation 19b [47].

$$E_L(G) = \sum_{i=1}^{n} \lambda_i^2 \tag{19a}$$

$$E_L(G) = \sum_{i=1}^{n} d_i^2 + d_i \tag{19b}$$

Equation 19a is computationally expensive when large networks are used due to the computational requirements involved in finding the eigenvalues of the Laplacian.

Laplacian centrality relates to the change in connectivity; the difference between a graph's current energy and the energy when a vertex, i, is removed [45]. Qi goes from the definition of Laplacian energy in Lazić [47] to a related definition of Laplacian centrality, or the drop in Laplacian energy, L_{E_i}, following a node removal, defined in Equation 20.

$$L_{E_i} = (\Delta E)_i = d_i^2 + d_i + 2 \sum_{v_j \in N(i)} d_j \tag{20}$$

$N(i)$ is the set of adjacent vertices of vertex i in the graph G. d_i is then the degree of node i in G. d_j then is the degree of the node j, which is a neighbor of node i. When applied to other well studied social networks, Laplacian Centrality seems to perform well, while providing perspective on the intermediate structure of a network.

The Laplacian centralities for the example network in Figure 2 are shown in Table 8.

Table 8. Example Network Laplacian Centralities

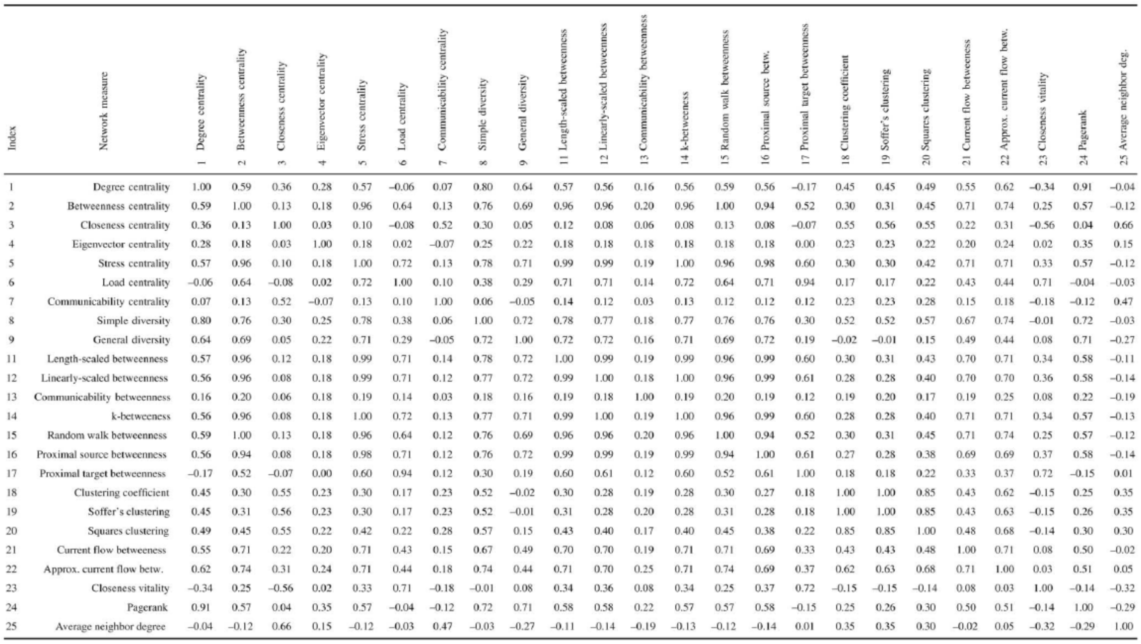
Node	Laplacian	Normalized
1	28	0.67
2	8	0.00
3	30	0.73
4	38	1.00
5	10	0.07
6	20	0.40

Nodes 3 and 4 have the highest Laplacian centralities with values of 30 and 38 respectively.

2.6 Measure Stability

In the arena of SNA, there are countless measures used to determine influential actors, network functionality, and clusterability. Guzman *et al.* studied twenty-five social networks measures to determine the relationship between them [6]. The results from the research are shown in Table 9.

Table 9. Guzman's Network Measure Rank Correlations [6]



Index	Network measure	1 Degree centrality	2 Betweenness centrality	3 Closeness centrality	4 Eigenvector centrality	5 Stress centrality	6 Load centrality	7 Communicability centrality	8 Simple diversity	9 General diversity	11 Length-scaled betweenness	12 Linearly-scaled betweenness	13 Communicability betweenness	14 k-betweeness	15 Random walk betweenness	16 Proximal source betw.	17 Proximal target betweenness	18 Clustering coefficient	19 Soffer's clustering	20 Squares clustering	21 Current flow betweenness	22 Approx. current flow betw.	23 Closeness vitality	24 Pagerank	25 Average neighbor deg.
1	Degree centrality	1.00	0.59	0.36	0.28	0.57	–0.06	0.07	0.80	0.64	0.57	0.56	0.16	0.56	0.59	0.56	–0.17	0.45	0.45	0.49	0.55	0.62	–0.34	0.91	–0.04
2	Betweenness centrality	0.59	1.00	0.13	0.18	0.96	0.64	0.13	0.76	0.69	0.96	0.96	0.20	0.96	1.00	0.94	0.52	0.30	0.31	0.45	0.71	0.74	0.25	0.57	–0.12
3	Closeness centrality	0.36	0.13	1.00	0.03	0.10	–0.08	0.52	0.30	0.05	0.12	0.08	0.06	0.08	0.13	0.08	–0.07	0.55	0.56	0.55	0.22	0.31	–0.56	0.04	0.66
4	Eigenvector centrality	0.28	0.18	0.03	1.00	0.18	0.02	–0.07	0.25	0.22	0.18	0.18	0.18	0.18	0.18	0.18	0.00	0.23	0.23	0.22	0.20	0.24	0.02	0.35	0.15
5	Stress centrality	0.57	0.96	0.10	0.18	1.00	0.72	0.13	0.78	0.71	0.99	0.99	0.19	1.00	0.96	0.98	0.60	0.30	0.30	0.42	0.71	0.71	0.33	0.57	–0.12
6	Load centrality	–0.06	0.64	–0.08	0.02	0.72	1.00	0.10	0.38	0.29	0.71	0.71	0.14	0.72	0.64	0.71	0.94	0.17	0.17	0.22	0.43	0.44	0.71	–0.04	–0.03
7	Communicability centrality	0.07	0.13	0.52	–0.07	0.13	0.10	1.00	0.06	–0.05	0.14	0.12	0.03	0.13	0.12	0.12	0.12	0.23	0.23	0.28	0.15	0.18	–0.18	–0.12	0.47
8	Simple diversity	0.80	0.76	0.30	0.25	0.78	0.38	0.06	1.00	0.72	0.78	0.77	0.18	0.77	0.76	0.76	0.30	0.52	0.52	0.57	0.67	0.74	–0.01	0.72	–0.03
9	General diversity	0.64	0.69	0.05	0.22	0.71	0.29	–0.05	0.72	1.00	0.72	0.72	0.16	0.71	0.69	0.72	0.19	–0.02	–0.01	0.15	0.49	0.44	0.08	0.71	–0.27
11	Length-scaled betweenness	0.57	0.96	0.12	0.18	0.99	0.71	0.14	0.78	0.72	1.00	0.99	0.19	0.99	0.96	0.99	0.60	0.30	0.31	0.43	0.70	0.71	0.34	0.58	–0.11
12	Linearly-scaled betweenness	0.56	0.96	0.08	0.18	0.99	0.71	0.12	0.77	0.72	0.99	1.00	0.18	1.00	0.96	0.99	0.61	0.28	0.28	0.40	0.70	0.70	0.36	0.58	–0.14
13	Communicability betweenness	0.16	0.20	0.06	0.18	0.19	0.14	0.03	0.18	0.16	0.19	0.18	1.00	0.19	0.20	0.19	0.12	0.19	0.20	0.17	0.19	0.25	0.08	0.22	–0.19
14	k-betweeness	0.56	0.96	0.08	0.18	1.00	0.72	0.13	0.77	0.71	0.99	1.00	0.19	1.00	0.96	0.99	0.60	0.28	0.28	0.40	0.71	0.71	0.34	0.57	–0.13
15	Random walk betweenness	0.59	1.00	0.13	0.18	0.96	0.64	0.12	0.76	0.69	0.96	0.96	0.20	0.96	1.00	0.94	0.52	0.30	0.31	0.45	0.71	0.74	0.25	0.57	–0.12
16	Proximal source betweenness	0.56	0.94	0.08	0.18	0.98	0.71	0.12	0.76	0.72	0.99	0.99	0.19	0.99	0.94	1.00	0.61	0.27	0.28	0.38	0.69	0.69	0.37	0.58	–0.14
17	Proximal target betweenness	–0.17	0.52	–0.07	0.00	0.60	0.94	0.12	0.30	0.19	0.60	0.61	0.12	0.60	0.52	0.61	1.00	0.18	0.18	0.22	0.33	0.37	0.72	–0.15	0.01
18	Clustering coefficient	0.45	0.30	0.55	0.23	0.30	0.17	0.23	0.52	–0.02	0.30	0.28	0.19	0.28	0.30	0.27	0.18	1.00	1.00	0.85	0.43	0.62	–0.15	0.25	0.35
19	Soffer's clustering	0.45	0.31	0.56	0.23	0.30	0.17	0.23	0.52	–0.01	0.31	0.28	0.20	0.28	0.31	0.28	0.18	1.00	1.00	0.85	0.43	0.63	–0.15	0.26	0.35
20	Squares clustering	0.49	0.45	0.55	0.22	0.42	0.22	0.28	0.57	0.15	0.43	0.40	0.17	0.40	0.45	0.38	0.22	0.85	0.85	1.00	0.48	0.68	–0.14	0.30	0.30
21	Current flow betweeness	0.55	0.71	0.22	0.20	0.71	0.43	0.15	0.67	0.49	0.70	0.70	0.19	0.71	0.71	0.69	0.33	0.43	0.43	0.48	1.00	0.71	0.08	0.50	–0.02
22	Approx. current flow betw.	0.62	0.74	0.31	0.24	0.71	0.44	0.18	0.74	0.44	0.71	0.70	0.25	0.71	0.74	0.69	0.37	0.62	0.63	0.68	0.71	1.00	0.03	0.51	0.05
23	Closeness vitality	–0.34	0.25	–0.56	0.02	0.33	0.71	–0.18	–0.01	0.08	0.34	0.36	0.08	0.34	0.25	0.37	0.72	–0.15	–0.15	–0.14	0.08	0.03	1.00	–0.14	–0.32
24	Pagerank	0.91	0.57	0.04	0.35	0.57	–0.04	–0.12	0.72	0.71	0.58	0.58	0.22	0.57	0.57	0.58	–0.15	0.25	0.26	0.30	0.50	0.51	–0.14	1.00	–0.29
25	Average neighbor degree	–0.04	–0.12	0.66	0.15	–0.12	–0.03	0.47	–0.03	–0.27	–0.11	–0.14	–0.19	–0.13	–0.12	–0.14	0.01	0.35	0.35	0.30	–0.02	0.05	–0.32	–0.29	1.00

Of the 25 measures tested, 14 were highly correlated, or had a correlation above 0.85, with at least one other measure. However, when the measures outlined in section 2.2 were compared to each other, the correlations were very low. Eigenvector centrality and closeness centrality for example, only had a correlation of 0.03. Betweenness and closeness centralities have a correlation of 0.13. Degree centrality has the highest correlation with the other measures at 0.59 with betweenness, 0.36 with closeness, and 0.28 with eigenvector, but does not exceed to 0.85 threshold set by Guzman. The correlations for the centrality measures outlined are shown in Table 10.

Table 10. Centrality Measures Correlation Matrix

	Degree	Closeness	Betweenness	Eigenvector
Degree	1.0	0.36	0.59	0.28
Closeness		1.0	0.13	0.03
Betweenness			1.0	0.18
Eigenvector				1.0

This shows the measures deliver different information about the structure of the network of interest and the roles or importance of the network actors. This result indicates the four measures could be effective in determining separate objectives. This possible correlation, or lack of, with other measures could be beneficial to the node insertion analysis from a risk and benefit trade off perspective, as well as from a utility perspective, shown in Section 2.8. It is important to have uncorrelated risks and benefits in this analysis to achieve independent utilities for each node insertion possibility. The previously mentioned Laplacian centrality is relatively new and therefore, not a part of Guzman's' correlation analysis.

For larger networks the availability of alternative, more quickly calculated measures is important and crucial for the creation of faster algorithms and heuristics. The similarities are important to know from the node insertion risk and benefit point of view as well. If two similar measures are used in the analysis, one to determine risk and the other to determine benefit, the values for each of these objectives could be extremely correlated, resulting in inaccurate representations of the two objectives. For this reason, the measures for each of these objectives need to be as uncorrelated as possible.

2.7 Network Disruption

Previous research in SNA has dealt with the determination of key nodes and edges within a network, along with the functionality of the network given the removal of this node(s) or edge(s). It seems that these techniques could be leveraged to determine either high risk or high benefit nodes for insertion as well.

Malik *et al.* [48] and Corley *et al.* [49] correlate a vital arc or node to be one that the removal of which results in the largest increase in the shortest path from node i to j. This includes algorithms to determine the k most vital arcs, which is much more difficult than finding the single most vital arc, especially when there is not a designated source or sink node (which is the case for most social networks). Malik *et al.*'s algorithm solves in $O(m + n \log n)$ time where m is the number of arcs and n is the number of nodes [50].

With the assumption of an unweighted, undirected network, the k most vital arcs become trivial because each arc is weighted the same. However, there could be arcs or nodes where their removal results in a huge detour for travel to a certain node, more so than the use of triad in the network. This might happen when the redundancy of a node is low.

Other research has dealt with the removal of arcs and nodes not corresponding to an increase in a specific shortest path, but with the separation of nodes into multiple components, known as the isolation set problem. First introduced by Bellmore in 1970 [51], the purpose behind this problem is to eliminate arcs between nodes of interest, or distinguished sets, effectively isolating them from the other components within a network, with a minimum cost objective.

Herbranson [5] extended the original formulation to include costs of resources to cut arcs, in addition to the inclusion to both vertex cut-sets and vertex-edge cut-sets as possible solutions. Partial isolation sets are cuts that do not completely separate the distinguished sets, but isolate them to a certain percentage, P_i [5]. A distinguished set is the collection of nodes that a decision maker wishes to isolate from each other.

Herbranson's isolation set solution, with distinguished sets {{2,4},{6,7,8},{10,17,16}} is shown in Figure 11. In Figure 11, the isolations set solution shows the separation of nodes two and four from nodes six, seven, and eight.

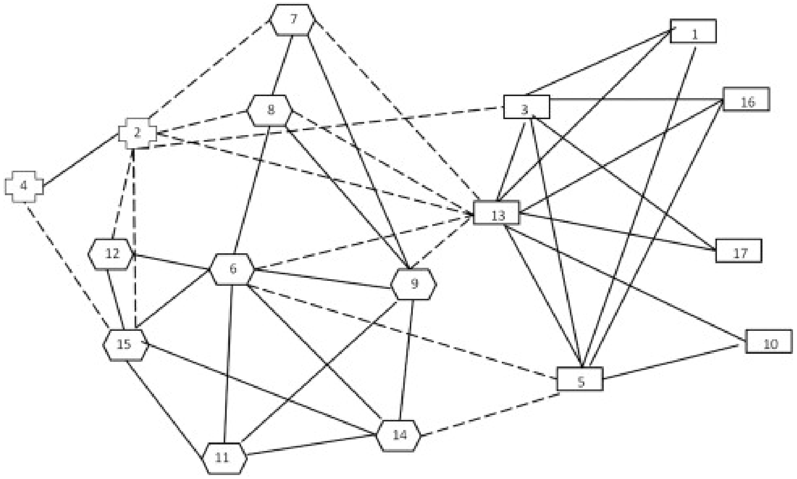


Figure 11. Herbranson Isolation Set Solution [5]

By applying isolation set techniques to node insertion, recommendations can be made with the future disruption of the network in mind, and does not limit the analysis to the benefit or risk of information gathering.

2.8 Risk and Benefit Within Networks

While the overall functionality of a network can be calculated using the measures previously provided, the determination of the importance of an actor is a much more difficult task, and is much of the subject of this research. Bonacich identified early on that the importance of overall actors cannot be based on simple measures, but necessitates the use of families of measures [31]. While different in application from Bonacich, the work of Jackson and Wolinsky [52] added to this idea by applying utility to each individual node based on geodesic paths and costs of maintenance for relationships. The main purpose behind the research was to determine the overall utility of possible connections in the network, which is explained sufficiently in Section 2.10. The utility function developed is shown in Equation 21 for each node i.

$$u_i(g) = w_{ii} + \sum_{j \neq i} \delta^{d_{ij}} w_{ij} - \sum_{j} c_{ij} \tag{21}$$

w_{ii} is the weight of a node in a network; δ is the benefit value of a connection within the range $[0, 1]$ and d_{ij} is the length of the shortest path from node i to node j and w_{ij} is the weight of a connection from node i to node j, or its "intrinsic value" [52]. Jackson and Wolinsky maintain that c_{ij} is the overall cost of maintaining arc (i, j) and is assumed to be nonnegative. They also apply this utility to the overall stability of networks, identifying how networks will change based upon the utility of an ego and the actors around it. This topic is explained in Section 2.10.

2.9 Clandestine Networks

The ideas mentioned in previous sections make an important assumption in order to provide seemingly valuable information for analysts. Early work in social networks assumed perfect information of the network connections, which includes no missing edges and no missing actors. Clandestine networks, or networks where the arcs and nodes are not known with 100 percent certainty, no longer make this assumption, but provide a network that is not assumed to be perfect. There is an expectation of missing information. Clandestine networks, or dark networks, include "terrorist organizations, drug-trafficking rings, arms-smuggling operations, gang enterprises", but are not limited to just criminal organizations [53]. The study of these networks, where the actors involved are actively trying to avoid detection, allows for a higher degree of fidelity within analysis, in addition to allowing for measure stability analysis given a possible alternative network reality.

However, imperfect information about a network could result in inaccurate conclusions that might be detrimental to the assets that organizations have inserted because of the analysis. Therefore, it is important to include the possibility of having incorrect and imperfect networks into node insertion analysis.

Carley *et al.* [20] identifies the dangers involved in networks with uncertainty from a quantitative standpoint. After finding the "key actors" of an uncertain network, and comparing to analysis with a full network with zero uncertainty, the results showed problems with assuming 100% certainty within a network. For a network including only 75% of the true nodes, the key actor was identified 65% of the time. Alternatively, a network with 75% certainty on the edges only allows for correct key actor classification approximately 57% of the time [20]. While the probabilities of finding the key actors are greater than random (or $\frac{1}{N}$ where N is the total number of nodes), it is significantly less than when the true network is known with 100% certainty. The findings imply that network and actor measures are not robust to imperfect information about the network.

Smith *et al.* in 2013 [54] introduced the idea of space-time threat propagation to determine the probability of a threat from an actor given the identification of an actor. The actual detection of covert actors is outside of the scope of this analysis, but it important to highlight the problems that could occur with clandestine networks and how these are being addressed regardless.

2.10 Dynamic Networks

When dealing with terrorist networks, a simplifying assumption is that the current network remains the same for an extended amount of time is often used. This simplification degrades the fidelity of SNA immensely because of the inherently cellular and dynamic nature of clandestine networks. When this assumption is relaxed, the analysis focuses on how influences and interdictions of a network will affect its structure, specifically the connections between actors. Carley identifies events like death, promotion, recruitment, innovation, goal changes, and acquisitions as reasons for a change in the networks, but these are by no means all-encompassing [55].

Jackson and Wolinsky [52] also attempt to describe possible changes based on the utility of the possible connections, which is described in Section 2.8. Actors within a network will create connections based on a mutual benefit, or a "pairwise stability", between them [52]. It is important to note that when a network attempts to become more stable, it does not necessarily reach the most stable structure, but will reach an overall higher utility structure based on having pairwise stability.

The dynamic nature of terrorist networks adds another dimension to node insertion analysis, specifically the time aspect. However, it also allows for more precise recommendations. If the dynamics of networks are taken into account, and predicted with a relatively high amount of certainty, a node insertion can be recommended not only upon an actor, with a certain relationship, but also at a specific time, to take advantage of a weakened or vulnerable state within the network.

While the proactive node insertion is important, also crucial is the predicted effect an insertion will have on a network following the implementation. If relationships or the network functionality changes drastically because of an insertion, it is likely that course of action could be detrimental to government and/or military interests as well as the covert operative to be inserted.

2.11 Literature Review Summary

This chapter outlined previous works crucial to this research. This included network type, centrality measures, network functionality, community analysis, and introductions to clandestine and dynamic networks.

III. Methodology

3.1 Overview

The following methodology describes the application of the various Social Network Analysis (SNA) measures described in Chapter 2, ultimately to deliver recommendations for future node insertion operations. Using a weighted additive model approach (WAM), these metrics are combined, organized by benefit and risk, and weighted based on decision-maker (DM) input. This is an extension of the Jackson-Wolinsky model for node utility [52], introduced in Section 2.8.

Section 3.2 outlines the notation used for the networks used in addition to the measures applied. Section 3.3 extends the work of Qi *et al.* with Laplacian centrality to the node insertion problem. Section 3.4 outlines a specific model for node insertion while Section 3.5 applies this model to an small overt network. Section 3.6 then outlines an extension of the specific model from Section 3.4 to a general form for the node insertion problem. Section 3.7 then applies the specific methodology to the Zachary Karate Club Network, a well known and well used network in SNA.

Figure 12 shows the process flow for this analysis; this chapter explains the family of measures portion of the chart.

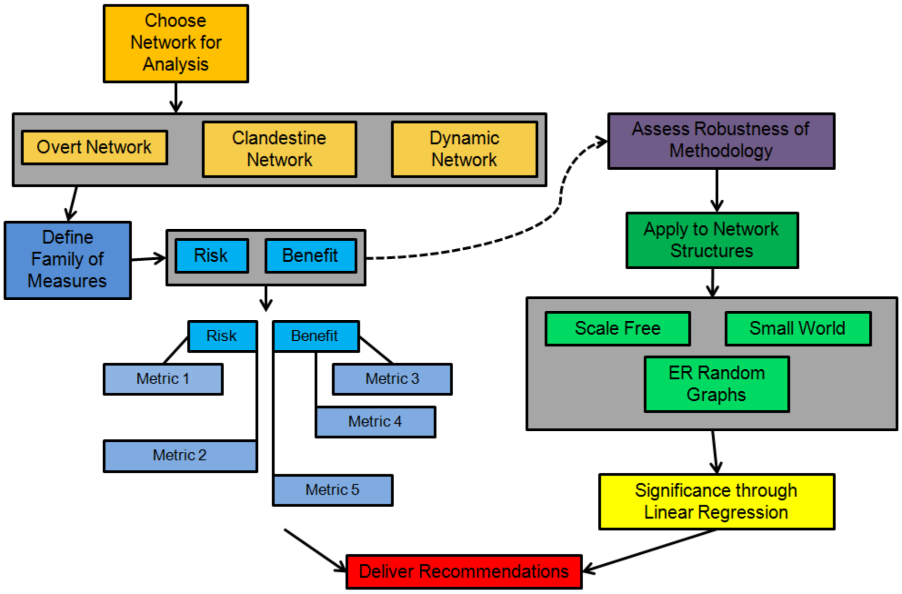


Figure 12. Process Flow

This methodology begins with identifying a single network for analysis, which includes real-world social networks. The next level shows the definition of both risk and benefit, which are decomposed into a collection of local and global centrality criteria. The risks and benefits used can differ within each application of node insertion, specifically to provide a higher recommendation fidelity based on the node insertion purposes.

Following the identification of risks, benefits, and weights for each measure, the WAM is applied to a social network of interest with various characteristics, particularly network structure. Results from this are collected in an attempt to understand the effects of different characteristics of the network on the optimal solution to the node insertion problem. The primary focus of the research is on the analysis of overt networks, both scale free and small world. These structure types align to networks found most commonly in SNA. The final step in the analysis deals with sensitivity parameters such as the inherent cost of adding an arc to an inserted node, and subject-matter expert (SME) node weighting. The pruned process flow specific to this analysis is shown in Figure 18.

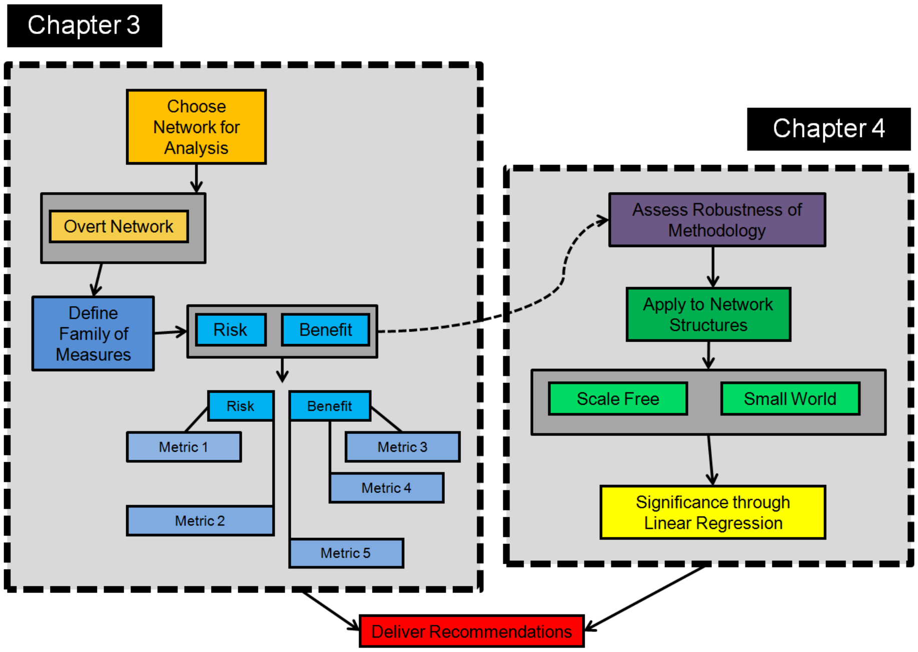

Chapter 3
Choose Network for Analysis
Overt Network
Define Family of Measures
Risk
Benefit
Risk
Benefit
Metric 1
Metric 2
Metric 3
Metric 4
Metric 5
Chapter 4
Assess Robustness of Methodology
Apply to Network Structures
Scale Free
Small World
Significance through Linear Regression
Deliver Recommendations

Figure 13. Pruned Process Flow

The final step is to deliver node insertion recommendations for the network; a portfolio of options for a DM to consider.

3.2 Notation

Each graph of interest is defined by $G = (V, E)$. For the purposes of this research, the terms "node" and "vertex" are used interchangeably. The same occurs with the terms "edge" and "arc". V is the set of all nodes in G, $V = \{v_1, v_2, ..., v_n\}$, where v_i represents node i. E is the set of all arcs in G, $E = \{a_{12}, a_{23}, ..., a_{ij}\}$. a_{ij} is the arc that connects nodes v_i and v_j such that $a_{ij} \in E$.

Node insertion scenario j, which includes the set of nodes involved with insertion j, is defined as x_j such that $x_j \subseteq V$. v_{x_j} is the inserted node for scenario x_j. Any arc involved in an insertion scenario is designated as a_{xj}, where a relationship exists between the inserted node v_{x_j} and v_j. $G^*_{x_j} = (V^*_{x_j}, E^*_{x_j})$ where $V^*_{x_j} = V \cup v_{x_j}$ and $E^*_{x_j} = E \cup E_{x_j}$, where E_{x_j} is the set of arcs associated with scenario x_j and $E_{x_j} \neq \emptyset$. With this, there must be at least one relationship for the inserted node v_{x_j}.

3.3 Laplacian Energy Applied to Node Insertion

In Section 2.5 Laplacian centrality is defined as the drop in energy that results when a node is "deactivated" within network G [45]. In order to capture this drop in energy Qi *et al.* [45] uses Equation 20, which is extended to the node insertion problem and the creation of a graph $G^*_{x_j}$. If a graph $G^*_{x_j}$ is created, where $G^*_{x_j}$ is the graph which includes a node insertion scenario x_j and $G \subset G^*_{x_j}$, then it can be shown that the $E_L(G) < E_L(G^*_{x_j})$ by applying the proof in Lazić [47]. The change in Laplacian energy then is the Laplacian centrality of v_{x_j} in $G^*_{x_j}$, defined as L_{x_j}. With this, the work of Qi *et al.* [45] can be extended to the building of networks and the activation of nodes, not just node removal, while still using the original equation provided. This allows for the elicitation of the importance of a node directly from the network itself, based on an intermediate network measure.

An example graph $G^*_{x_j}$, created from the network G shown in Figure 2, is shown in Figure 14 where v_{x_j} is the inserted node associated with scenario x_j.

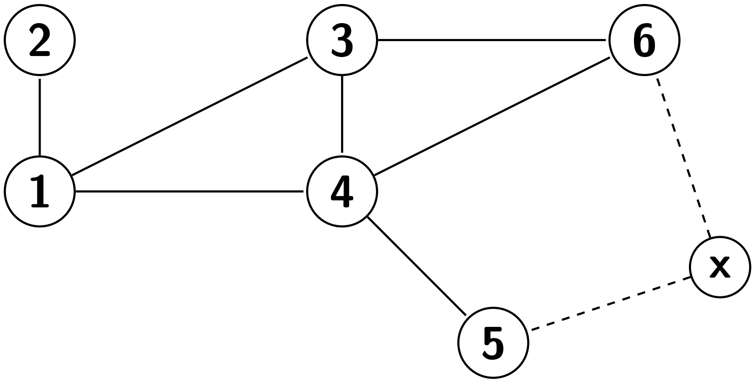


Figure 14. Example G^*

For the inserted node in Figure 14 the degree is two, with the sum of the degrees of its neighbors equaling five. With this, the energy generated from adding a node with connections to v_5 and v_6 results in $l_{x_j} = 16$, calculated using Equation 20. This follows because $E_L(G) = 54$ and $E_L(G^*_{x_j}) = 70$.

3.4 Specific Node Insertion Model

For this analysis, there are two possible node insertion purposes, which are intelligence collection and future network disruption. In order to complete the different objectives demanded by these purposes the benefits and risks associated with each are inherently different. Table 11 shows the different purposes, along with associated benefits and risks. The focus of this research is information gathering.

Table 11. Node Insertion Purposes

		Risk	Benefit
Purpose	Information Gathering	Min Closeness $v_i \in x_j$	Max Laplacian Centrality v_{x_j}
		Min ΔM	Max Degree $v_i \in x_j$
		Min Inherent Costs	
	Future Network Disruption	Min Laplacian Centrality v_i	Max Closeness v_{x_j}
		Min Degree $v_i \in x_j$	Max ΔM
		Min Inherent Costs	

Using the Laplacian centrality applied in Section 3.3 the specific WAM for this analysis can be formulated. This is shown in Equation 22 and corresponds to a focus on information gathering.

$$U_{x_j} = w_B B_{x_j} - w_R R_{x_j} \quad \forall \quad j \tag{22a}$$

$$\phi_i = w_s d_i + (1 - w_s) p_i \tag{22b}$$

$$B_{x_j} = w_1 L_{x_j} + w_2 \sum_{i \in x_j}^{n} \phi_i \quad \forall \quad j \tag{22c}$$

$$R_{x_j} = w_3 m_{x_j} + \sum_{i \in x_j}^{n} w_4 C_i + w_5 z_{x_j} \quad \forall \quad j \tag{22d}$$

U_{x_j} then is the overall utility for x_j. B_{x_j} is the benefit value for x_j, while R_{x_j} is the risk value for x_j. ϕ_i is the individual node bias which is made up the degree d_i for v_i and the SME rating, p_i, for v_i, which are both normalized. The weight w_s defines the importance of both of these measures. L_{x_j}, the change in Laplacian energy, is as defined previously with w_1 as its associated weight, and w_2 as the weight of the summed individual node biases. Different weighting corresponds to both differing measure emphasis, levels of risk aversion, and competing node insertion purposes (intelligence collection or future network disruption). m_{x_j} is defined as the absolute change in assortative mixing occurring between G and $G^*_{x_j}$, the calculation for which is shown with Equation 24.

$M(G)$ is the assortative mixing by degree for a graph G while $M(G^*_{x_j})$ is the assortative mixing by degree for $G^*_{x_j}$, previously described in Section 2.3. C_i is this closeness of v_i with w_3 and w_4 are defined as the weights for m_{x_j} and C_i respectively. z_{x_j} is the inherent cost of creating x_j and is shown in Equation 23 where c is some fixed cost parameter and $|E_{x_j}|$ is the cardinality of the edges added in scenario x_j.

$$z_{x_j} = |E_{x_j}| c \tag{23}$$

$$m_{x_j} = |M(G^*_{x_j}) - M(G)| \tag{24}$$

Each of the values calculated for each x_j will be normalized using Equation 25.

$$\text{Normalized } s_i = \frac{s_i - min(S)}{max(S) - min(S)} \tag{25}$$

S is the set of all scores s_i. This normalization technique is used because we are only concerned with the relative risks and benefits between each insertion scenario, not an actual cost or profit for each of these measures.

To keep a two-level normalization within the network model, the only restriction on the weights is that they sum to one within each level. The first level of normalization occurs with the Laplacian centrality for each scenario x_j. The second level of normalization occurs with total benefit, which includes the normalized Laplacian centrality and the summation of the degree and SME weighting. On the risk side, the first level of normalization occurs with the closeness values for graph G and the change in assortative mixing for each insertion scenario. The second level includes the normalization of the summation of normalized closeness, assortative mixing, and non-normalized inherent costs. With this, the weights on the degree and SME input should sum to one. The weights on the individual node bias and Laplacian centrality, w_1 and w_2 respectfully, should also sum to one. The same should occur with weights w_3, w_4, and w_5.

With each of these benefits and risks, the specific model created includes a local measure (degree), intermediate measure (Laplacian centrality), global measure (closeness), and a full network measure (assortative mixing), which allows for the true structure of the network to be captured within each dimension. The utility of each node insertion strategy for this model is then described with Equation 26.

$$U_{x_j} = \overbrace{\underbrace{\text{degree}}_{\text{local}} + \underbrace{\text{Laplacian}}_{\text{intermediate}}}^{\text{benefit}} - \overbrace{\underbrace{\text{closeness}}_{\text{global}} - \underbrace{\Delta\text{assortivity}}_{\text{full network}}}^{\text{risk}} - \text{cost} \tag{26}$$

Degree and Laplacian centrality were used as benefits because of the focus on information collection within this analysis. A number of contacts and high importance of indirect contacts directly affects the amount and strength of information that could flow through an inserted node. Laplacian centrality also accounts for communities which might exist within a network. With that being said, residing along the quickest lines of communication could make an inserted operative vulnerable, which increases risk when inserting. For this reason, closeness centrality was designated as a risk measure. Assortivity was also used as risk because it captures the types of connections made in a network. If an operative attempts to make irregular connections within a network, it could raise suspicion from within the network and increase overall risk.

3.5 Methodology Applied To Example Network

In order to thoroughly explain the proposed methodology, the example network previously shown in Figure 2 is used as the network of interest. The different steps in methodology will be implemented to this network, ultimately resulting in the identification of a portfolio of possible node insertion scenarios.

The analysis begins with the elicitation of weights for each v_i. As defined previously, the weights for each node are made up of structural characteristics and SME input. The characteristic of focus for this analysis is the degree of v_i. The total weights for each v_i are a weighted combination of the normalized degrees and normalized SME weights. These values, along with the SME weights, are shown in Table 12. For this example, degree and SME input are equally weighted.

Table 12. Individual Node Weights

Node	Degree	Normalized Degree	SME Weight	Normalized SME	Total Bias
1	3	0.667	0.1190	0.0000	0.1772
2	1	0.000	0.4984	0.4513	0.0000
3	3	0.667	0.9597	1.0000	1.0000
4	4	1.000	0.3404	0.2634	0.6682
5	1	0.000	0.5853	0.5547	0.0850
6	2	0.333	0.2238	0.1247	0.0055

Each x_j will receive an additional weight in benefit given which nodes make up the insertion scenario. For example, x_3 would receive an additional benefit of 1.0 because the inserted node forms a relationship with v_3, while x_2 would receive no additional benefit because the inserted node forms a relationship with v_2 which has no individual bias.

An assumption exists on the actual insertion of nodes. Operationally, it would be unrealistic to attempt to create relationships with every node within a large network. The risk involved in creating this large number of relationships and the risk involved with remaining inconspicuous while maintaining them is extremely large. Knowing this, there is a limit on the number of relationships that can be created for an inserted node. The assumption then is that the maximum number of relationships that can be added is a function of the total number of nodes within the network.

For this methodology, the function chosen to limit the number of relationships possible for each x_j is the average degree of graph G. If more than this number of relationships is added, the inserted node could attract a large amount of suspicion, which would dramatically increase the overall risk . This functions also aids in limiting the design space to a tractable level.

Every possible insertion scenario is combined into the matrix **X**, called the edge addition matrix. For the original graph G, in Figure 2 **X** is shown in Equation 27 for the addition of up to two arcs, resulting in twenty-one possible scenarios.

$$\mathbf{X} = \begin{array}{rcccccc} & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 1 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 1 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \\ 7 & 1 & 1 & 0 & 0 & 0 & 0 \\ 8 & 1 & 0 & 1 & 0 & 0 & 0 \\ 9 & 1 & 0 & 0 & 1 & 0 & 0 \\ 10 & 1 & 0 & 0 & 0 & 1 & 0 \\ 11 & 1 & 0 & 0 & 0 & 0 & 1 \\ 12 & 0 & 1 & 1 & 0 & 0 & 0 \\ 13 & 0 & 1 & 0 & 1 & 0 & 0 \\ 14 & 0 & 1 & 0 & 0 & 1 & 0 \\ 15 & 0 & 1 & 0 & 0 & 0 & 1 \\ 16 & 0 & 0 & 1 & 1 & 0 & 0 \\ 17 & 0 & 0 & 1 & 0 & 1 & 0 \\ 18 & 0 & 0 & 1 & 0 & 0 & 1 \\ 19 & 0 & 0 & 0 & 1 & 1 & 0 \\ 20 & 0 & 0 & 0 & 1 & 0 & 1 \\ 21 & 0 & 0 & 0 & 0 & 1 & 1 \end{array} \tag{27}$$

Each column in the matrix **X** represents a node in G, and each row represents a different x_j. The presence of a one in a specific row is the presence of a_{xi} in $G^*_{x_j}$. Using **X**, a $G^*_{x_j}$ is then created for each x_j, and the Laplacian centrality of v_{x_j} in each $G^*_{x_j}$ is calculated using Equation 20.

The Laplacian centralities for each v_{x_j} are shown in Table 13, along with their normalized values, calculated using Equation 25.

Table 13. Laplacian Centralities of v_{x_j}

x_j	Laplacian	Normalized
1	10	0.22
2	6	0.00
3	10	0.22
4	12	0.33
5	6	0.00
6	8	0.11
7	18	0.67
8	22	0.89
9	**24**	**1.00**
10	18	0.67
11	20	0.78
12	18	0.67
13	20	0.78
14	14	0.44
15	16	0.56
16	**24**	**1.00**
17	18	0.67
18	20	0.78
19	20	0.78
20	22	0.89
21	16	0.56

Scenarios x_9 and x_{16} exhibit the highest Laplacian centralities. This corresponds to relationships with v_1 and v_4 for scenario x_9 and v_3 and v_4 for scenario x_{16}. It is also important to note that scenarios x_2 and x_5, which include relationships with v_2 and v_5 respectively, provide the lowest benefit.

Of note in these results is the Laplacian centrality for scenarios x_9 and x_{16} exceeds the Laplacian centrality of the sum of their parts. For example, scenario x_{16} includes relationships with v_3 and v_4, which includes a_{x3} and a_{x4}, for a Laplacian centrality of 24. If two nodes are inserted, one with a single relationship with v_3 and the other with a single relationship to v_4, their combined Laplacian centrality is 22. This non-linear increase will always occur with Laplacian centrality due to the d_i^2 term in the Laplacian centrality calculation, shown in Equation 20. With this, $L_{x_j} > \sum L_i \quad \forall i \in x_j$. Figure 15 provides the network diagram for this result. A similar relationship is found with closeness. As multiple relationships are created with an inserted node, the possibility of creating new shortest paths also exists. This creates a relationship in which $C_{x_j} \geq \sum C_i \quad \forall i \in x_j$. It is not a strict inequality because the addition of multiple arcs does not necessarily change the shortest path.

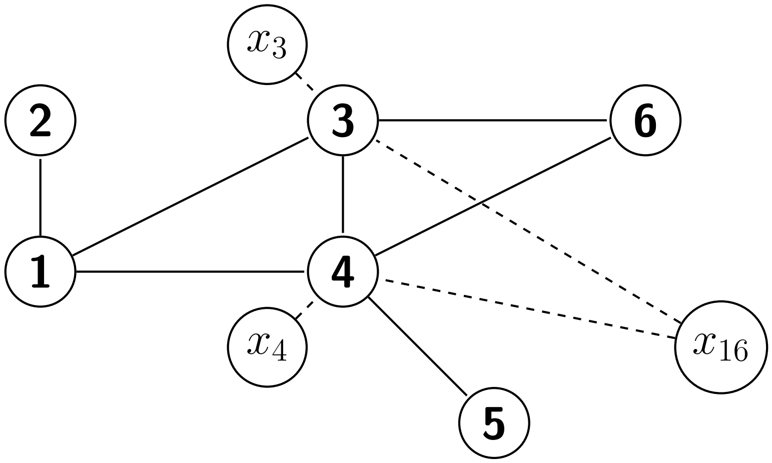


Figure 15. Laplacian Insertion Property

To determine the risk of an insertion scenario, we first define an inherent cost of creating a relationship, z_{x_j}. This is similar to the c_{ij} identified by Jackson and Wolinsky [52], and originally described in Section 2.8, but only captures the action of creating a_{xi} and not the risk associated with that node based on the structure of the network. This is accomplished using Equation 23 previously shown in Section 3.4.

For the example network an inherent cost of 0.2 is used for each added relationship. Changing the inherent cost for the addition of an arc would only shift the highest utility insertion strategies to those involving a higher or lower number of arcs depending on whether or not the change decreases or increases respectively. The inherent cost is not normalized because it remains on the same scale as the other measures, unlike closeness and betweenness which lie on different scales. It is important to remember that the inherent costs can exceed one depending on the number of relationships added for a particular x_j.

The second step of the risk formulation is the calculation of closeness centrality for every node in G. Since this calculation involves the identification of all shortest paths it provides huge computational load to the analysis, especially with large networks. However, the risk is determined prior to insertion, so only the closeness values of the nodes in G need to be calculated, not the closeness of each node in every $G^*_{x_j}$. Equation 6a is used to calculate the closeness, which assumes a connected graph. For the example network the closeness values, originally shown in Table 2, are reproduced in Table 14 along with the normalized values, calculated using Equation 25.

Table 14. Example Network Closeness Centralities

Node	Closeness	Normalized
1	0.143	0.686
2	0.091	0.000
3	0.143	0.686
4	**0.167**	**1.000**
5	0.100	0.120
6	0.111	0.267

v_4 has the highest closeness centrality but is closely followed by v_1 and v_3. These also happen to be the nodes that result in the largest Laplacian centrality with scenarios x_9 and x_{16}. These costs are added to the total risk for insertion scenario x_j based on the set of nodes involved in x_j.

The third step in determining cost is to find the change in the assortivity coefficient occurring between G and $G^*_{x_j}$ calculated using Equations 11 and 24, which calculates degree assortivity. Given this differentiation, the groups for each node are purely made up by the degrees of each of the nodes. Following the creation of $G^*_{x_j}$, the group that v_{x_j} falls into is also determined. The absolute change in assortative mixing from G to each $G^*_{x_j}$ is shown in Table 15, and defined as η_{x_j}.

Table 15. Δ Assortative Mixing from G to $G^*_{x_j}$

x_j	m_{x_j}	Normalized
1	0.087	0.191
2	0.258	0.584
3	0.004	0.000
4	0.142	0.316
5	**0.392**	**0.890**
6	0.188	0.424
7	0.125	0.278
8	0.013	0.021
9	0.130	0.289
10	0.125	0.278
11	0.004	0.001
12	0.125	0.278
13	0.019	0.036
14	**0.378**	**0.860**
15	0.270	0.610
16	0.071	0.154
17	0.236	0.533
18	0.221	0.498
19	0.115	0.255
20	0.016	0.027
21	**0.439**	**1.000**

Scenarios x_{21}, x_5, and x_{14} and have the highest change in assortative mixing, which indicates the highest possible risk values for this measure.

In order to remove the possibility of inherent weighting between risks and benefits just based on the scale of values, the benefits and risks are normalized using unity based normalization from Equation 25. The total benefit for each is the summation of each x_j's Laplacian centrality value and each individual node weight. The total risk is the summation of inherent cost for x_j, the closeness for each node that x_j involves, and the change in assortative mixing that occurs with x_j. The normalized values are shown in Table 16.

Table 16. Benefits and Risks Associated With v_{x_j}

x_j	Benefit	Risk
1	0.164	0.179
2	0.000	0.153
3	0.411	**0.000**
4	0.367	0.444
5	0.025	0.488
6	0.057	0.200
7	0.386	0.448
8	0.797	0.601
9	0.753	**1.000**
10	0.412	0.497
11	0.444	0.385
12	0.633	0.448
13	0.589	0.368
14	0.248	0.647
15	0.279	0.561
16	**1.000**	0.874
17	0.659	0.736
18	0.690	0.851
19	0.615	0.623
20	0.646	0.557
21	0.305	0.975

The highest benefit comes from scenario x_{16}, which includes v_3 and v_4. The lowest risk comes from scenario x_3, which also has a moderate benefit of 0.411. The highest risk comes from scenario x_9, with a total benefit of 0.753.

In order to calculate the utility for each x_j, Equation 28a is used, where $w_B = w_R = 0.5$. These utilities are shown in Table 17.

Table 17. Utility For Each x_j

x_j	Utility
1	-0.007
2	-0.076
3	**0.205**
4	-0.039
5	-0.231
6	-0.071
7	-0.031
8	0.098
9	-0.123
10	-0.042
11	0.029
12	0.093
13	0.110
14	-0.200
15	-0.141
16	0.063
17	-0.038
18	-0.080
19	-0.004
20	0.044
21	-0.335

The optimal scenario is x_3 which includes only v_3 with a total utility of 0.205. In total there are seven scenarios where the benefits outweigh the risk, resulting in a positive utility for the scenario. These are shown in Table 18, which describes a portfolio of the best node insertion scenarios, along with the nodes involved in each insertion scenario.

Table 18. Example Network Node Insertion Portfolio

x_j	Nodes	Utility
3	v_3	0.205
13	v_2, v_4	0.110
8	v_1, v_3	0.098
12	v_2, v_3	0.093
16	v_3, v_4	0.063
20	v_4, v_6	0.044
11	v_1, v_6	0.029

A relationship with v_3 is prominent in the top node insertion strategies, but a relationship with v_2 appears in multiple strong strategies as well. This portfolio suggests a number of significantly different strategies to prepare for decision-maker input. The third, and fourth strategies in the portfolio also seem to have similar utility values suggesting they are essentially interchangeable, which allows for flexibility in the final decision.

3.6 Generalized Model

While a specific model was used for this analysis, which included degree, Laplacian, and closeness centralities, in addition to assortative mixing, a generalized methodology was also developed. Equation 28 shows the generalized form of the utility function identified in Equation 22, which is an extension to Jackson and Wolinsky's model [52], originally shown in Equation 21.

$$U_{x_j} = w_B B_{x_j} - w_R R_{x_j} \quad \forall \quad j \tag{28a}$$

$$\phi_i = \sum_{s}^{S} w_s b_s^i \quad \forall \quad i \tag{28b}$$

$$B_{x_j} = w_b \sum_{i \in x_j}^{n} \phi_i + w_c \sum_{k}^{K} w_k \gamma_k^{x_j} \quad \forall \quad j \tag{28c}$$

$$R_{x_j} = \sum_{l}^{L} w_l \eta_l^{x_j} \quad \forall \quad j \tag{28d}$$

The definitions of each of these variables correspond to the original definitions outlined in the specific WAM shown in Section 3.4, but are not limited to degree, Laplacian centrality, closeness, or assortative mixing but can follow any function that the decision maker designates as a risk or a benefit. Individual benefit and risk values are calculated to allow for assessment of trade-offs within the decision maker's (DM) decision space. The overall utility is used for the direct comparison of node insertion scenarios and ultimately allows for the creation of a prioritized list of these scenarios.

3.7 Methodology Applied to Zachary's Karate Club Network

One well known social network comes from Zachary 1977 [56] and is a representation of the relationships between members of a dissolved karate club, simply called Zachary's Karate Club Network (ZKCN). Originally collected between 1970 and 1972 for the purpose of studying "fission" in small networks from a anthropology perspective, this network has become widely used in social network research [56]. Using the specific WAM outlined in Section 3.4 a portfolio of node insertion strategies for this network will be identified and presented to further demonstrate the methodology.

The parameters and weighting for the node insertion analysis of the ZKCN remain the same as for the analysis of the example network, with the exception of the individual node bias. A random SME weighting scheme was used to create the individual node bias within the network and is shown in Table 19. The total individual node bias with the inclusion of the normalized degree of each node is shown in Table 20.

Table 19. ZKCN SME Weighting

Node	Weight	Normalized	Node	Weight	Normalized
1	0.148	0.149	18	0.228	0.230
2	0.055	0.055	19	0.498	0.503
3	0.851	0.859	20	0.901	0.910
4	0.561	0.566	21	0.575	0.580
5	0.930	0.939	22	0.845	0.854
6	0.697	0.704	23	0.739	0.746
7	0.583	0.588	24	0.586	0.592
8	0.815	0.824	25	0.247	0.249
9	0.879	0.888	26	0.666	0.673
10	0.989	0.999	27	0.083	0.084
11	0.001	0.000	28	0.626	0.632
12	0.865	0.874	29	0.661	0.667
13	0.613	0.619	30	0.730	0.737
14	0.990	1.000	31	0.891	0.900
15	0.528	0.533	32	0.982	0.992
16	0.480	0.484	33	0.769	0.777
17	0.801	0.809	34	0.581	0.587

Table 20. Total Individual Node Bias for ZKCN

Node	Degree	Normalized	Node Bias	Normalized	Node	Degree	Normalized	Node Bias	Normalized
1	16	0.938	0.543	0.657	18	2	0.063	0.146	0.114
2	9	0.500	0.278	0.294	19	2	0.063	0.283	0.301
3	10	0.563	0.711	0.887	20	3	0.125	0.517	0.622
4	6	0.313	0.439	0.515	21	2	0.063	0.321	0.354
5	3	0.125	0.532	0.642	22	2	0.063	0.458	0.541
6	4	0.188	0.446	0.524	23	2	0.063	0.404	0.467
7	4	0.188	0.388	0.445	24	5	0.250	0.421	0.490
8	4	0.188	0.506	0.606	25	3	0.125	0.187	0.170
9	5	0.250	0.569	0.693	26	3	0.125	0.399	0.460
10	2	0.063	0.531	0.640	27	2	0.063	0.073	0.015
11	3	0.125	0.063	0.000	28	4	0.188	0.410	0.475
12	1	0.000	0.437	0.512	29	3	0.125	0.396	0.457
13	2	0.063	0.341	0.380	30	4	0.188	0.462	0.547
14	5	0.250	0.625	0.769	31	4	0.188	0.544	0.658
15	2	0.063	0.298	0.322	32	6	0.313	0.652	0.807
16	2	0.063	0.273	0.288	33	12	0.688	0.732	0.916
17	2	0.063	0.436	0.511	34	17	1.000	0.794	1.000

When dealing with a small network, the evaluation of each possible node insertion scenario is tractable. For a thirty node network, allowing up to three relationships to be added, the number of combinations is 4,525, the calculation for which is shown in Equation 29.

$$\binom{30}{1} + \binom{30}{2} + \binom{30}{3} = 30 + 435 + 4060 = 4525 \tag{29}$$

As we allow for the inclusion of up to n relationships for an inserted node, the possibilities grow to over 1 billion possible relationship combinations. For a 1,000 node network, the number of scenarios grows to nearly this magnitude after only three possible relationships are allowed, and passes this threshold with the possibility of four relationships. For this reason, it is not feasible to evaluate every potential node scenario to gain an optimal solution. Instead, for the sake of having a tractable decision space, the guarantee of optimality is sacrificed. To eliminate the possibility of not finding an optimal solution in tractable time, we limit the number of relationships within each x_j in order to guarantee an optimal solution over each scenario investigated.

The average degree for this network is 4.588, so the scope of this specific instance includes every contingency up to the creation of four relationships. With this specification there are 52,955 node insertion scenarios to evaluate.

The top node insertion scenario for this network is x_{5899}, which includes a relationship with three nodes: v_{17}, v_{33} and v_{34}. The total utility is 0.115, which is 0.227 greater than the expected value of a random insertion, -0.1125, corresponding to a 201.9% increase.

The portfolio for the top ten node insertion candidates is shown in Table 21.

Table 21. ZKCN Insertion Portfolio

x_j	Nodes	Utility	% Δ Utility
5899	v_{17}, v_{33}, v_{34}	0.1147	201.9%
31800	$v_6, v_{17}, v_{33}, v_{34}$	0.1133	200.7%
20810	$v_3, v_{17}, v_{33}, v_{34}$	0.1128	200.3%
20350	$v_3, v_{14}, v_{33}, v_{34}$	0.1125	200.0%
7404	v_1, v_3, v_{17}, v_{34}	0.1107	198.4%
50519	$v_{17}, v_{26}, v_{33}, v_{34}$	0.1097	197.5%
5439	v_{14}, v_{33}, v_{34}	0.1094	197.2%
34725	$v_7, v_{17}, v_{33}, v_{34}$	0.1077	195.7%
21489	$v_3, v_{31}, v_{33}, v_{34}$	0.1076	195.6%
987	v_1, v_{17}, v_{34}	0.1071	195.2%

v_{33} and v_{34} appear in each of the top ten node insertion scenarios except for x_{7404}, which only includes v_{33}. v_{17} makes six appearances. It is noteworthy that the second, third, and fourth scenarios in the portfolio, x_{20350}, x_{20180}, and x_{31800}, each have essentially the same utility. This allows for decision maker flexibility given that these three scenarios include different nodes. A portfolio including the top candidates for each number of possible relationships, up to four, is shown in Table 22.

Table 22. ZKCN Insertion Portfolio by Number of Relationships

# Relationships	x_i	Nodes	Utility	% Δ Utility
1	3	v_3	0.0802	171.3%
2	217	v_6, v_{34}	0.0946	184.1%
3	5899	v_{17}, v_{33}, v_{34}	0.1147	201.9%
4	31800	v_6, v_{17}, v_{33}, v_{34}	0.1133	200.7%

According to Table 22 the utility of the top insertion candidates increased greatly from the addition of only two arcs to three or four arcs added, but not a large change occurred between three and four relationships for the insertion scenario.

The ratio between risk and benefit also gives insight for each of these scenarios. These risk ratios, or the quotient between risk and benefit, for each of the top ten insertion scenarios, are shown in Table 23.

Table 23. Risk Ratios for ZKCN Portfolio

x_j	Ratio (R_{x_j}/B_{x_j})
5899	0.6334
987	0.6544
5439	0.6812
31800	0.7061
50519	0.7089
34725	0.7166
20810	0.7405
7404	0.7437
20350	0.7580
21489	0.7625

When comparing risk ratios, values above one mean risks outweigh benefits. Alternatively, values below one mean benefits outweigh risks. Each of these scenarios has a risk ratio below one, which corresponds to the results shown previously. The optimal solution, scenario x_{5899}, also has the lowest risk ratio, with value of 0.6334. The risk for this insertion is 0.6334 times as large as the benefit. Equivalently, the benefit is 1.577 times as large as the risk.

When comparing Tables 21 and 23 differences do exits. The tenth solution overall, scenario x_{987}, has the second lowest risk ratio, while scenario x_{21489} has the highest risk ratio of the top ten scenarios.

It is also interesting to note that the top two insertion scenarios in terms of risk ratio, are both insertions that only require three inserted relationships. The risk ratio associated with the expected risk and expected benefit of a random insertion is 1.462, suggesting that the expected risk of a random insertion is 1.462 times as large as the expected benefit of a random insertion.

For the top insertion scenario, x_{5899}, both the Laplacian and closeness centralities, along with their overall ranks, are shown in Table 24.

Table 24. ZKCN Benefit and Cost Comparison for x_{5899}

Node	Closeness	Rank	Laplacian	Rank
17	0.3265	34	6	28
33	0.6152	4	156	3
34	0.6838	1	306	1

Figure 16 shows the plots of each node insertion scenario and their associated utility. The average utility for the network is shown by the red line. The black line is representative of zero utility. Because within this analysis benefit and utility were equally weighted, any point above this line corresponds to a scenario with more benefit than risk. It is important to note that there were a significantly higher amount of risky scenarios than there were beneficial ones. Figure 16 also shows that multiple scenarios separate themselves significantly from the group. These higher utility scenarios are addressed in Table 21. The pattern in Figure 16 exists due to the sequencing of MATLAB's *nchoosek* function.

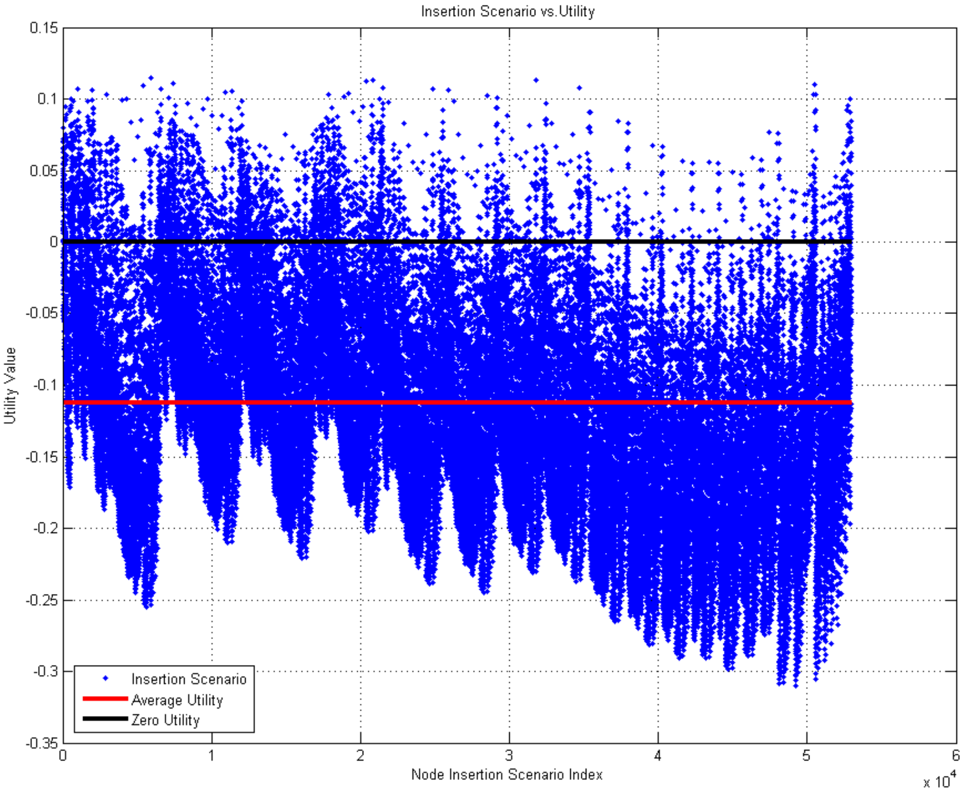


Figure 16. ZKCN Utility Plot

Since risks and benefits were present in the analysis, it was important to make sure these measures were uncorrelated. Figure 17 shows a Risk vs. Benefit plot for the ZKCN utility values.

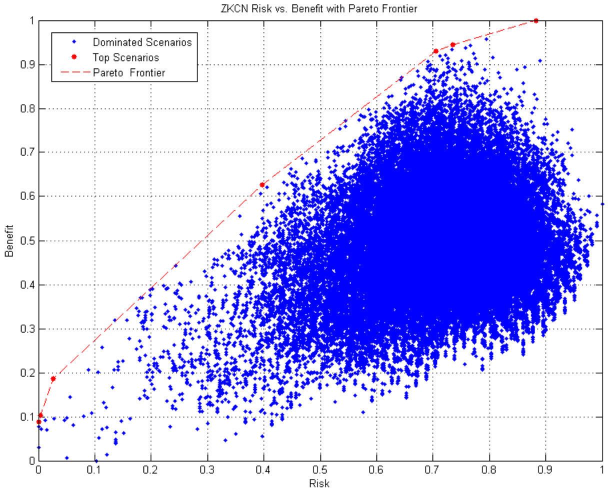


Figure 17. ZKCN Risk vs. Benefit Plot with Pareto Frontier

As shown in Figure 17, there seems to be a slight upward trend in the data, suggesting some slight correlation exists between the risk and benefit. Using the benefit values as a response and the risk values as the independent variable within a linear regression yielded a significant model. The R^2 value was 0.0564 suggesting that just over 5% of the variation within the risk values was captured by the model. While risk was significant in determining benefit, the fit to the response was lacking.

The correlation between the measures was 0.24, which is just below 0.30, a threshold considered significant in most social sciences. This significant relationship was also dependent on the inherent costs of adding additional arcs because the Laplacian centrality of an inserted node increases with the addition of more arcs, just as the inherent costs increases with the addition of arcs. To control this, the risks and benefit values were broken up by number of relationships. With this control added, the results differed. Table 25 shows these correlation values.

Table 25. Benefit and Risk Correlation by # of Inserted Relationships

Relationship #	Correlation
1	-0.0566
2	-0.352
3	-0.264
4	-0.0637

This analysis initially applied the methodology to the ZKCN with equal weighting benefit and risk. In order to assess the sensitivity of these weights, different weighting schemes were also used. These different weighting schemes allowed for the Pareto frontier to be established, which is also shown in Figure 17. Any red point in Figure 17 is a dominating insertion scenario, or one that is optimal for at least one weighting scheme, which involves changing only the weights on benefit and risk, w_B and w_R respectively. Both the scenarios associated with these points and their respective weighting are shown in Table 26.

Table 26. w_B and w_R Sensitivity Analysis Results

x_j	w_B	w_R
6	0.00	1.00
6	0.05	0.95
6	0.10	0.90
6	0.15	0.85
4	0.20	0.80
3	0.25	0.75
3	0.30	0.70
3	0.35	0.65
3	0.40	0.60
3	0.45	0.55
5899	0.50	0.50
20350	0.55	0.45
20350	0.60	0.40
20350	0.65	0.35
10895	0.70	0.30
7540	0.75	0.25
7540	0.80	0.20
7540	0.85	0.15
7540	0.90	0.10
7540	0.95	0.05
7540	1.00	0.00

Table 26 shows that certain solutions dominate for a large range of weighting schemes, suggesting more stable solutions. The optimal solution for equal weights appeared only once, whereas scenario 7540 dominated from a benefit weight of 0.75 to a benefit weight of 1.00. It is also interesting to note that for the equally weighted solution, three relationships were added. For all other dominating scenarios, the optimal solution included either the lowest number of possible relationships, or the highest number of possible relationships.

3.8 Methodology Summary

This chapter focuses on the explanation of both the specific and general utility models for node insertion. This includes the extension of Laplacian centrality to node activation. The specific formulation was explained through application to an example network. A secondary application to the ZKCN resulted in an optimal insertion strategy and a portfolio of other high utility strategies for the decision-maker stage. A Pareto frontier was generated to assess the sensitivity of the weights on the risk and benefit portions of the methodology. The methodology provides marked improvement in insertion scenario utility, from an information collection perspective, over a random insertion.

IV. Analysis

4.1 Introduction

In this chapter, the node insertion methodology presented in Chapter 3 is applied to multiple networks of differing type, size, and structural characteristics. Section 4.2 outlines the experimentation plan in addition to describing the graph generation process. Section 4.3 presents the results of the experimentation in addition to providing robustness analysis. This process is reiterated in Figure 18.

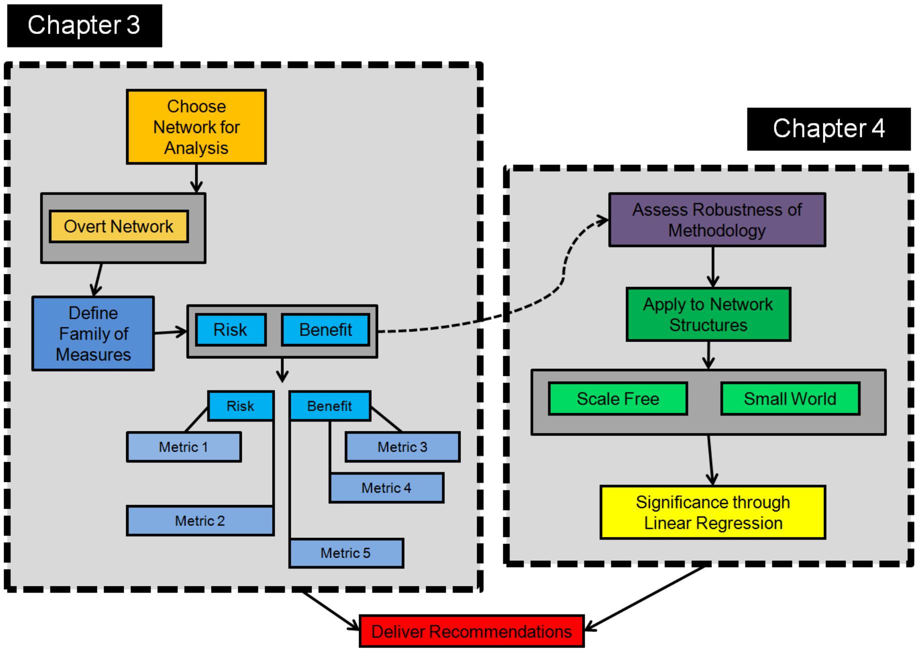

Chapter 3
Choose Network for Analysis
Overt Network
Define Family of Measures
Risk
Benefit
Risk
Benefit
Metric 1
Metric 2
Metric 3
Metric 4
Metric 5
Chapter 4
Assess Robustness of Methodology
Apply to Network Structures
Scale Free
Small World
Significance through Linear Regression
Deliver Recommendations

Figure 18. Pruned Process Flow

4.2 Experimentation

In order to cover the different types of network structures, and even the subtle differences between networks of the same type, the experimentation plan shown in Table 27 was implemented. This experimentation plan served which, if any, of the different network and node characteristics are crucial in determining the make-up of the best node insertion strategies. Ultimately this will determine the overall robustness of the methodology outlined in Chapter 3 in dealing with networks of different structures and characteristics.

Table 27. Node Insertion Pilot Experimentation Plan

Scale Free	Nodes	Exponent		Replicates
	30	2.2		2
	30	2.8		
	200	2.2		
	200	2.8		
Small World	Nodes	Rewiring	Average Degree	Replicates
	30	0.10	2	2
	30	0.25	2	
	200	0.10	4	
	200	0.25	4	

For scale free networks the network size and exponent α are the only variables used because they govern the extent to which a network follows scale free properties. For small word networks, network size, rewiring probability and average degree are the sole variables. While the rewiring probability is important for the generation of small world networks, it is difficult to determine this probability from an operational small world network. However, a larger the rewiring probability results in a network that more closely resembles an Erdös-Réyni random graph, two of the characteristics of which are a low global clustering coefficient and a low average local clustering. Because of their high negative correlation, this average local clustering value is used as a proxy factor for the rewiring probability.

While the power law exponent can be calculated for a small world network, the coefficient of variation (CV) is used as a proxy factor for the power law exponent. CV is calculated using Equation 30. σ is the variance of degrees while μ is the average degree.

$$CV = \frac{\sigma}{\mu} \tag{30}$$

With these factors and levels a $2^2 + 2^3$ full factorial experiment with two replicates is used, resulting in a total of twenty-four experiments. These experiments consist of graphs created using Morris' Prescribed Node Degree, Connected Graph (PNDCG) Algorithm [57] and a small world connected network generator, the code for which is shown in Appendix A. Due to randomness, the exponent for each of the scale-free graphs will differ, resulting in a semi-orthogonal experimentation plan. The characteristics of each of these experiments are shown in Table 28.

The principal response for this analysis is the percent increase in utility gained from using the optimal insertion strategy versus the expected utility gained from a random insertion. Because of the qualitative nature for assessing possible undercover insertions is most likely random, this response identifies the increase in information collection potential resulting from the optimal insertion strategy. Using this response, along with the characteristics of the network the responses were gathered from, a regression model can be created to determine the significance of the different experimentation factors in determining the utility of the top node insertion strategy.

Inherent cost remains at a level of 0.2 for the all experiments. In an additional attempt to reduce noise within the experiments, the SME weights for the individual nodes were held constant between networks of the same size. While this portion of the bias remained constant, due to the randomness involved in the creation of these networks, the degree of each of a network's nodes were affected. This resulted in different individual biases for each node overall, but kept the noise to a minimum.

Table 28. Experiment Network Descriptions

Experiment #	Network Type	Nodes	Edges	Average LC	Density	Average Degree	CV
1	1	30	49	0.2022	0.1126	3.3	1.0511
2	1	30	39	0.2364	0.0897	2.6	1.2959
3	1	30	45	0.1819	0.1034	3	1.2533
4	1	30	36	0.0112	0.0828	2.4	1.7322
5	1	200	300	0.0935	0.0151	3	1.8476
6	1	200	253	0.0849	0.0127	2.5	2.2108
7	1	200	357	0.1562	0.0179	3.6	2.0659
8	1	200	215	0.0349	0.0108	2.2	2.5333
9	2	30	30	0.0000	0.0690	2	0.2274
10	2	30	30	0.0000	0.0690	2	0.2936
11	2	30	30	0.0000	0.0690	2	0.2936
12	2	30	30	0.0000	0.0690	2	0.2274
13	2	30	60	0.4478	0.1379	4	0.0928
14	2	30	60	0.2978	0.1379	4	0.1970
15	2	30	60	0.4156	0.1379	4	0.1137
16	2	30	60	0.3000	0.1379	4	0.1970
17	2	200	200	0.0000	0.0101	2	0.1876
18	2	200	200	0.0000	0.0101	2	0.2879
19	2	200	200	0.0000	0.0101	2	0.1662
20	2	200	200	0.0108	0.0101	2	0.3325
21	2	200	400	0.4080	0.0201	4	0.1302
22	2	200	400	0.3203	0.0201	4	0.1772
23	2	200	400	0.4223	0.0201	4	0.1121
24	2	200	400	0.3243	0.0201	4	0.2036

Table 28 summarizes the different characteristics of the networks used for experimentation. It is important to note that even networks of the same size and overall type have differing characteristics. Experiments 17 and 18 for example are both scale free, with 200 nodes, and densities of 0.0101, but differ in terms of their CV, suggesting that the degree distributions do differ slightly. Even replicated networks, such as experiments 1 and 3, resulted in different overall characteristics.

4.3 Analysis

For each of the experiments performed, the top insertion strategy, expected utility value, top insertion utility value, and overall computation time were recorded. These values are shown in Table 29 and Table 30. The results in Table 29 were found using the *parfor()* parallel processing function in MATLAB on an Intel(R) Xeon E5-1620 3.6 GHz processor with eight cores and 32 GB memory. The final four experiments involved the evaluation of 66,018,450 possible insertion scenarios, totaling over seventy hours of computation time for each experiment on the machine used for experiments 1-20. For this reason a more powerful machine was used for experiments 21-24, which allowed for data to be gathered in tractable time. The results in Table 30 were found using a dual Intel Xeon E5-2650v2 workstation with 192 GB of RAM.

Table 29. Experimental Results: Experiments 1-20

Experiment	Network Type	Expected Utility	Top Insertion Utility	Utility Δ	% Utility Δ	# Arcs	Comp Time (s)	Comp Time (h)
1	1	-0.0694	0.1685	0.2379	343%	3	1.7954	0.0005
2	1	-0.1764	0.1722	0.3486	198%	1	0.3073	0.0001
3	1	-0.1429	0.1405	0.2834	198%	3	1.7269	0.0005
4	1	-0.2080	0.1860	0.3940	189%	2	0.3318	0.0001
5	1	-0.1406	0.2512	0.3918	279%	3	974.9646	0.2708
6	1	-0.2561	0.1444	0.4005	156%	2	18.5989	0.0052
7	1	-0.2496	0.1716	0.4213	169%	3	1158.9660	0.3219
8	1	-0.2074	0.2187	0.4261	205%	2	17.1319	0.0048
9	2	-0.0576	0.0969	0.1545	268%	2	0.4637	0.0001
10	2	0.0689	0.2174	0.1485	215%	2	0.4176	0.0001
11	2	-0.0163	0.1420	0.1583	**972%**	2	0.4314	0.0001
12	2	0.0429	0.2132	0.1702	396%	2	0.4545	0.0001
13	2	0.0836	0.2992	0.2156	258%	4	4.5680	0.0013
14	2	0.0157	0.1595	0.1438	**914%**	4	4.5612	0.0013
15	2	0.0879	0.2969	0.2090	238%	4	4.4112	0.0012
16	2	0.0571	0.1915	0.1344	235%	4	4.6613	0.0013
17	2	0.0972	0.2550	0.1577	162%	2	16.0465	0.0045
18	2	0.0666	0.2763	0.2097	315%	2	15.5402	0.0043
19	2	0.1001	0.2676	0.1675	167%	2	15.6466	0.0043
20	2	0.0548	0.2825	0.2277	415%	2	15.5407	0.0043

Table 30. Experimental Results: Experiments 21-24

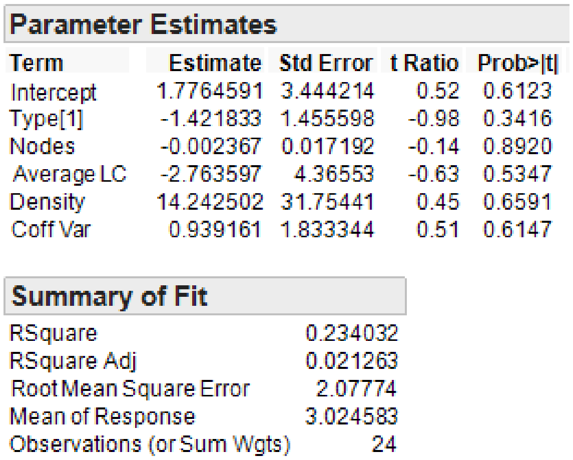
Experiment	Network Type	Expected Utility	Top Insertion Utility	Utility Δ	% Utility Δ	# Arcs	Comp Time (s)	Comp Time (h)
21	2	0.0807	0.2563	0.1756	218%	4	20986.547	5.830
22	2	0.0533	0.2261	0.1728	324%	4	21808.150	6.058
23	2	0.0994	0.2687	0.1693	170%	4	20643.413	5.734
24	2	0.0636	0.2259	0.1624	255%	4	22119.201	6.144

The results shown in Tables 29 and 30 show huge improvement on insertion utility when compared to the expected utility for a random insertion. Each experiment yielded at least a 150% increase in utility, which we relate to an increase in information collection potential. Some networks experienced increases over 900%. Experiments 11 and 14 yielded 972% and 914% increases respectively. While the table shows a marked increase over the average utility for small world networks, the percent change in utility remains similar to the results found with scale free networks. This response allowed for the assessment of the largest proportion changes in terms of utility, meaning the use of this methodology should improve information collection while minimizing the risk of an inserted asset. The utility plots for Experiments 1-20 are shown in Appendix

B.

Robustness Assessment.

For the regression, the factors of interest were number of nodes, average local clustering, density, coefficient of variation, and network type. The results of the regression are shown in Figure 19.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.7764591	3.444214	0.52	0.6123
Type[1]	-1.421833	1.455598	-0.98	0.3416
Nodes	-0.002367	0.017192	-0.14	0.8920
Average LC	-2.763597	4.36553	-0.63	0.5347
Density	14.242502	31.75441	0.45	0.6591
Coff Var	0.939161	1.833344	0.51	0.6147

Summary of Fit

RSquare	0.234032
RSquare Adj	0.021263
Root Mean Square Error	2.07774
Mean of Response	3.024583
Observations (or Sum Wgts)	24

Figure 19. Experimentation Regression Results

None of the five independent variables were significant in determining the overall percent change in utility from the expected value of a random insertion to the optimal insertion strategy. This suggests that the model outlined provides a robust analysis approach that is not affected by network type, size, or other inherent structure characteristics. Figure 20 shows the residuals for the initial regression. It seems that the variance of error term is not constant, which is an assumption for linear regression.

To remedy this problem, an inverse transformation was applied to the response following the application of the Box-Cox Test, the results of which are shown in Figure 21. The model also seems to have high variance inflation factors (VIFs) which result from dependent factors in the model. While the assumption of independent factors is important for linear regression, the goal of this assessment is not to predict utility, but to determine methodology robustness. The prediction of the overall percent increase in utility has no value in terms of this analysis. Even if this response could be predicted, there would be no insight provided on the insertion scenario resulting in this percent increase, and thus, no operational value. Therefore, less emphasis was put on decreasing these high VIFs.

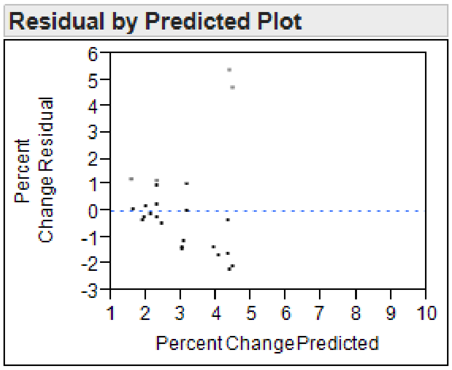


Figure 20. Error Variance

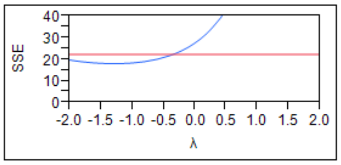
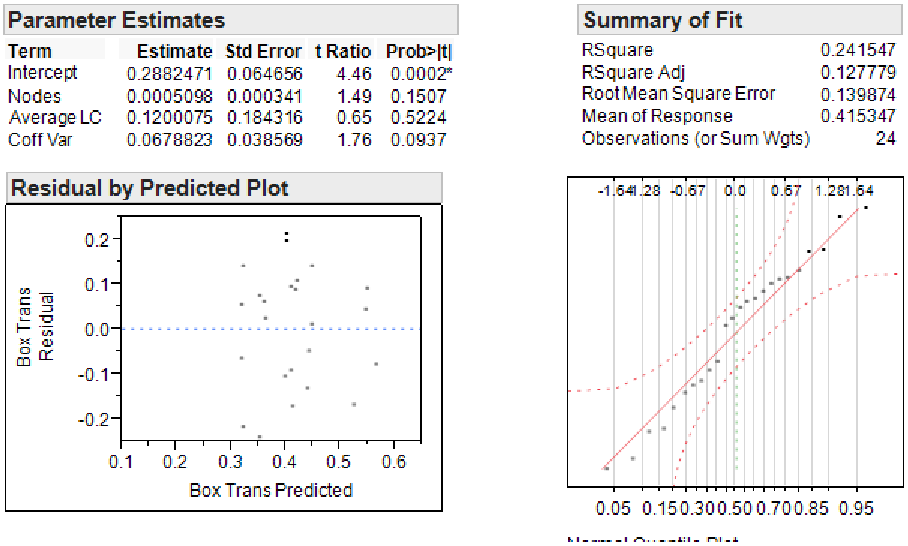


Figure 21. Box-Cox Transformation

The regression following this transformation and removal of highly multicollinear factors yielded the same results and followed both the constant variance and normality assumptions on the error term. This validation is shown in Figure 22.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.2882471	0.064656	4.46	0.0002*
Nodes	0.0005098	0.000341	1.49	0.1507
Average LC	0.1200075	0.184316	0.65	0.5224
Coff Var	0.0678823	0.038569	1.76	0.0937

Summary of Fit

RSquare	0.241547
RSquare Adj	0.127779
Root Mean Square Error	0.139874
Mean of Response	0.415347
Observations (or Sum Wgts)	24

Figure 22. Transformed Model Results

While the elicitation of the inherent costs of creating relationships within a network is outside of the scope of this analysis, it is important to note that as the inherent cost increases, the number of relationships involved in the top insertion candidate decrease, which is a trivial conclusion.

4.4 Analysis Summary

This chapter investigates the methodology presented in Chapter 3. A designed experiment of twenty-four randomly generated networks was implemented to test model performance, specifically from a robustness perspective. The results indicate that the methodology developed is robust to network type and other network characteristics.

V. Conclusions

This research was able to identify a quantitative methodology for the evaluation of insertion scenarios from a risk-benefit perspective. Current undercover operations seem to only utilize subjective and qualitative definitions of risk and benefit and do not consider the structure of the network minimally, if at all. Chapter 2 outlined past research crucial to the formulation of this methodology and its application. The most crucial of these ideas included network type, Laplacian centrality, and past network utility models. Chapter 3 then outlined the methodology itself, also applying it to a real-world network. Chapter 4 assessed the robustness of the methodology through the use of linear regression.

5.1 Contributions

Social Network.

This research was able to extend Laplacian centrality to the insertion of nodes while maintaining the Laplacian energy relationship outlined previously by Qi *et al.* [45]. An specific extension to Jackson and Wolinsky's [52] node utility model was also provided, followed by a generalized formulation of the specific model. The specific methodology implemented the Laplacian energy extension along with other local, intermediate, global, and full network measures, in order to determine the utility of differing node insertion scenarios, specifically for the Zachary Karate Club Network, a staple network in the SNA community.

Operational.

Operationally, this research combined qualitative and quantitative methods for the comparison of different node insertion scenarios. Previously, primarily qualitative methods were used purely based on subject-matter expertise. With the methodology provided, SME input is elicited and incorporated into the model to act as a benefit for inserted relationships. This research also showed that attempting to create a higher number of inserted relationships does not yield the highest overall utility in terms of information collection potential; due to the amount of risk that is inherent with the creation of more relationships in a network.

The utility of any insertion strategy provided can not only be assessed, but also quantitatively compared to different insertion strategies, within a single network. The creation of a portfolio of options for a specific network is also possible to allow flexibility in the decision-maker process, allowing for the most intelligent insertion strategy.

5.2 Future Research

The first extension would be to Clandestine Networks, where the certainty within a network varies. Carley *et al.* [20] previously identifies the risks involved in determining vital actors within a clandestine network, so a methodology already exists that could be adapted to address this problem. Because real-world networks are inherently uncertain; the pursuit of this research would provide greater fidelity for the node insertion problem.

This research does not consider the secondary or tertiary effects of an insertion in to a network, and assumes that new relationships are created simultaneously and instantaneously. If an extension to dynamic networks was made, changes to other parts of the network (over time) could be quantified. In addition, the actual process of insertion could be analyzed. One could also obtain the answer to the question of when to create which relationships, ultimately providing step-by-step instruction for the inserted actor.

Currently the research is also limited to one network, where an insertion must be made. In real-world counterterrorism operations, there are hundreds of possible networks to choose from. One of these networks might be better suited for insertion than others based on size, structure, or even qualitative characteristics of the actors within it. For this reason, a methodology for the comparison of insertion scenarios between networks would be beneficial.

While degree assortivity was used as the network functionality metric for the specific application of node insertion, changing the groupings, perhaps to qualities of each actor, could provide more realistic results for the node insertion problem.

The inherent cost values used in both the specific application and experimental portion of the analysis were purely placeholders. If more analysis could be done on true costs of building relationships within an enemy network, the insertion strategies would provide more accurate results. In addition, just as Jackson and Wolinsky provide a c_{ij} to identify the cost of keeping a relationship between v_i and v_j, these inherent costs could change for each relationships, or even increase depending on if how many relationships the inserted node already maintains.

With the application to a real-world, operational network as a case study, it would be beneficial to look at differing the individual node bias metrics, applying legitimate SME weight elicitation techniques, and creating different benefits and risk measures for more node insertion purposes..

For this research, the maximum number of relationships allowed was governed by the average degree of the network. Allowing for more than this value would increase the computational time to find the optimal solution, but also allows for the elicitation of insertion strategies that are near the true optimal for the network and are not bound by the average degree. With this extension, the application of heuristics would be necessary due to the combinatorial nature of the node insertion problem. In addition, the largest network involved in the experimentation was a 200 node network. The analysis of larger networks, perhaps an order of magnitude larger, could provide more insight on the node insertion problem, which would be made possible by heuristic application.

The research focuses on a composite network, where all of the different layers are combined into a single adjacency matrix. Extending to layered networks could provide a higher degree of fidelity and make the methodology more operationally relevant by assessing the benefit of insertions into different layers of the network. These layers could be based on different types of affiliations or other means of network disaggregation.

Because Laplacian centrality is relatively new, its correlation with other SNA measures is not known quantitatively. One could apply the methodology of Guzman *et al.* [6]to provide quantitative and definitive correlation measures for Laplacian centrality. In addition, there seemed to be correlation between the leading eigenvalue of the adjacency matrix and network functionality. While this relationship was not explored it could provide faster computation for overall functionality of a network.

Previously mentioned was the inherent domination of an insertion strategy compared to the sum of its parts, specifically with Laplacian centrality. Node insertion has not been quantitatively analyzed prior to this research, so the tendencies for other centrality measures in terms of this problem are not known. Future research could include the analysis of other centrality and structural measures, which would then lead into research possibly involving multiple insertions or even the creation of an insertion centrality measure, which might be able to differentiate more influential insertion strategies based on the structure of the network.

5.3 Overall Conclusions

While current Social Network Analysis focuses on network resiliency, interdiction, and influence, more insight into the counterterrorism landscape can be gained by the evolution of these tactical methods. This involves the use of not node changes within a network, but node insertion. The current system for insertion of undercover operatives into criminal networks seems to be ineffective, basing the risks and benefits of an insertion on qualitative methods. These methods most likely result in a random insertion scenario that lacks potential for information collection in addition to providing unnecessary risk based on the structure of the network. Network insertions made in an effective way provide the greatest opportunity for successful counterterrorism operations.

Appendix A. Model Implementation Code

```
function [edgeAddition totalLapEnergyG totalChange risk utility ...
    best benefit indiv w t bb cc dd] = ...
    LaplacianEnergyFixed_parfor(A, weights)
%elimiate memory problems by eliminating yy alltogether, and just ...
    using the
%m matrix. still gets large but does not have to have a number ...
    rows, just
%the number of arcs added in that scenario across...

if ischar(A) ==1
    fid = fopen(A);
    C = textscan(fid, '%s %s');
    fclose(fid);

    C{1}(1) = [];
    C{1}(1) = [];
    C{2}(1) = [];
    C{2}(1) = [];

    rows = str2double(C{1}(:))+1;
    cols = str2double(C{2}(:))+1;

    list = [rows cols];
    A = EdgeToAdj(list);

    clear C list rows cols
end

[a b] = size(A);
```

```

tic;

%nonNormDegreeG = sum(A)';
degreeG = sum(A)';
%degreeG = (nonNormDegreeG - min(nonNormDegreeG)) / ( ...
    max(nonNormDegreeG) - min(nonNormDegreeG) );
avgdeg = sum(sum(A))/a;
degreeMatG = [degreeG degreeG.^2];
clear degreeG
sumDegreeMatG = sum(degreeMatG');

%get average degree
medge = sum(degreeMatG(:,1));
avgDegree = medge/a;


%individual node weight
%w = rand(1,a)*10;
dw = degreeMatG(:,1)';
dw =(dw - min(dw)) / ( max(dw) - min(dw) );

%subject matter expert weights
smew = weights';
%smew = [0.119 0.4984 0.9597 0.3404 0.5853 0.2238];
smew = (smew - min(smew)) / ( max(smew) - min(smew) );

delt = 0.5;
nonNormw = delt*(dw)+(1-delt)*smew;
w = (nonNormw - min(nonNormw)) / ( max(nonNormw) - min(nonNormw) );

%total Laplacian Energy for Graph G
```

```
totalLapEnergyG = sum(sumDegreeMatG);

%G Assortivity
Q = GetAssort(A,degreeMatG(:,1)');

%closeness
nonNormbc=closeness(A);
bc = (nonNormbc - min(nonNormbc)) / ( max(nonNormbc) - ...
    min(nonNormbc) );


%we transform graph G to graph G* where G* is the graph with ...
    inserted nodes
%and edges. To do this we use an edge addition list

%numRel = maxmimum number of relationships to consider
numRel = floor(avgDegree);
%numRel = 2;
[total indiv] = nchooseuptok(a,numRel);
indiv = [0;indiv];

%edge addition matrix. holds all possible scenarios
edgeAddition = zeros(total,numRel);

%inherent costs
inherent = 0.2;

v = [1:a];
count = 1;

%where to start with the new m matrix. using parfor, parralel ...
    processing
```

```
rowCount1 = 1;
output=cell(total,numRel);
for i= 1:numRel

    % get all possible combinations for relationships
    m = nchoosek(v,i);
    [c d] = size(m);


    rowCount2 = rowCount1 + indiv(i+1)-1;

    edgeAddition([rowCount1:rowCount2],[1:i]) = m(:,[1:i]);

    %increment row start to create ematrix
    rowCount1 = rowCount2 + 1;
    parfor j = 1:c

        out = somefun(A,m,a,inherent,d,bc,w,j,i)
        %increment count
        count = count + 1;
        output{j,i}=out;
    end

    clear m
end
par_out = [];
for q = 1:numRel
par_out=[par_out; cell2mat(output(:,q))];
end

totalLapEnergyGPrime=par_out(:,1);
QG=par_out(:,2);
```

```
NonweightedBen=par_out(:,3);
NontotalCost1=par_out(:,4);
totalCost2=par_out(:,5);


%importance of node scenario i, normalized
NonNormTotalChange = totalLapEnergyGPrime-totalLapEnergyG;
totalChange = (NonNormTotalChange - min(NonNormTotalChange)) / ( ...
    max(NonNormTotalChange) - min(NonNormTotalChange) );
totalChange = totalChange';

%absoluate mod change from G to G'
NonNormModChange = abs(QG-Q);
modChange = (NonNormModChange - min(NonNormModChange)) / ( ...
    max(NonNormModChange) - min(NonNormModChange) );

%benefit for each node added

%normalizing indiv node bias
weightedBen = (NonweightedBen - min(NonweightedBen)) / ( ...
    max(NonweightedBen) - min(NonweightedBen) );

%squared
squared = 1;

%normalizing closeness cost
totalCost1 = (NontotalCost1 - min(NontotalCost1)) / ( ...
    max(NontotalCost1) - min(NontotalCost1) );

%utility = totalChange-totalCost;
benefit = totalChange'+weightedBen;
risk = totalCost1 + totalCost2 + modChange;
```

```

%normalizing risk and benefit
benefit = (benefit - min(benefit)) / ( max(benefit) - ...
   min(benefit) );
risk = (risk - min(risk)) / ( max(risk) - min(risk) );

%weighting risk and benefit
wb = .50;
wr = 1-wb;
utility = wb*benefit-wr*risk;

[c loc] = max(utility);
best = [c loc];

t = toc;
%s = sort(utility);

%figure out way to plot without doing by hand
%only works with up to 4 arcs
%for i=1:total
%   switch color(i)
%   case 1
%       hold on
%       plot(i,utility(i),'ko')
%   case 2
%       hold on
%       plot(i,utility(i),'go')
%   case 3
%       hold on
%       plot(i,utility(i),'bo')
%   case 4
%       hold on
%       plot(i,utility(i),'ro')
%   end
%end

totalMat = [1:total]';

figure()
plot(totalMat,utility,'.');

%labels and axis and plots and everything
hold on
%plot(totalChange,'x')
hold on
%axis([0, total + 1,min(utility)-10, max(totalChange)+10])
title('Insertion Scenario vs. Value')
ylabel('Value')
xlabel('Node Insertion Scenario')
%labels = cellstr(num2str([color]'));
%labels = cellstr(num2str([totalMat]));
%text(double(totalMat(:,1)), double(utility(1,:)'), labels, ...
   'VerticalAlignment','bottom', ...
%                              'HorizontalAlignment','right')
grid on

bb = best(1);
cc = mean(utility);
dd = edgeAddition(best(2),:);
end
```

Appendix B. Small World Generator Code

```
function [A density avgd edges] = generateSmallWorld2(n,d2,p)

%create a regular matrix that has n nodes and a total degree of d2
A = zeros(n,n);
d = d2/2;
%d=d2;
for i = 1:n
    if i+d≤n
    A(i,[i+1:i+d]) = 1;
    else A(i,:) = 1;
        A(i,[i+d-n+1:i]) = 0;
    end
end
A = A+A';

%rewire
%removing rewired edges

%get random probs
rewire = rand(1,n);
rewire = rewire<p;

%find arcs to remove
%remove = ceil(rand(1,n)*d2);
remove = rand(n,n).*A;

%find arcs to add
add = ceil(rand(1,n)*n);

```

```
for j = 1:n
    if rewire(j)== 1
        %make sure the added arcs do not already exist for the ...
            current node
        while A(j,add(j)) == 1 || add(j) == j
            add(j) = ceil(rand()*n);
        end

        %remove highest random edge value
        %keep connected graph
        check = sum(A);
        [m loc] = max(remove(j,:));

        %check for connectedness
        a = 2;
        while a > 1

        A(j,loc) = 0;
        A(loc,j) = 0;

        [ci sizes] = components(sparse(A));
        [a b] = size(sizes);

        %if new deletion makes unconnected, readd the deleted arc and
        %choose new node for rewiring from the second component
        if a > 1

        %find shorcut to add from other component
        while ci(add(j)) == ci(j) || (A(j,add(j)) == 1 || add(j) ...
            == j)
            add(j) = ceil(rand()*n);
        end

        remove(j,loc) = 0;
        remove(loc,j) = 0;

        A(j,add(j)) = 1;
        A(add(j),j) = 1;

        else

        remove(j,loc) = 0;
        remove(loc,j) = 0;

        %add shortcuts
        A(j,add(j)) = 1;
        A(add(j),j) = 1;
        end

        [ci sizes] = components(sparse(A));
        [a b] = size(sizes);
        end


   end
end

edges = (sum(sum(A))/2);
density = edges/(n*(n-1)/2);
avgd = sum(sum(A))/n;
end
```

Appendix C. Scale Free PNDCG Inputs

30 Nodes, $\alpha = 2.8$		30 Nodes, $\alpha = 2.2$	
IS_DIRECTED	N	IS_DIRECTED	N
RANDOM_SEED	0	RANDOM_SEED	0
NUM_NODES	30	NUM_NODES	30
PCT_DEVIATION_FROM_EX	100	PCT_DEVIATION_FROM_EX	100
EX_FILE	ex_U_1000.txt	EX_FILE	ex_U_1000.txt
DEGREE_DIST	2	DEGREE_DIST	2
POWER_DIST_EXP	2.8	POWER_DIST_EXP	2.2
DEGREE_DIST_FILE	degree.txt	DEGREE_DIST_FILE	degree.txt
PCT_CLUSTERING	0	PCT_CLUSTERING	0
NUM_OUTPUT_FILES	5	NUM_OUTPUT_FILES	5
OUTPUT_FILE_START_NUM	6	OUTPUT_FILE_START_NUM	6
OUTPUT_DATA_FILE	outputgraph2.txt	OUTPUT_DATA_FILE	outputgraph7.txt

200 Nodes, $\alpha = 2.8$		200 Nodes, $\alpha = 2.2$	
IS_DIRECTED	N	IS_DIRECTED	N
RANDOM_SEED	0	RANDOM_SEED	0
NUM_NODES	200	NUM_NODES	200
PCT_DEVIATION_FROM_EX	100	PCT_DEVIATION_FROM_EX	100
EX_FILE	ex_U_1000.txt	EX_FILE	ex_U_1000.txt
DEGREE_DIST	2	DEGREE_DIST	2
POWER_DIST_EXP	2.8	POWER_DIST_EXP	2.2
DEGREE_DIST_FILE	degree.txt	DEGREE_DIST_FILE	degree.txt
PCT_CLUSTERING	0	PCT_CLUSTERING	0
NUM_OUTPUT_FILES	5	NUM_OUTPUT_FILES	5
OUTPUT_FILE_START_NUM	42	OUTPUT_FILE_START_NUM	42
OUTPUT_DATA_FILE	outputgraph8.txt	OUTPUT_DATA_FILE	outputgraph13.txt

Appendix D. Quad Chart

A Risk Based Approach to Node Insertion Within Social Networks

Chancellor A.J. Johnstone*, Jennifer L. Geffre, James F. Morris
Sponsor: NASIC/SMRB

Background

- Current counterterrorism (CT) techniques involve node changes: disruption, interdiction, and influence
- Effective, but the integration of alternative methods could provide additional insight to the CT decision landscape
- Current guidelines for undercover operations based primarily on qualitative methods and provides less than optimal insertion strategies
- Need for intelligent insertions apparent

Problem Statement

This research aims to provide a structured methodology for overt network infiltration through the application of node insertion. This includes the formulation of risk and benefit measures from the perspective of the inserted node, with a focus on information collection.

Motivation

"The American people and interests will not be secure from attacks until this threat is eliminated–its primary individuals and groups rendered powerless, and its message relegated to irrelevance"

Model

Specific model includes a local measure, intermediate measure, global measure, and a full network measure

$$\overbrace{\underbrace{\text{degree}}_{\text{local}} + \underbrace{\text{Laplacian}}_{\text{intermediate}}}^{\text{benefit}} - \overbrace{\underbrace{\text{closeness}}_{\text{global}} - \underbrace{\Delta\text{assortivity}}_{\text{full network}} - \text{cost}}^{\text{risk}}$$

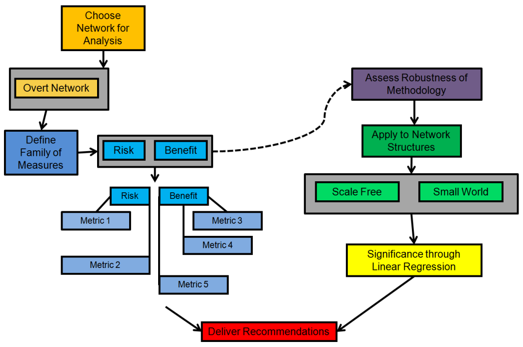
- Specific Model:

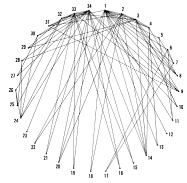
$$U_{x_j} = w_B B_{x_j} - w_R R_{x_j} \qquad \forall j$$
$$\phi_i = w_s d_i + (1 - w_s) p_i$$
$$B_{x_j} = w_1 L_{x_j} + w_2 \textstyle\sum_{i \in x_j} \phi_i \qquad \forall j$$
$$R_{x_j} = w_3 m_{x_j} + w_4 \textstyle\sum_{i \in x_j} C_i + w_5 z_{x_j} \qquad \forall j$$

- General Model:

$$U_{x_j} = w_B B_{x_j} - w_R R_{x_j} \qquad \forall j$$
$$\phi_i = \textstyle\sum_s^S w_s b_s^i \qquad \forall i$$
$$B_{x_j} = w_b \textstyle\sum_{i \in x_j} \phi_i + w_c \sum_k^K w_k \gamma_k^{x_j} \qquad \forall j$$
$$R_{x_j} = \textstyle\sum_l^L w_l \eta_l^{x_j} \qquad \forall j$$

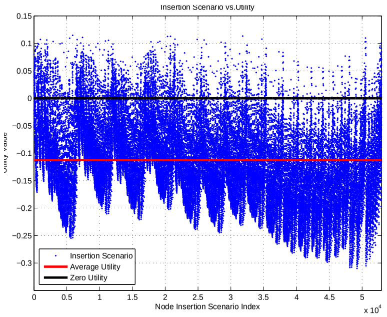
Methodology


Choose Network for Analysis
Overt Network
Define Family of Measures
Risk
Benefit
Risk
Benefit
Metric 1
Metric 2
Metric 3
Metric 4
Metric 5
Assess Robustness of Methodology
Apply to Network Structures
Scale Free
Small World
Significance through Linear Regression
Deliver Recommendations

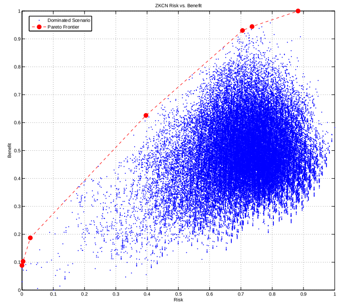


Visual of Zachary Karate Club, the social network chosen for application of methodology.

Zachary Karate Club Network Application



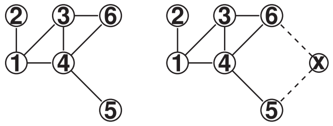
Points above the black line show insertion scenarios which provide more benefit than risk. The red line is where current undercover operations lie in terms of information collection potential.



By shifting emphasis from minimizing risks to maximizing benefits a Pareto frontier is created. Red points identify dominating insertion scenarios for at least one weight scheme.

Laplacian Extension

- Given a graph G we can create a graph $G^*_{x_j}$ which includes node insertion scenario x_j
- $G \subset G^*_{x_j}$ so $E_L(G) < E_L(G^*_{x_j})$
- The change in Laplacian energy from G to $G^*_{x_j}$ is the Laplacian centrality of inserted node



$E_L(G) = 54$ $\qquad$ $E_L(G^*_{x_j}) = 70$

- The Laplacian energy of the inserted node is 16

Laplacian centrality extends passed node inactivation to node insertion using the same equation presented by Qi *et al*

Robustness Assessment

- Linear regression used to assess robustness of methodology to:
  - Network type, size, density, clustering, average degree, and coefficient of variation
- No variables significant at $\alpha = 0.05$ level, methodology robust to these network characteristics
- Not a predictive model, assumptions for linear regression not necessary

Contributions

- Extended Laplacian centrality to node insertion
- Provided quantitative methodology for node insertion, extension of Jackson and Wolinsky's utility
- Applied methodology to the Zachary Karate Club, delivering optimal insertion and portfolio of near optimal insertions
- Determined robustness of methodology to network structure characteristics through linear regression

Bibliography

1. V. Krebs, “Uncloaking terrorist networks.” `http://journals.uic.edu/ojs/index.php/fm/article/view/941`, 2002. Accessed: 2014-11-22.

2. M. E. Newman, D. J. Watts, and S. H. Strogatz, “Random graph models of social networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 1, pp. 2566–2572, 2002.

3. M. E. Newman, “Models of the small world,” *Journal of Statistical Physics*, vol. 101, no. 3-4, pp. 819–841, 2000.

4. C. M. Rocco S and J. E. Ramirez-Marquez, “Vulnerability metrics and analysis for communities in complex networks,” *Reliability Engineering & System Safety*, vol. 96, no. 10, pp. 1360–1366, 2011.

5. T. J. Herbranson, R. F. Deckro, J. W. Chrissis, and J. T. Hamill, “Considering the isolation set problem,” *European Journal of Operational Research*, vol. 227, no. 2, pp. 268–274, 2013.

6. J. D. Guzman, R. F. Deckro, M. J. Robbins, J. F. Morris, and M. A. Ballester, “An analytical comparison of social network measures,” *IEEE Transactions on Computational Social Systems*, vol. 1, no. 1, pp. 35–45, 2014.

7. L. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science.* Empirical Press, 2004.

8. “Terrorism-related inadmissibility grounds (trig).” `http://www.uscis.gov/laws/terrorism-related-inadmissability-grounds/terrorism-related-inadmissibility-grounds-trig`. Accessed: 2014-11-22.

9. C. Hagel, “Quadrennial defense review,” 2014.

10. White House, “National strategy for counterterrorism,” 2011.

11. J. C. Johnson, “The 2014 quadrennial homeland security review,” 2014.

12. U.S. Joint Chiefs of Staff, “Joint doctrine for information operations,” *Joint Publication 3-13*, 1998.

13. D. Benjamin, *Strategic Counterterrorism.* Brookings Institution, 2008.

14. K. M. Carley, J. Reminga, and S. Borgatti, “Destabilizing dynamic networks under conditions of uncertainty,” in *International Conference on Integration of Knowledge Intensive Multi-Agent Systems, 2003.*, pp. 121–126, IEEE, 2003.

15. U.S. Federal Bureau of Investigation, “Guidelines on undercover operations.” `http://vault.fbi.gov/`, 1994. Accessed: 2014-11-24.

16. C. J. Fijnaut and G. T. Marx, *Undercover: Police Surveillance in Comparative Perspective.* Martinus Nijhoff Publishers, 1995.

17. J. Dobyns and N. Johnson-Shelton, *No Angel: My Harrowing Undercover Journey to the Inner Circle of the Hells Angels.* Random House LLC, 2010.

18. D. Wagner, "Hells angels: The federal infiltration," *The Arizona Republic*, 2005.

19. K. Droban, *Running with the Devil: The True Story of the ATF's Infiltration of the Hells Angels.* Globe Pequot, 2007.

20. K. M. Carley, M. Dombroski, M. Tsvetovat, J. Reminga, and N. Kamneva, "Destabilizing dynamic covert networks," in *Proceedings of the 8th International Command and Control Research and Technology Symposium*, 2003.

21. A.-L. Barabási, R. Albert, and H. Jeong, "Mean-field theory for scale-free random networks," *Physica A: Statistical Mechanics and its Applications*, vol. 272, no. 1, pp. 173–187, 1999.

22. A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, no. 5, 2003.

23. M. Newman, *Networks: An Introduction.* Oxford University Press: New York, 2010.

24. A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

25. P. Erds and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

26. F. Chung, "A whirlwind tour of random graphs," *Encyclopedia on Complex Systems, Springer*, 2008.

27. D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

28. M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

29. A. Bavelas, "A mathematical model for group structures," *Human Organization*, vol. 7, no. 3, pp. 16–30, 1948.

30. K. S. Cook, R. M. Emerson, M. R. Gillmore, and T. Yamagishi, "The distribution of power in exchange networks: Theory and experimental results," *American Journal of Sociology*, pp. 275–305, 1983.

31. P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, pp. 1170–1182, 1987.

32. M. G. Everett and S. P. Borgatti, "Induced, endogenous and exogenous centrality," *Social Networks*, vol. 32, no. 4, pp. 339–344, 2010.

33. L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, pp. 215–239, 1978.

34. M. A. Beauchamp, "An improved index of centrality," *Behavioral Science*, vol. 10, no. 2, pp. 161–163, 1965.

35. M. Herland, P. Pastran, and X. Zhu, "An empirical study of robustness of network centrality scores in various networks and conditions," in *2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 221–228, IEEE, 2013.

36. P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

37. S. P. Borgatti, C. Jones, and M. G. Everett, "Network measures of social capital," *Connections*, vol. 21, no. 2, pp. 27–36, 1998.

38. R. D. Luce and A. D. Perry, "A method of matrix analysis of group structure," *Psychometrika*, vol. 14, no. 2, pp. 95–116, 1949.

39. R. M. Karp, *Reducibility Among Combinatorial Problems.* Plenum Press: New York, 1972.

40. M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

41. R. S. Burt, *Structural Holes.* Harvard Univerity Press, 1992.

42. L. D. Sailer, "Structural equivalence: Meaning and definition, computation and application," *Social Networks*, vol. 1, no. 1, pp. 73–90, 1979.

43. P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Physical Review E*, vol. 65, no. 5, pp. 056109–1–056109–14, 2002.

44. I. Gutman, "The energy of graph," *Ber. Math-Statist*, vol. 103, pp. 1–22, 1978.

45. X. Qi, R. D. Duval, K. Christensen, E. Fuller, A. Spahiu, Q. Wu, Y. Wu, W. Tang, and C. Zhang, "Terrorist networks, network energy and node removal: A new measure of centrality based on laplacian energy," *Social Networking*, vol. 2, no. 01, p. 19, 2013.

46. R. Merris, “Laplacian graph eigenvectors,” *Linear Algebra and its Applications*, vol. 278, no. 1, pp. 221–236, 1998.

47. M. Lazić, “On the laplacian energy of a graph,” *Czechoslovak Mathematical Journal*, vol. 56, no. 4, pp. 1207–1213, 2006.

48. K. Malik, A. Mittal, and S. K. Gupta, “The k most vital arcs in the shortest path problem,” *Operations Research Letters*, vol. 8, no. 4, pp. 223–227, 1989.

49. H. Corley and D. Y. Sha, “Most vital links and nodes in weighted networks,” *Operations Research Letters*, vol. 1, no. 4, pp. 157–160, 1982.

50. E. Nardelli, G. Proietti, and P. Widmayer, “A faster computation of the most vital edge of a shortest path,” *Information Processing Letters*, vol. 79, no. 2, pp. 81–85, 2001.

51. M. Bellmore, G. Bennington, and S. Luhore, “A network isolation algorithm,” *Naval Research Logistics Quarterly*, vol. 17, no. 4, pp. 461–469, 1970.

52. M. O. Jackson and A. Wolinsky, “A strategic model of social and economic networks,” *Journal of Economic Theory*, vol. 71, no. 1, pp. 44–74, 1996.

53. J. Xu and H. Chen, “The topology of dark networks,” *Communications of the ACM*, vol. 51, no. 10, pp. 58–65, 2008.

54. S. T. Smith, K. D. Senne, S. Philips, E. K. Kao, and G. Bernstein, “Covert network detection,” *Lincoln Labratory Journal*, vol. 20, pp. 47–61, 2013.

55. K. M. Carley, “Dynamic network analysis,” in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pp. 133–145, Committee on Human Factors, 2003.

56. W. W. Zachary, “An information flow model for conflict and fission in small groups,” *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

57. J. F. Morris, J. W. O’Neal, and R. F. Deckro, “A random graph generation algorithm for the analysis of social networks,” *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, no. 2, pp. 265–276.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704–0188*

1. REPORT DATE *(DD–MM–YYYY)* 26–03–2015	**2. REPORT TYPE** Master's Thesis	**3. DATES COVERED** *(From — To)* Oct 2013 — Mar 2015

4. TITLE AND SUBTITLE

A Risk Based Approach to Node Insertion
Within Social Networks

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Johnstone, Chancellor A.J., Second Lieutenant, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/ENS)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-15-M-136

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

Dr. James Morris
National Air and Space Intelligence Center
Behavioral Influences Division
Wright Patterson Air Force Base, OH, USA
Email: james.morris.5@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

NASIC/SMRB

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

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A.
Approved for Public Release; distribution unlimited.

13. SUPPLEMENTARY NOTES

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Social Network Analysis (SNA) is a primary tool for counter-terrorism operations, ranging from resiliency and influence to interdiction on threats stemming from illicit overt and clandestine network operations. In an ideal world, SNA would provide a perfect course of action to eliminate dangerous situations that terrorist organizations bring. Unfortunately, the covert nature of terrorist networks makes the effects of these techniques unknown and possibly detrimental. To avoid potentially harmful changes to enemy networks, tactical involvement must evolve, beginning with the intelligent use of network infiltration through the application of the node insertion problem. The framework for the node insertion problem includes a risk-benefit model to assess the utility of various node insertion scenarios. This model incorporates local, intermediate and global SNA measures, such as Laplacian centrality and assortative mixing, to account for the benefit and risk. Application of the model to the Zachary Karate Club produces a set of recommended insertion scenarios. A designed experiment validates the robustness of the methodology against network structure and characteristics. Ultimately, the research provides an SNA method to identify optimal and near-optimal node insertion strategies and extend past node utility models into a general form with the inclusion of benefit, risk, and bias functions.

15. SUBJECT TERMS

social network analysis; SNA; node insertion; Laplacian; undercover; counterterrorism; risk; benefit

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Maj Jennifer L. Geffre (ENS)
U	U	U	UU	111	**19b. TELEPHONE NUMBER** *(include area code)* jlgeffre@gmail.com; (719)-393-2708