3-2004

Aggregation Techniques to Characterize Social Networks

Sara E. Sterling

Aggregation Techniques to Characterize Social Networks

THESIS

Sara E. Sterling
Captain, USAF

AFIT/GOR/ENS/04-12

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

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

AFIT/GOR/ENS/04-12

Aggregation Techniques to Characterize 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 in Operations Research

Sara E. Sterling
Captain, USAF

March, 2004

AFIT/GOR/ENS/04-12

Aggregation Techniques to Characterize Social Networks

Sara E. Sterling

Captain, USAF

Approved:

Dr. Richard F. Deckro Thesis Advisor	Date
Dr. James W. Chrissis Committee member	Date

Abstract

Social network analysis focuses on modeling and understanding individuals of interest and their relationships. Aggregation of social networks can be used both to make analysis computationally easier on large networks, and to gain insight in subgroup interactions. Aggregation requires determining appropriate closely knit subgroups as well as choosing a measure or measures to represent the network data.

This thesis provides the analyst with several techniques for using aggregation to analyze the characteristics of social networks. The contribution of this research lies in its ability to analyze a wide variety of social network structures and available data through two methods for subgroup detection and application of two network measures. These techniques are demonstrated first on notional social networks, then on open source information for the terrorist group, Jema'ah Islamiyah. Since analysts rarely have perfect information of the network structure, an exploration of the effects of missing arcs on subgroup detection is presented.

Acknowledgments

I would like to thank everyone who helped me complete this research. My advisor, Dr. Deckro, constantly encouraged me to explore new areas. I credit him with the breadth and depth of the research in this thesis. As my reader, Dr. Chrissis provided invaluable insight to the thesis document. My classmates offered me opportunities to both excel and be humbled. My kitty, Kitty, gave me unconditional love and constant purring. Most importantly, my husband gave me the love and support I needed every day.

Sara E. Sterling

Table Of Contents

Page

Page

Page

Page

List of Figures

Figure Page

List of Tables

Page

Table Page

Aggregation Techniques to Characterize Social Networks

Chapter 1 - Introduction

1.1 Problem Setting

The analysis of social networks examines the complex interactions of individuals and organizations within specific environments. These associations can encompass both formal relationships, such as an office organizational structure, and informal relationships, such as family or religious ties. Because individuals exist in multiple contexts, their decisions in one network may affect the other individuals in that network, and also all other networks in which they are involved. For example, the captain of a sports team may be considered an individual of power in that context. If the captain tells the team members to boycott a brand name product, those players who are easily influenced by their captain may then bring that ideology to other contexts in which they themselves have influence. Thus, the idea of boycotting the brand percolates throughout the larger social network of overlapping contexts, connected by those individuals in multiple environments.

The ability to effectively model both these formal and informal relations is critical to many aspects of information operations and other types of operations, including the war on terrorism. Understanding relationships between individuals in a group may provide insight to the individual interactions within the group and actions of the group as a whole. Tracking these relationships may provide the analyst with information about ongoing changes in the network. For example, such information might assist the analyst in identifying recruiters, individuals who are consistently introducing other new individuals to the network. Alternatively, new connections between otherwise isolated terrorist cells may indicate imminent organized simultaneous attacks, such as those on the Bali nightclubs, Turkish synagogues, or U.S. embassies in Africa (http://www.cnn.com). Further, by understanding the structural position of people or groups of high influence in a social network it is possible to determine optimal targets of influence. An action may not require capturing or

disabling the individual who is considered to be most important or most influential in the network. An action may be considered successful if an individual is simply excluded from the network, or is fed disinformation which the target individual then disseminates to other members of the network. The key individuals may perhaps be influenced through another person in the network who is close to the leader, but not so securely protected.

A social network may be modeled as a graph in which the nodes represent individuals or groups and arcs represent relationships that exist between individuals or groups. The weight on an arc indicates the measure of association between the two nodes it connects.

Social networks distinguish themselves from the general class of graphs by possessing three properties: 1. high local clustering, 2. low average diameter, and 3. vertex degree distributions which follow the power law (Comellas, *et al*: 2000, Watts: 1999). High local clustering is common in social situations, because two acquaintances of a person are likely to be acquaintances themselves. If Alice and Bob are both Charlie's friends, then they themselves are likely to meet and become acquaintances. Small average diameter is the theory behind the popular notion "six degrees of separation" - that two people, no matter how seemingly remote they are, can be linked by only a small number of intermediaries. The third common characteristic of social networks states that the degree distribution on the vertices follows a power law. This means most people have a small number of acquaintances, and very few people have a large group to which they are immediately connected.

Sociologists have long tried to model these relationships between people, but their models do not always employ rigorous mathematical methods. Typical sociological studies have included surveying the group of interest to form a network representing linkages where all existing arcs have a unity weight. This weighting shows only whether some relationship exists between individuals or groups, but gives no indication of the level of that association. Mathematicians have provided a rigorous foundation for analysis of all types of networks, though they rarely focus on these special structures to try to interpret meaning in the context of social networks. Accurately modeling and understanding social networks requires skills, knowledge and techniques from multiple disciplines.

An early rigorous work on social networks which brings the analysis out of solely the social science realm comes from Duncan Watts' book, *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Watts defines small worlds as social networks which possess the two properties of high local clustering and small average diameter. He showed that networks with these properties are pervasive in the world, from movie actors working in films together, to research scientists coauthoring papers, to seemingly unrelated areas as the neurological structure of nematode worms.

Captain Rob Renfro's Ph.D. dissertation, *Modeling and Analysis of Social Networks*, was one of the first research in the open literature to define a non-unity weight on the arcs. He developed a ratio measure of social closeness, which gives, for example, a higher measure to an individual considered to have more influence in a group. Renfro also considered an individual in more than one context. Since interpersonal relationships are influenced in many aspects of life, (work, family, sports groups, religious organizations), modeling social networks requires a more robust model to accurately represent social interactions in multiple contexts. Using maximum flow algorithms, Renfro identifies the most important person in the network to be the one that has the greatest ability to influence the others in the network.

Many social networks include an overwhelming number of people and groups, making the task of finding an individual of interest extremely difficult. A common technique for reducing the size of a network is through node aggregation. Two nodes connected by an arc in the network can be grouped together into a new, aggregated node which represents the two individual nodes in the new network. This process can be performed iteratively until the network is sufficiently small, yet still offers the required fidelity to conduct the desired analysis. Ideally, such aggregation of nodes would occur by grouping individuals that are, in fact, closely associated.

The most tightly knit set of individuals occurs in a clique, in which each member of the set is mathematically defined as having a direct relationship with every other member in the set. Since each person interacts with every other person, a clique is well-suited for aggregation into a single node. It is therefore important to find these structures in a network. However, sets of individuals

in which each pair have a direct relationship are rare in true social network. Therefore, it often is desirable to find near-cliques and other closely knit subgroups in the network. Identification of subgroups provides the analyst with further insight on the network structure such as whether the subgroups share liaison individuals, or are distinct and can communicate using cut-out individuals. Alternatively, individuals not aggregated into subgroups may still be vulnerable, and serve as targets of exploitation.

If all interactions within the network are well-known, then aggregation by closest relationship is simply a matter of finding those collections of individuals that are most tightly bonded. Once a single aggregated node with the properties of interest has been located, it can be disaggregated, or expanded, into the individuals which make up the group. The process of eliminating arcs and replacing multiple individual nodes with a single group node during aggregation, reduces the amount of information about the relationships between individuals on the aggregated network. Renfro's dissertation suggests a method for aggregating nodes to more quickly find people of higher influence (defined by his ratio measure of social closeness) in the social network, without losing any valuable information.

1.2 Problem Statement and Approach

Renfro's measure improves upon the single unity measures used elsewhere in the literature, but there remain more robust ways to measure individuals of high importance in the social network. Although a metric measure for relationships is unlikely (the triangle inequality and symmetry of relationship often cannot be satisfied), other measures can be placed on the nodes to provide information about relationships within the social network. Node measures indicate relative importance of an individual or subgroup in the network. Standard arc measures are binary – nonexistent if there is no relationship between two nodes, unity weight if there is a relationship. This thesis extends Renfro's ratio measure on the arcs to allow a continuous measure on the nodes.

The social networks under study in this thesis are large webs of interconnected individuals and subgroups. Due to the large size of these social networks, the ability to aggregate nodes, while still

being able to accurately determine individuals of interest, is critical. Large networks are not the only ones worth aggregating. If an analyst is interested in interactions and relationships between subgroups, aggregation of individuals into subgroups is appropriate, regardless of the size of the network. The results of this thesis offer techniques for aggregation based on the measures assigned to nodes or arcs in the network.

This thesis provides several methods for detecting appropriate subgroups for aggregation, and measures for assessing individual and subgroup positions in the network, along with the conditions when such aggregation can be performed successfully. No single method is the perfect choice for every objective – each has its own strengths and weaknesses for different scenarios. To ensure robustness of the aggregated network, whatever measure is being performed on the network must convey the same information whether it is being performed on the disaggregated network of individuals, or a smaller, denser version with aggregated nodes representing subgroups. Therefore, for each of the node aggregation methods developed in the thesis, it is necessary to also provide conditions under which aggregation does not affect the results of appropriate network performance measures.

Whether it is of interest to place a weight on the nodes in the network or on the arcs is also scenario dependent. A weighted node indicates the level of importance of the individual or subgroup represented by the node. Alternatively, a weighted arc indicates the strength of the relationship between individuals the arc connects. Several techniques detailing how a weighting on the arcs can be rewritten as weighting on the nodes are explored in Chapter 2, and then used in the remainder of this thesis. Thus, it is not necessary to make a distinction between arc and node measures when determining whether a method for node aggregation is robust.

This thesis focuses on structural characteristics of terrorist groups. While generally social networks possess the three properties of high local clustering, low average diameter, and power law vertex degree distribution, the parameters of those properties can differ. The groups of interest to the national security structure often keep individuals ignorant of one another for security reasons. Therefore, high local clustering may not be appropriate when modeling these groups. Of interest

in this thesis are groups that pose a threat to the United States, such as nation-state terrorists, transnational terrorists, computer hackers, and drug operators, among others. The methods, however, also apply to business organizations, political groups, or tribal memberships. Many of the techniques can even be partially extended to the analysis of physical networks such as computer systems or power grids.

1.3 Research Assumptions

Initially it is assumed the analyst has perfect knowledge of the network, including who the individuals are, as well as their relationships with others in the network (arc measure) or level of importance in the network (node measure). Further, it is assumed that each person in the network also has full knowledge of the relationships between everyone else. This requirement is necessary for evaluating network measures. These assumptions allow the use of standard network algorithms such as shortest path or maximum flow, among others. For example, consider Alice sending a message to Bob. It is assumed that a message between two individuals always traverses the shortest path. The assumption of perfect knowledge is then relaxed to test the effects of removing pieces of information from the network.

1.4 Thesis Format

The remainder of this thesis begins with an overview of the sociological and mathematical literature of social networks in the Literature Review in Chapter 2. It includes methods used in practice to model social networks, and several techniques for aggregating nodes. Chapter 3 offers clique detection methods as well as several methods for aggregating nodes to aid in the analysis of social networks. Chapter 3 also provides indications for when each is appropriate, and conditions under which aggregation will retain accuracy of network performance measures. This methodology is demonstrated on four notional scenarios in Chapter 4 and a real-world example in Chapter 5. Chapter 6 extends selected results to networks of imperfect information. Conclusions and limitations of the work are offered in Chapter 7, with recommended areas of future research.

Chapter 2 - Literature Review

2.1 Introduction

This chapter reviews the pertinent literature of social network analysis, specifically that relating to modeling social networks and methods for aggregation of large networks. To that end, it is first necessary to understand how network structures are modeled using graph theory. Once a graph structure is defined for social interactions, it may be desirable to condense the structure into a smaller, more manageable representation for performing analysis. Therefore, this literature review offers an overview of methods for aggregating individuals in a social network into groups, as well as measures for relationships between and among groups and individuals.

2.2 Introduction to Social Networks

Social networks have long been studied by sociologists and anthropologists, but one of the first efforts at mathematical modeling outside of the social sciences was introduced by Duncan Watts in *Small Worlds: The Dynamics of Networks Between Order and Randomness*. The small world concept - any two people on earth can be connected through only a small number of intermediaries - has been made popular through the "Six-Degrees-of-Kevin-Bacon" game. Watts introduced the small world concept to improve the traditional methods of performing social network analysis. He claims that traditional sociological and anthropological approaches to obtain data through surveys, observations, or questionnaires, have several inherent problems: respondents are often poor at estimating the number of people with whom they have a relationship, the number of such relationships can change over time, and an individual's definition of relationship is subjective (Watts, 1999: 23). Though researchers have considered small worlds since the 1960s, Watts provided it a rigorous theoretical foundation.

Small world networks are unique within graph theory due to their structure. They reside between perfectly ordered graphs, such as lattices, and perfectly disordered random graphs (Watts, 1999: 41). Social network analysis must consider both local and global properties. Watts discovered

small worlds to be based on one local and one global property. Locally, these networks exhibit dense clustering; globally, the existence of ties between seemingly far away nodes reduces the diameter of the network, meaning that dissimilar people can be connected through a small number of nodes and arcs. Watts found that this structure is exhibited by a wide variety of subjects: movie actors (Watts, 1999: 140), high-voltage electric lines (Watts, 1999: 147), the nervous system of the Caenorhabditis elegans worm (Watts, 1999:153), scientific paper coauthorships and worldwide web links (Watts, 1999: 160).

Watts' work is revolutionary for its rigor in small world analysis. Since social networks are an application of small worlds (i.e. Six-Degrees-of-Kevin-Bacon, or scientific collaborations), Watts' research forms a good foundation for an operations research analysis of social networks. However, his work has several limitations that require further research. For example, he considers only arcs of binary weight – if a relationship exists between two nodes, then those two nodes are connected by an arc of unity weight, else there is no arc (alternatively, an arc of weight 0).

Some assumptions of small worlds are inappropriate for terrorist networks. For example, traditional social network analysis has assumed that assortative mixing, a property of networks in which "people prefer to associate with others who are like them", holds (Newman, Feb 2003: 1). This can be seen in large cities that have ethnic neighborhoods. Closely associated with the property of assortative mixing is homophily, which is the property of social groupings which leads to localized clustering in the network representations. This is the idea that if Alice and Bob are friends, and Alice and Charlie are friends, then Bob and Charlie have a high probability of becoming friends. This may be true in many social situations, but is not necessarily appropriate in the study of terrorist networks, where many individuals are intentionally kept ignorant of one another to protect the security of the network as a whole.

Renfro addresses some of these limitations. He defines "measures of social closeness that are ratio in nature" (Renfro, 2001: 66). This is the first research in the open literature which expands the weight of the relationship between individuals or groups to be other than binary. He then uses

optimization techniques to find the most influential individual in the network. He accomplished this through a "mapping of social network analysis terms to mathematical programming, and specifically flow modeling" (Renfro, 2001: 67). He also extends the single context limitation in Watts' work to multidimensional flows of influence through the social network. Renfro does this by noting:

> "Sharing capacity on the edges in a social network implies that either capacity of the edge is an aggregate of multiple contexts, or based on a known sociological or psychological property of the measure influence where one context directly manifests itself in another context" (Renfro, 2001: 73).

2.3 Introduction to Graph Theory

Much of what made Watts' work the first rigorous modeling of small worlds was his ability to define social networks in the language of graph theory. He represents individuals and their relationships using mathematical structures investigated in graph theory. The definitions in this section come from West's *Introduction to Graph Theory*. West offers the following basic structural definitions (West, 2001: 2):

> A *graph* G is a triple consisting of a *vertex set* $V(G)$, an *edge set* $E(G)$, and a relation that associates with each edge two vertices (not necessarily distinct) called its endpoints ... When u and v are endpoints of an edge, they are *adjacent* and are *neighbors*.

The vertices and edges of a graph are referred to as the *nodes* and *arcs*, respectively, of a network. $|V(G)|$, or simply $|V|$ is the number of nodes in the vertex set. In the context of social networks, individuals (or aggregated groups of individuals) are the nodes. The arcs in a social network represent some form of relationship or association between two individuals or groups. These associations signify whatever data the analyst gathers, such as family relationships, number of times individuals share a phone call, or speed of communications, for example.

Since some of these relationships may exist in only one direction (i.e. Alice always initiates phone calls with Bob), the direction of the relationship must be considered. West defines the following graph structures that can be used in social network modeling to address this situation (West, 2001: 53):

> A *directed graph* or *digraph* G is a triple consisting of a *vertex set* $V(G)$, an *edge set* $E(G)$, and a function assigning each edge an ordered pair of vertices. The first vertex of the ordered pair is the *tail* of the edge, and the second is the *head*; together, they are the

endpoints. We say that an edge is *from* its tail *to* its head .

One method of storing the structural data of a graph is in a matrix. Two common representations are the adjacency matrix and incidence matrix, defined by West:

> The *adjacency matrix* of G, written $A(G)$, is the $n \times n$ matrix in which entry $a_{i,j}$ is the number of edges in G with endpoints $\{v_i, v_j\}$. The incidence matrix $M(G)$ is the $n \times m$ matrix in which entry $m_{i,j}$ is 1 if v_i is an endpoint of e_j and otherwise is 0. If vertex v is an endpoint of edge e, then v and e are incident. The degree of vertex v ... is the number of incident edges (West, 2001: 6).

Since the networks this thesis considers do not have multiple edges between the same pair of vertices, entries in the adjacency matrix will all be either 0 (if no relationship exists between those two individuals or groups) or 1 (if a relationship does exist). Consider an organization in which Bob works for Alice. If an edge (i, j) in the network represents "i is boss of j", a digraph representation of this relationship would show an edge from Alice to Bob and the entry $a_{Alice,Bob}$ in the adjacency matrix would be 1. Alternatively, if the edge (i, j) represents "i works for j", the edge would be from Bob to Alice and $a_{Alice,Bob} = 0$ while $a_{Bob,Alice} = 1$.

This method of storing data is appropriate when it is necessary only to know who has a relationship with whom, and the level of that relationship is unimportant. However, when a level of relationship is known between the nodes in the network, that information can be stored in a similarity matrix, $S(G)$, in which the entry $s_{i,j}$ is the weight of the relationship from i to j. Thus, in the previous example where Alice exerts greater influence over Bob, $s_{Alice,Bob} > s_{Bob,Alice}$. These weights can be integers, continuous, or even fuzzy, as the data gathered for the analysis requires.

There are several important structures which may exist within a graph's topology. "A *clique* in a graph is a set of pairwise adjacent vertices" (West, 2001: 4). A clique on n vertices is noted K_n. In a social network context, a clique indicates a group of people, all of whom have a direct relationship with every other person in the group. "A *path* is a simple graph whose vertices can be ordered so that two vertices are adjacent if and only if they are consecutive in the list" (West, 2001: 5). Communication through a social network from one individual or group to another is often assumed to follow the shortest path connecting the individuals or groups.

"The *complement* G' of a simple graph G is the simple graph with vertex set $V(G)$ defined by $\{u, v\}$ in $E(G')$ if and only if $\{u, v\}$ is not in $E(G)$" (West, 2001: 4). "A graph G is *connected* if it has a $\{u, v\}$-path whenever $\{u, v\}$ is in $V(G)$ (otherwise, G is disconnected)" (West, 2001: 20). The concept of connectedness in social networks is important, as it indicates whether there is some known path between each pair of vertices. If the edges represent a relationship between individuals or groups, then a connected network means every person in the network can interact directly or indirectly with any other person in the network.

Several of the following properties of graphs are important in the study of social networks. Traditional graph theory defines "the *length* of a ... path ... is its number of edges" (West, 2001: 20). Again, considering the path of communication from Alice to Bob, the length of the path indicates the distance the message must travel. "The *chromatic number* of a graph G, written $\chi(G)$, is the minimum number of colors needed to label the vertices so that adjacent vertices receive different colors" (West, 2001: 5). Graphs are almost always colored so that the colors, or labels, on each vertex are distinct from the colors on every adjacent vertex. If the graph is colored in this manner, a higher chromatic number is indicative of a highly dense area (possibly over the entire network, though not necessarily). In a social network, a set of n vertices colored n unique colors signifies a group of tightly knit individuals. These sets of nodes are candidates for aggregation, since they form a group of individuals with a high density of interactions.

Graphs can become quite large. If a part of the graph is of interest, the analysis can focus on a subgraph of the graph G, which is a smaller portion of the graph G. West defines a subgraph more rigorously: "A *subgraph* of a graph G is a graph H such that $V(H)$ is a subset of $V(G)$ and $E(H)$ is a subset of $E(G)$ and the assignment of endpoints to edges in H is the same as in G" (West, 2001: 6). Alternatively, to condense a large graph in order to make it easier to analyze, aggregation or clustering techniques are commonly used.

2.4 Node Aggregation

The purpose of node aggregation is for the analyst to be able to work with a smaller, less detailed graph. Aggregation can be used to reduce large graphs by a node and/or link reduction. Van Miegham describes one common clustering method, *hierarchical clustering*, in which the network is partitioned into subsets of nodes (Van Miegham, 1999: 2115). Each subset becomes an aggregated node in a reduced network, which represents the partitioning in an efficient layered form called a *hierarchy*. This partitioning is a recursive process, where each new layer in the hierarchy indicates another step in the reduction of the network. This process continues until the entire network has been aggregated into a single node.

In *Graphs as Structural Models*, Godehardt claims, "The principle of cluster analysis procedures – when detecting and outlining groups – is that of optimization" (Godehardt, 1990:15). Godehardt lists four types of clustering, or classification: disjoint, nondisjoint, hierarchic, and quasi-hierarchic. Disjoint and hierarchic classifications both forbid overlapping subsets, i.e. a vertex can only be in one aggregated node. Nondisjoint and quasi-hierarchic classifications allow overlapping. As may be the case in many social networks, Godehardt states, "In some classification problems, it is convenient to ask for homogeneous groups and to allow objects to belong to more than one class at the same time" (Godehardt, 1990: 42)

In hierarchic and quasi-heirarchic classification, "the objects and groups are arranged and graphically represented in the form of a genetic tree" (Godehardt, 1990: 42). At the lowest level, each node is an individual. Each level aggregates from the previous, until at the highest level only one node remains, which is the aggregation of every original node.

Godehardt argues "the analyst needs a basis for the valuation of a classification, a clustering criterion since he is interested in a 'good partition' of the sample into clusters, which should be as 'natural' and problem-oriented as possible" (Godehardt, 1990: 43). To that end, Godehardt presents some measures of homogeneity. However, most of his measures and methods rely on the distances between nodes to be metric. For social networks, this is not a reasonable assumption, since

transitivity and the triangle inequality do not necessarily hold. For example, if Alice is Bob's boss, Alice's relationship to Bob is not the same as Bob's relationship to Alice.

2.5 Blockmodeling

Blockmodeling is a common aggregation technique. Informally, Breiger notes: "social contacts can be conceptualized in terms of 'blocks', in each of which the members are in 'structurally equivalent' position (because of their similar ties with third parties) even when they are not in direct contact with each other" (Breiger, 1991: xiii).

Wang and Wong's seminal work, "Stochastic Blockmodels for Directed Graphs," decomposes the adjacency matrix of a network into "submatrices, or blocks, each of which represent ties from individuals in some B_r to individuals in some B_s" (Wang and Wong, 1987: 8), where B_r and B_s are blocks, or sets of nodes in the network. The goal in using these blocks to aggregate networks is to decompose the adjacency matrix into blocks of highest possible density. These blocks of high density show where sets of nodes have many edges between them. If the blocks are on the diagonal, such that the high density is at the intersection of B_s and B_s, then this indicates that B_s is a tightly knit group. Highly dense blocks on the diagonal suggest sets of nodes that are cliques or near cliques, and are therefore candidates for aggregation. Alternatively, if the blocks are off the diagonal, such that the high density is at the intersection of B_s and B_r, then the two sets of nodes have many interactions between them. This situation may occur, for example, when two neighboring families share many of the same activities such as sports groups, school classes, and church groups.

While the blocks seem to offer a structure to aid finding closely associated groups of individuals, finding these blocks of higher density is not an easy task. Nagpaul suggests using TABU search to place collections of nodes into groups such that the within-block variance is minimized (Nagpaul, 2002: 224). Batagelj's paper "Notes on blockmodeling" provides an extensive list of the types of connections between sets of vertices in a graph, and then uses those, along with some convoluted rules, to rearrange the adjacency matrix to produce blocks (Batagelj, *et al*, 1999: 502-3).

Bonacich and McConaghy address the concern "blockmodeling is sometimes regarded as a mere clustering method" by claiming "in actuality, it is a theoretically grounded and algebraic approach to the analysis of the structure of relations" (Bonacich and McConaghy, 1980: 489). They define blockmodels in terms of "the direct and compound relations among individuals or social positions" (Bonacich and McConaghy, 1980: 490) and then uses algebraic techniques to examine the structure of these social relations.

Similar to blockmodeling, White and Harary introduce the concept of cohesive and adhesive blocks which measure how a group of people stay together, either by the influence of individuals in the network (cohesion) or through the relationships between individuals (adhesion). Their paper, entitled "The Cohesiveness of Blocks in Social Networks: node connectivity and conditional density" offers models of cohesion and adhesion and provides the graph theoretic foundation for their measures. They use this methodology to determine how social networks will most likely split apart if crucial links are broken. The example the authors provide is of a group in which the two leaders have a fundamental difference of opinions and each member of the network must decide which leader they want to follow.

The results reported in the White and Harary paper show the methodology based on the cohesion and adhesion measures to be fairly accurate in determining which group members will follow a given leader. These results may be useful in determining the effect of removing a node from a network, cutting communications between leaders of terrorist cells, or driving a wedge between leaders to split the network. The results are predictive in nature, providing a guide of which individuals will follow each leader

Chang and Fung introduce a methodology to replace a cluster of nodes in a Bayesian network with a single node, without changing the underlying joint distribution of the network. They define

> "... a probabilistic Bayesian Network [as] a directed, acyclic graph in which the nodes represent random variables, and the arcs between the nodes represent possible probabilistic dependence between variables. A network as a whole represents the joint probabil-

ity distribution between random variables." (Chang and Fung, 1989: 265)

Their method joins adjacent nodes and aggregates combinable groups. "A group of nodes is called combinable if for every pair of nodes in the group there exists no path between such pair which contains a node outside the group" (Chang and Fung, 1989: 266). This definition of combinable groups is an extension of the classical clique, in which every node must be adjacent to every other node. This relaxation is appropriate in situations where it is not necessary for every individual to have a direct relationship with every other individual, but that they do have an indirect connection through others members of the group. Consider, for example, a group of drug dealers which has a local provider, three middle distributors, and ten local dealers. Suppose the distributors and dealers all know each other. The main provider may not have direct contact with any of the lower level dealers, but he is connected to each of them through the middle distributors. The main provider may not have a direct relationship with every individual, but he does belong in the group, and should be in the aggregated node if this drug group is aggregated together

Seidman and Foster restructure a network with aggregated nodes in order to examine within- and between-subgroup interactions or relations (Seidman and Foster, 1978: 139). They note anthropologists tie people together using family bonds, while sociologists often use a more empirical approach. Sociologists tend to look for a person's role within a sub-society; these social grouping tend to be categorical and are obtained through surveys, which are murky at best. In general, the sociologists are searching for "sets of individuals who are tied to each other more closely than to non-members" (Seidman and Foster, 1978: 140); specifically, they look for cliques.

Seidman and Foster argue cliques are not an appropriate structure to study, since they are difficult to find in large networks, and are in general unlikely to exist at all. They want to find a structure that retains the property of relatively short path between all pairs of nodes, but that path length does not have to be 1, as it is in the classical definition of a clique. This would imply near direct communication between all members of the group. One of their concerns about the classical definition of cliques is robustness: the clique falls apart if only a small number of arcs are removed.

To keep the idea of a closely-knit group, but remove some of the rigid restrictions, Seidman and Foster define a k-plex to be "a graph with n vertices in which each vertex is connected by a path of length 1 to at least $n - k$ of the other vertices" (Seidman and Foster, 1978: 142). Further, they offer an algorithm for finding k-plexes in a graph. As with Chang and Fung's combinable groups, the k-plex offers another relaxation of the clique concept.

In a terrorist group, some individuals may be kept purposely ignorant of each other for operational security reasons, so it is often difficult to find cliques. If a terrorist cell is preparing a coordinated attack, each individual may know only his piece, and not know the members of a cell simultaneously attacking a target on the other side of town or even that such an attack is to occur. Consider a scenario of seven individuals in which each of the two bombing groups has two individuals, and three more individuals are coordinating the attacks. In this scenario, suppose each of the three coordinators know all four of the bombers, but the pairs are unaware of each other. Each individual has a direct connection with at least $n-2$, or 5, other individuals. Therefore, this terrorist cell forms a 2-plex.

When aggregating individuals into groups, it is preferable to group such that sets of individuals most like each other end up together. It was noted earlier that cliques define a group of people who are close to each other, but finding them in a graph is difficult. Sometimes, the actual level of relationship between two people is not known, or is qualitative, such as "Alice and Bob are good friends" or "Alice and Bob communicate frequently". Fuzzy cliques offer some potential solutions for these problems in social network analysis.

2.6 Fuzzy Cliques

The original concept of the clique in a social network, a structure in a graph where each person has a direct connection with every other person in the group, from Luce and Perry's 1949 seminal work is too rigid for social network analysis. Yan Xiaoyan describes five limitations of the original clique definition: redundant connections, rigid definition of membership, uniform structure, network weights, and computational complexity (Yan, 1988: 360-362). One of the greatest practical

limitations is that finding cliques is an NP-complete problem, and therefore may be prohibitive for large networks.

Yan contends that cliques as Luce and Perry define them have too many edges. It is too difficult to find subgraph with $n(n-1)$ edges. The limitations of the classical definition do not offer gradations of membership in the clique. Yan writes, "For any given clique and any given node, there are only two possibilities. The node either belongs to the clique or does not belong to the clique" (Yan, 1988: 360-1). A person can be removed from the clique by losing a direct connection to just one other person in that clique. This requirement that each node be directly connected to every other node produces an uninteresting structure in which no member of the clique is distinguishable from any other. The original definition of clique puts a binary weight on each edge – an edge has weight 1 if a relationship exists – which Yan argues is unrealistic. To be able to assess relationships in a more quantitative manner, Yan recommends using edges with integers or fuzzy strengths. Yan recalls, "In traditional set theory, given any object x and a set A, there are two possibilities: object x either belongs to A or does not," defining a membership function in which $M(x) = 0$ or 1 (Yan, 1988:366). In a fuzzy set, an element can take on any value in [0,1] (Yan, 1988: 366). These membership values between 0 and 1 may become useful if the analyst is unsure of how integrated a member is in the network.

As an example, let $M(x) = 0$ if x is not a drug smuggler, 1 if he is. Consider, an individual who transports a package of cocaine from Colombia to the United States. The individual is clearly a smuggler, but only small time, so merits a membership value of 0.9 Consider instead a farmer in Afghanistan who is caught carrying opium across the border to Pakistan. When caught he claims a local warlord offered him money if he transported a package and asked no questions. The farmer, though not completely innocent, is only a mule, and his membership value is lower, say 0.4.

Yan defines "the strength of a path ... as the smallest fuzzy strength of all the fuzzy strengths on the arrows of the path" (Yan, 1988: 375) and "the length of a path is ... the number of arrows on

the path/the strength of the path" (Yan, 1988: 375). The length increases as more edges separate nodes.

Finally, Yan defines a *fuzzy clique* (Yan, 1988: 377-8):

> "... a maximum strongly connected node subgroup in which each node is connected to all the others directly or indirectly, regardless of the number of intermediate nodes. The core members are those nodes whose distances to and from all clique members are less than or equal to a given fuzzy or non-fuzzy number D"

The members of a clique that are not in the core are peripheral members. Under his definition of fuzzy clique, Yan offers several measures relating nodes to cliques and cliques with each other. He defines a *node-clique coefficient* which gives a measure for a node's relationship to a clique of which it is not a member, a *clique-clique coefficient* which provides a measure of comparison between cliques, and a *node membership coefficient* which measures a node's position within the clique of which it is a member.

To determine the node-clique coefficient, let $n_1, n_2, ..., n_m$ be members of a clique, C, and let n be a node in the network that is not a member of C. The coefficient $K(C,n) = \sum_i \frac{1}{Q_i} / m$, where Q_i is the directed distance from n to n_i (or n_i to n). By Yan's definition of clique, since n is not in the clique C, either there is a path from n to n_i or n_i to n, but not both. Else, n would be in C. The node-clique coefficient tells the analyst how close an individual is to a given clique of which it is not a member. The individual with the largest node-clique coefficient with a given clique may be the individual who passes information to that set of individuals in the network.

To determine the clique-clique coefficient, let C_1, C_2 be two cliques; let the nodes in C_1 be $n_1, n_2, ...n_m$ and let the nodes in C_2 be $n'_1, n'_2, ...n'_k$. Then the clique-clique coefficient $J(C1, C2) = \sum_i \sum_j \frac{1}{Q_{ij}} / m * n$ where Q_{ij} is the path length from n_i in C_1 to n'_j in C_2(or n'_j in $C_2$to n_i in C_1). The clique-clique coefficient offers a measure for the aggregated relationship between two groups of individuals in the network. If communication occurs between groups other than through the official hierarchy, this coefficient will show how much these informal relationships add to the relationship between groups.

The value of a member to the clique of which it is a member is $P'/P \leq 1$, where P' is the number of nodes in the clique whose distance from the member is less than some threshold D and P is the number of members in the clique. Recall that the members of the clique whose distance from every other member in the clique is less than D are in the *core* of the clique. Varying the value of D changes not only the size of the core but also the value of a node within the clique. It does, however, keep the relative measures of nodes, and provides a method for distinguishing members within the clique. Depending on what the weight on the arcs represents in the network, this membership value may tell the analyst which individual in the group is most susceptible, or instigates the most phone calls, or brings the most new members into the group (depending on what the weight on the arcs represents).

Since Yan's definition of a fuzzy clique does not allow a node to be in more than one clique, the computational problems associated with finding all (overlapping) cliques in a network do not exist. Yan offers the following algorithm for detecting fuzzy cliques (Yan, 1988: 382):

Fuzzy clique detection algorithm

1. Calculate the distance matrix for the network
2. $k := 1$
3. Identify all nodes in the kth clique by finding a maximum strongly connected subgraph
4. Calculate the membership values for all the clique members according to the given D
5. Are all the nodes in the network classified into cliques? If yes, stop; otherwise let $k := k + 1$ and go to step 3.

Yan also offers a definition of the clique network of a weighted network, w, as "a weighted network in which the nodes are the cliques in w, and the connections and their weights are the clique-clique coefficients in w" (Yan, 1988: 382). This provides a method for aggregating fuzzy cliques into a smaller, denser network, in which analysis can be performed more quickly due to the smaller size. In this aggregated network, the edge weight between the aggregated nodes (each representing a clique) is the clique-clique coefficient. If the analyst chooses to leave some cliques disaggregated, the edge weight between an individual node in disaggregated clique and an aggregated node is the

node-clique coefficient described above. Further information about fuzzy networks and examples of this method can be found in Appendix A.

2.7 Node Measures

Suppose the measure of interest is not the relationship between individuals in a network, but a member's position in the network as compared to others. To that end, there are several methods for determining the centrality of an individual within the network. Three well known measures of node centrality are degree centrality, closeness centrality and betweenness centrality.

The *degree centrality* measure for a node is the degree of the node (Shaw, 1954). Any individual has a high degree centrality is able to directly communicate with a large number of other individuals. The individual does not have to rely on other members of the network to convey information, and so has some measure of direct influence over a large percentage of the people. However, it is not necessarily the case that the person with the highest degree centrality measure is a person of great importance.

For example, in a large office, the person who delivers the mail has direct contact with almost every other person in the office every day. He may not be an important person in the company, nor may it be obvious he can exert any influence over the employees he sees on a regular basis. He may be the invisible person whose name no one knows. No one knows the janitor who comes to empty the trash every evening, but he sees every scrap of paper people throw away. Alternatively, he may be an individual who has built a personal relationship with everyone over time. If the mail deliverer is then susceptible to outside influence, he can use this web of personal contact and trust to disseminate that outside idea. Thus this degree centrality measure can be indicative of an individual in a high position of power, it may also prove to be a person low in the organizational structure who may or may not have a great deal of influence in the network. It is worth noting that the degree centrality measure is a local property: an individual's measure is dependent only upon his relationships with his immediate contacts. This measure gives no indication of a person's relationship to other individuals more than a path of length one away.

The *closeness centrality* measure for a node is the sum of the shortest paths to every other node (Beauchamp, 1965). This measure represents independence, "the possibility to communicate with many others depending on a minimum number of intermediaries" (Gomez *et al*, 2003: 28). As with degree centrality, an individual with a high closeness centrality needs not rely on a great number of other people in the network to disseminate information. However, the closeness centrality measure is a global property, making it reasonable to compare each person's measure in the entire network, regardless of the path length between them. An individual with a high closeness centrality measure can distribute information most quickly. In a network with two groups separated by an individual who acts as an intermediary between the groups, that intermediary has a high closeness centrality. If an outsider familiar with a terrorist cell needs to disseminate information quickly, he would desire access to the person with a high closeness centrality measure, all other things being equal, in order to get out the information as quickly as possible.

The *betweenness centrality* measure counts the number of shortest paths a node is on (Freeman, 1977). This is an indication of the control a node has on communication in the network, "the possibility to intermediate in the communication of others" (Gomez *et al*, 2003: 29). This measure is also a global measure, as it indicates the level to which an individual's removal disrupts the connectivity of the entire network. In the example above of the intermediary between the two groups, the intermediary has an extremely high betweenness centrality, since his removal disconnects the two groups. If the analyst's goal is to severely hurt communications in a network, then he should seek those individuals with high betweenness centrality. This does not require removing the individual from the network. Instead, it may be desirable to use this individual's position to insert incorrect information into the network, or distort the communications between groups.

Gomez, *et al* suggest using the Shapley value as a node measure. The Shapley value is a game theoretic power index that indicates marginal contribution of a player in the game. Using the Shapley value, Gomez, *et al* suggests one can tell which coalitions may form within the network. Game theory states coalitions form from a group of individuals with a common goal; the coalition

allows the individuals to gain a higher payoff as a group then they would as individuals (Morris, 1994:149). Since the members of the coalition share a common goal, they can be aggregated together in a social network.

2.8 Imperfect Information

In many social networks of interest, relationships between some individuals are unknown. Analytic network techniques require perfect knowledge of the network topology. When relationships are known to exist, but there is only a limited knowledge of the level of that relationship, the analyst can put fuzzy measures on the arcs and apply Yan's methodology for fuzzy cliques. Parsons' research focuses on interpreting imprecise information in databases. He also recommends using fuzzy sets when the actual value is unknown. Parsons claims:

> "Most of the work on the modeling of imprecise information within databases has involved the use of fuzzy sets and fuzzy logic. Fuzzy set theory is a generalization of normal set theory in which it is recognized that the kinds of classes of objects one encounters in the real world do not always have precisely defined criteria of membership." (Parsons, 1996: 357)

When data is missing altogether, different methods are required. Philip Roth sagely advises, "The best possible method of dealing with missing data is to avoid the problem" (Roth, 1994: 538). Given that missing data is inevitable, however, Roth offers methods for performing analysis on data sets with missing data, including several simple methods, hot deck imputation and maximum likelihood estimates.

The simple methods suggested by Roth are *listwise deletion*, *pairwise deletion*, and *mean substitution*. When implementing listwise deletion, the analyst eliminates all data with any amount of information. Suppose the data consists of three types of data, A, B, and C, and some of the data for B is missing. Listwise deletion requires the analyst delete either the entire B data, or the A and C data points where ever B is also missing.

Implementation of pairwise deletion allows the analyst to use the data available for the statistics that can be calculated. For example, suppose again that the data consists of three types of data, A, B, and C, and much of the data for B is missing. The analyst can still find correlations between A

and C. This method can often lead to inconsistent correlations and covariances. Mean substitution allows the analyst to "use the mean value of a variable in place of missing data values for the same variable" (Roth, 1994: 540). Since mean values are used whenever missing data points are encountered, the variance and covariance estimates tend to be unrealistically low. (Roth, 1994: 539-540). For all of these methods, it is necessary for the analyst to be aware where information for the network is actually missing. To possess this knowledge with certainty may be impossible for social network analysis.

With the limitations of the simple techniques listed above, "A growing number of researchers are choosing to estimate missing data values based on other variables in the data" often via regression (Roth, 1994: 544). One such technique is hot deck imputation. Roth defines this as follows:

> "Hot deck imputation is a strategy that has become popular in survey research. The underlying principle is that researchers should replace a missing value with the actual score from a similar case in the current data set." (Roth, 1994: 544)

Hot deck imputation has several advantages over the simple techniques suggested by Roth. First, it uses realistic values to replace missing data. Second, the replacement values are not means, so the variable distributions will not be distorted. Roth recommends using this approach as being "particularly helpful when data are missing in certain patterns" (Roth, 1994: 544).

Hot deck imputation also has several disadvantages. There is little theoretical or empirical work done to test its accuracy. It categorizes continuous variables, which causes a loss of robustness. Since replacement values are taken to be the same as similar data points, standard errors are difficult to estimate. Roth does not, unfortunately, offer a way to determine a similar case for replacing the missing information.

Roth also provides a maximum likelihood estimation technique. He claims, "The relatively simple approach generally assumes that the observed data are a sample drawn from a multivariate normal distribution" (Roth, 1994: 545). Though not multivariate normal, it is known that the degree distribution of the nodes in a social network follows the power law. Thus, when faced with

a network whose degree does not follow a power law, an analyst can speculate about appropriately placed arcs which would bring the degree distribution into line with the power law.

Roth recommends considering two factors when choosing a missing data technique (MDT) - amount of missing data and pattern of missing data. He claims that if the amount of missing data is small, then the choice of MDT is unimportant. As justification for this statement, he states, "Monte Carlo studies suggest there is little difference in the parameter estimates and answers to research questions when less than 10% of the data are missing in random patterns or systematic patterns" (Roth, 1994: 553). The choice becomes more critical as the amount of missing data rises to 15 or 20%.

Unfortunately, in many cases of social network analysis, the analyst may not know what percentage of the data is missing. In this situation, it is reasonable to choose a method that pervades all levels, such as Maximum Likelihood techniques or hotdeck imputations. As previously noted, if the degree distribution does not follow a power law, or the diameter of the network is large, then arcs must be appropriately inserted to meet these social network conditions.

The other factor Roth advises considering when choosing an MDT is the pattern of missing data. Data missing at random is least problematic. When the pattern of missing data is not random, it can be either across or within subgroups. The simple MDTs are highly likely to misestimate correlations. Even the hot deck imputation approach is not recommended due to the effects it has on biasing the data. Although there has not been a great deal of research done in this area, Roth believes "the expectation maximization approach shows great promise" (Roth, 1994: 559). Further, Roth states, "Statisticians might ... recommend use of expectation maximization or maximum likelihood approaches that model the missing data problem based on previous knowledge of the distributional functions" (Roth, 1994: 560). As Parsons advises, it is important to consider what is happening when the analyst fills in a missing data point. He cautions the analyst (Parsons, 1996: 356):

> "Quite a number of schemes have been proposed [to deal with the missing value], most of which center around the null value, a placeholder for the missing value... While the use of a null value seems a very sensible way of handling the problem of representing incompleteness, it introduces a new problem – interpreting what the null value is representing"

2.9 Summary

This chapter has reviewed pertinent literature relating to modeling social networks, aggregation techniques and a brief overview of methods to overcome missing information. The structures and techniques of graph theory provide social network analysis a rigorous analytic foundation. Social networks can be classified as those subset of graphs possessing some combination of the following three properties: 1. high local clustering, 2. small average diameter, and 3. power law degree distribution.

Many aggregation techniques and structures are available, but the most appealing concept lies in blockmodeling. Blockmodeling essentially attempts to find sets of nodes that are in some way similar. If the analysis at hand covers a community recreation league, then different sport teams may constitute appropriate blocks. In this research, the blockmodeling concept is translated to rearranging rows and columns, representing nodes, in the adjacency matrix to obtain blocks of 1s on the diagonal.

Yan's paper on fuzzy cliques offers techniques applicable to social network analysis, especially when information is not known with certainty or the arcs are directed. If level of relationship between individuals or groups is uncertain, Yan recommends using a fuzzy number or a membership function. In general, clique detection requires undirected arcs, but Yan relaxes many of the constrictive restraints of pure cliques in his definition of fuzzy cliques as maximal connected subgraphs. While this definition may not be appropriate for undirected networks, it is for directed networks. Further details and an example of fuzzy cliques is in Appendix A.

The node measures introduced in this chapter (degree, closeness, and betweenness centrality) are used to transform information about relationships between nodes into information about a node. Application of multiple measures on a network can provide insight to an individual or aggregated subgroup's position in the network.

Chapter 3 applies many of these techniques to offer the social network analyst a method for characterizing social networks. To assist the analyst in performing this aggregation analysis,

Chapter 3 focuses on clique (and its less dense cousin, the k-plex) detection and applying degree and closeness centrality node measures to differentiate individuals or subgroups in the network. First, two techniques for determining appropriate cliques or subgroups for aggregation are offered, using the concepts of blockmodeling and node coloring. Then, the two node measures of degree centrality and closeness centrality measures are further explored for their contribution to social network analysis. The remainder of Chapter 3 provides the methodology for aggregating the appropriately determined subgroups and applying either the degree or closeness centrality measure. The aggregation process provides the analyst with information on individual and subgroup relationships as well as individual and subgroup relative positional importance in the network, with respect to the data defined on the network relationships. Chapter 4 then applies these techniques to notional social networks and Chapter 5 uses them for real world open source data for Jema'ah Islamiyah. Chapter 6 investigates the effect of missing information on network structures and the analytic techniques presented in Chapter 3 and demonstrated in Chapters 4 and 5.

Chapter 3 - Methodology for Subgroup Detection and Node Aggregation Techniques

3.1 Introduction

The aggregation techniques in this chapter benefit not only the analysis of large networks, but of any social network in which subgroup interactions are of interest. The techniques and calculations developed in this chapter demonstrate how different aggregation methods and levels of aggregation affect the information available in the aggregated network.

Aggregation of a social network requires two steps: 1. determine appropriate sets of individuals to be aggregated into subgroups, and 2. calculate measures for the networks of aggregated nodes. Section 3.3 offers methods for determining appropriate subgroups, and Section 3.4 details the four aggregation measures and when each is appropriate. Naturally, no single aggregation technique is appropriate for all social network structures; Section 3.2 characterizes the network to determine which techniques are appropriate for a given network.

3.2 Network Structures

The first step of aggregation is to determine appropriate groupings of individuals into subgroups, and is dependent on the relationships in the network. This research considers two structures: 1. distinct subgroups, connected only through cut-outs, and 2. liaison individuals who have membership in multiple subgroups. Methods for finding appropriate subgroups for both of these network structures are detailed in the next section. If there is no previous knowledge of possible subgroups in the network, both methods may be used for exploratory purposes.

The appropriate subgroups for aggregation are determined using one of the clique detection methods offered in Section 3.3. Cliques are an important structure, since they represent a set of individuals all of whom have a direct relationship with every other member in the set. If one susceptible member of a clique can be influenced by an outsider, that member has a direct line of communication with every other person in the clique.

However, in social networks, it is common to find only a few cliques, as one missing arc means the set of nodes fails to form a clique. Furthermore, when complete information of relationships in the network is not known, the missing arcs inhibit clique detection. For the research in this thesis, it is therefore necessary to consider subgroups that relax the pure clique constraints. To facilitate finding candidate sets to aggregate for social network analysis, the methods developed in this chapter allow the analyst to find groups in which not every pair of nodes has a direct relationship. Chapter 2 offers several relaxations of the pure clique definition, including k-plexes (Seidman and Foster, 1978), combinable groups (Chang and Fung, 1989), and fuzzy cliques (Yan, 1988).

The relaxation chosen for the methods below is the k-plex, in which each pair does not have to be directly related, but each of the n nodes in the group must be adjacent to at least $n-k$ other individuals in the group. A smaller k produces a denser group; $k = 1$ yields a pure clique, since in a clique on n nodes, each node is adjacent to each of the other $n-1$ other nodes. k-plexes are an appropriate relaxation for the clique structure in which it is known information is imperfect, or some members are intentionally kept ignorant of one another. When k is still relatively small compared to n, then the k-plex on n nodes is still fairly tightly connected, and a small number of missing direct relationships does not imply the group has lost cohesion.

The second step assigns a positional weight, via a node measure, to each node in the aggregated network. Assigning positional weights to individuals or subgroups in the network provides the analyst with relative measures of those individuals or subgroups. This offers information on relative importance of each node to the measure definition. In Chapter 2, several node measures were introduced; degree centrality and closeness centrality are further explored in the remainder of the thesis.

The degree centrality measure is a local property indicating the strength of an individual's relationship with immediate neighbors. The original introduction of this measure assumes each arc has a unity weight and the network is undirected (Shaw, 1954); thus the measure assigned to the node is simply the degree of the node. In this research, the degree centrality measure is extended to allow

for non-unity weight on the arcs and for the arcs to be directed. This extended degree centrality measure may be appropriate when considering an individual's relationship with his direct contacts. For example, this measure can be used to model a supervisor's relationship with his subordinates, or the interactions involving otherwise ignored individuals such as the mail delivery person. The boss may not be aware how much influence the person who passes out the mail has on all the office employees, but the mailman's low position in terms of money and respect by the supervisors may make him a susceptible target for outside influence. Thus, the typical office structure may not provide the necessary information for exploiting the network.

The closeness centrality measure is a global property indicating the efficiency with which an individual can disseminate information or materiel throughout the network. As with the degree centrality, the original implementation of this measure assumed the arcs are undirected and have unity weight (Beauchamp, 1965). The closeness measure is also expanded in this thesis to allow non-unity weights and directed arcs. This measure is appropriate when the analyst seeks an individual who can communicate easily with the rest of the network. Consider a scenario in which individuals or subgroups in a network are about to be apprehended, and only a few members in the group have a small closeness centrality measure. If those few individuals can be incapacitated, then the network loses much of its ability to communicate quickly and efficiently, preventing individuals not yet detained from being warned of imminent threat.

Once subgroups are known and an appropriate node measures chosen, the network can be aggregated into subgroups and the degree or closeness centrality measure can be assigned. Section 3.4 provides the necessary calculations for aggregation depending on subgroup structure and the node measure used. The analysis in this thesis assumes the network under consideration has arc weights. However, if arcs are known to exist, but their weights are unknown, they can all be assigned a unity weight. Alternatively, if some weights are known, but not all, the unknown weights can be approximated by giving them a fuzzy number (see Yan, 1998) or using a maximum likelihood function (see Roth, 1994).

Weights on arcs in a network give a relational measure between two nodes, while weights on nodes are positional. Depending on the purpose of the social network analysis at hand, the network may be aggregated whether the weight is on the arcs or on the nodes, and the two steps of aggregation and assigning a node measure can be done in either order. In general, if aggregation precedes assigning node measures, more information is masked in the aggregation step. Any weight on an arc with both endpoints in the same subgroup is lost in the aggregation step. However, if a node measure is assigned to each individual before aggregating the network into subgroups, the information from each individual is carried into the aggregation step and considered when aggregating the individual nodes into aggregated nodes.

At first glance, it may seem at this point that in the interest of saving information, aggregating nodes which have already been assigned measures is preferable. However, there are circumstances when it is important to understand a network that has been aggregated with weights on the arcs. If the analyst is interested only in the relationships between subgroups, then aggregating the social network with weighted arcs (instead of weighted nodes) provides the analyst with measures that are not clouded with information from within groups.

3.3 Subgroup Detection Techniques

This section develops two techniques for determining sets of individuals to be aggregated into subgroups. The first, in Section 3.3.1, requires an individual to be in only one subgroup, whereas the second, in Section 3.3.2, allows liaison individuals as members of multiple subgroups. Traditional clustering techniques, such as those implemented in UCINet or JMP, allow an individual in only one subgroup.

Both methods developed here first seek maximal cliques. A clique is said to be maximal if no larger clique contains it. Cliques are desirable for subgroups, since every pair of individuals has a direct relationship and no structural information about the subgroup is masked in the aggregation. If all aggregated nodes are cliques, then the structure of the disaggregated network is known. However,

if some of the aggregated nodes represent k-plexes (for $k \geq 2$) the exact structure of that subgroup is not completely known.

Pure cliques are likely to be rare in social networks, especially in networks of imperfect information. Once it has been determined that aggregating only pure cliques does not produce a sufficiently small aggregated network, it may be necessary to extend the clique subgroups to k-plexes. The clique detection methods detailed in Sections 3.3.1.2 and 3.3.2 demonstrate how to find maximal cliques, and can be expanded to include k-plexes. All of these methods are only suited for undirected networks. For directed networks, Yan's method, which defines a clique to be a maximally connected subgroup is appropriate, as described in Chapter 2 and further detailed in Appendix A.

3.3.1 Non-Overlapping Subgroups

Generating a set of non-overlapping cliques is accomplished by rearranging the adjacency matrix into blocks, utilizing the blockmodeling concept introduced in Chapter 2. A set of nodes forms a block at their intersection in the adjacency matrix. If every entry in that block is a 1 (except for the diagonal entries which are all 0), then the set of nodes forms a clique.

The method for finding non-overlapping cliques seeks to blockmodel by node coloring the complement of the network. Recall that node coloring a network requires placing a color, or label, on each node such that no two adjacent nodes share the same color. The greedy method outlined in this section colors each node with the smallest color available. A color is available for node i when no adjacent node to i is already labelled with that color. Since the complement G' of the graph G uses the same node set as G, and has an arc exactly where G does not, node coloring G' by this method finds maximal cliques in the graph itself.

`Complement Coloring Algorithm:`

1. Node color an uncolored node in G' by assigning it the smallest color available.
2. Repeat Step 2 until all nodes have been colored.
3. Order the nodes in the adjacency matrix by grouping nodes together by their color.

As coded in MATLAB for this thesis, this algorithm runs in $\mathbb{O}(n^6)$

Theorem 3.1 *The Complement Coloring Algorithm partitions the node set of a graph G into non-overlapping cliques of G. If one or more nodes is in multiple cliques, this algorithm finds only one of the multiple cliques.*

Proof. It is necessary to show the following: 1. every node is in exactly one clique, and 2. the cliques are non-overlapping.

1. Every node in the network is colored exactly one color in steps 1 or 2. By proving the set of nodes colored the same color are a clique, it will be shown that every node is in exactly one clique. Without loss of generality, consider the nodes colored i. This set of nodes can be colored i since no pair of them have an arc connecting them in G'. Therefore, all pairs have a connecting arc in G, forming a clique. This is true for every color used, so the network partitions into cliques.

2. Since no node receives more than one color, the cliques cannot overlap. ■

If two cliques share a node, this methods requires that individual to be in only one group. It is the user's choice whether to remove a node from one clique to enter it in another. If there is some outside knowledge of group dynamics, then that information can be applied to determining the clique to which the individual belongs. However, if individuals are in multiple groups, then it may not be appropriate to try to aggregate the network as if the groups are distinct, such as in Figure 1 on page 36. It would perhaps be preferable use the method in Section 3.3.2, allowing an individual to have membership in multiple subgroups.

The initial run through the method may find cliques $\{1, 2, 3\}$ and $\{4, 5, 6\}$. A reexamination of the adjacency matrix yields the clique $\{2, 3, 4, 5, 6\}$. Any column of a block off the diagonal that is all 1s can be added to the clique in those rows.

As another example, consider the following adjacency matrix representing a network G :

	1	2	3	4	5	6	7
1	**0**	**1**	**1**	0	0	0	0
2	**1**	**0**	**1**	0	0	0	0
3	**1**	**1**	**0**	1	1	1	0
4	0	0	1	**0**	**1**	**1**	**1**
5	0	0	1	**1**	**0**	**1**	**1**
6	0	0	1	**1**	**1**	**0**	**1**
7	0	0	0	**1**	**1**	**1**	**0**

The coloring of G' has picked up a K_3 of nodes $\{1, 2, 3\}$ and a K_4 of nodes $\{4, 5, 6, 7\}$, where K_n denotes a clique on n nodes. Since the off-diagonal block showing the arcs connecting the K_3 and K_4 has a column of 1s, node 2 can be augmented to the K_4 and it becomes a K_5. This is accomplished at the expense of the K_3 which would then only be an arc on nodes $\{1, 2\}$. Thus, the method did determine a maximal clique of which node 2 is a member, but not the maximum one, the K_5 consisting of nodes $\{2, 4, 5, 6, 7\}$.

This method finds only pure cliques. As previously stated, cliques are likely to be rare in social networks, especially when there is not perfect information and operational security does not allow members of a group to associate freely with other members. It is therefore necessary to relax the clique definition and seek subgroups where individuals know most, but not necessarily all, of the other members in the subgroup. The amount by which the cohesiveness of the subgroup can be relaxed into k-plexes is a subjective decision that must be made by the analyst.

Now that all the cliques have been found, it is necessary to consider the possibility that many of the individuals in the network are in a clique of size 1 - themselves. When analyzing social networks, it is highly likely that not have all of the information regarding interpersonal relationships is available, or that an otherwise tightly knit group may be a couple of arcs short of a clique. To address this situation, the following calculations provide the density of groups in which members that may or may not have a direct relationship with every other individual in the subgroup are added .

Three sets of calculations are presented. The first shows augmenting a nonmaximal clique with a node adjacent to every member of the clique simply increases the size of the clique and does not decrease the density of arcs in the subgraph. This shows that removing a node from one clique to add it to another if necessary does not affect the density of either clique. The second shows how the density of the arcs in the subgroup changes when a node is added to a clique adjacent to $n - 1$ of the nodes on a clique of size n. The decrease in density is dependent only on n. The third calculation extends the second to show the decrease in density when a node is added to a clique when it is adjacent to only $n - c$ $(c > 1)$ of the nodes in the clique. The largest c among the additional

nodes to the clique determine the k coefficient in the k-plex. Let c' be the largest c; the subgraph then obtained is a $(c'+1)$-plex.

In graph theory, it is generally assumed that a clique has density one, since every arc between all distinct pairs of nodes in the subgraph exist. This definition of density allows the range of acceptable densities for a graph to fall within $[0, 1]$, where 0 density is a set of isolated nodes with no arcs, and 1 density is a full clique. As the density of the subgraph decreases, the group becomes less well connected. It is therefore often necessary to find a trade-off between density that is not too low, but which condenses the graph enough to be able to perform analysis on the aggregated network quickly, still meeting the operational requirement.

If the clique is not already maximal, adding another node to the clique (identified with a column of 1s in an off-diagonal block) has a density equal to the sum of the following:

1. the clique on n nodes out of $(n+1)^2$: $\dfrac{n^2-n}{(n+1)^2-(n+1)}$
2. the row and column for the $n+1^{st}$ node: $\dfrac{2n}{(n+1)^2-(n+1)}$

The denominator in these fractions subtracts the 0s on the diagonal of the adjacency matrix, or self loops $(n+1)$ from the total number of possible arcs $(n+1)^2$, which includes the self-loops.

$$\begin{aligned}\frac{n(n-1)+2n}{(n+1)^2-(n+1)} &= \frac{n^2-n+2n}{n^2+2n+1-n-1} \\ &= \frac{n^2+n}{n^2+n} \\ &= 1\end{aligned}$$

Since adding this node simply made a larger clique, the outcome of density 1 is expected.

If the clique is maximal, to keep the density as high as possible add a node connected to $n-1$ of the nodes in the clique. Such a node can be found if a block off the diagonal has a column of $(n-1)$ ones and one zero. By adding this node to a clique of size n, the set of nodes has density:

$$\begin{aligned}
\frac{n(n-1)+2(n-1)}{(n+1)^2-(n-1)} &= \frac{n^2-n+2n-2}{n^2+2n+1-n-1} \\
&= \frac{n^2+n-2}{n^2+n} \\
&= \frac{(n+2)(n-1)}{n(n+1)} \qquad (3.1) \\
&= 1-\frac{2}{n^2+n} \qquad (3.2)
\end{aligned}$$

Equations (3.1) and (3.2) show the density of the new subgroup. Equation (3.2) shows specifically how the density decreases from that of a pure clique when a node adjacent to only $n-1$ other nodes is added to the subgroup.

In general, adding a node connected to $(n-c)$ of the nodes in the clique of size n gives that set density:

$$\begin{aligned}
\frac{n(n-1)+2(n-c)}{(n+1)^2-(n-1)} &= \frac{n^2-n+2n-2c}{n^2+2n+1-n-1} \\
&= \frac{n^2+n-2c}{n^2+n} \qquad (3.3) \\
&= 1-\frac{2c}{n^2+n} \qquad (3.4)
\end{aligned}$$

Equation (3.3) and (3.4) show the density of the new subgroup. Equation (3.4) shows specifically how the density decreases from that of a pure clique when a node adjacent to only $n-c$ other nodes is added to the subgroup.

The clear disadvantage to finding only non-overlapping maximal cliques is that many cliques may be ignored. Consider the network on six nodes shown in Figure 1. The two maximal cliques are $\{1,2,3\}$ and $\{2,3,4,5,6\}$. Building a clique starting with arc $\{1,2\}$ produces the cliques $\{1,2,3\}$ and $\{4,5,6\}$. Alternatively, the seven node network in Figure 2 has a more appropriate structure for aggregation into non-overlapping cliques, since the subgroups are distinct. The method finds $\{1,2,3,4\}$ and $\{5,6,7\}$. It is therefore necessary to consider the network topology and the mission requirements before deciding on a clique detection technique.

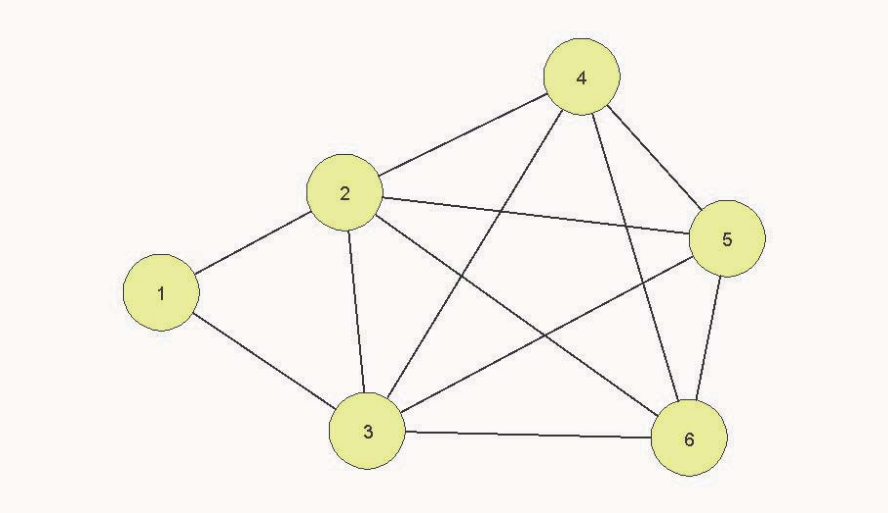

Figure 1. **Poor choice for non-overlapping clique detection method**

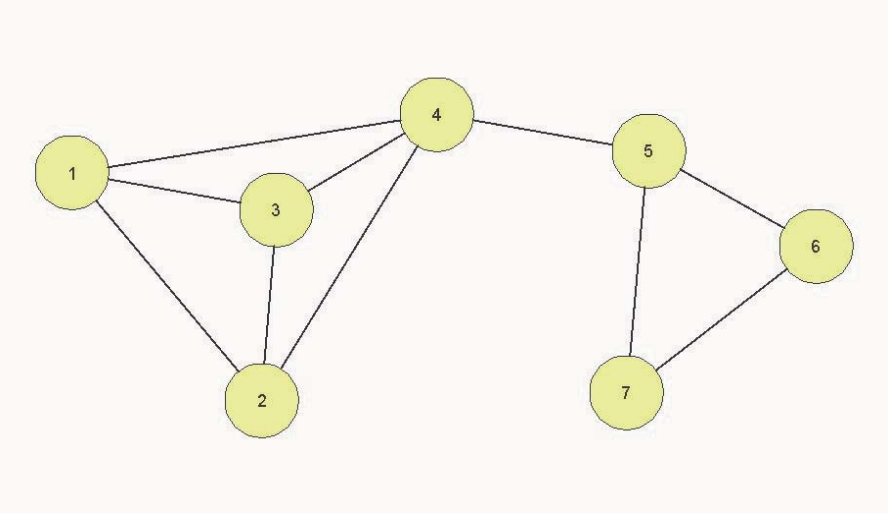

Figure 2. **Appropriate Network for Non-Overlapping Clique Method**

This implies that before selecting a method for determining which sets of nodes form appropriate aggregated subgroups, some preliminary checks of the network data must be performed. This could include producing a visual representation of the network, if possible. Any prior knowledge of intergroup relationships can be used. If nothing is known about the network or suspects the subgroups are connected through liaison individuals, then the techniques offered next for finding overlapping subgroups can provide insight to the network structure.

3.3.2 Clique Detection for Overlapping Subgroups

The method introduced in this section offered finds overlapping subgroups, appropriate when subgroups interact through liaison individuals. Seeking subgroups in this manner removes one of the concerns found in non-overlapping cliques: that some cliques may be missed entirely. It is, however, computationally more complex.

Finding every clique in a graph is a difficult problem. Every clique can be found in non-polynomial time by taking every subset of the node set and testing to see if there is an arc between every distinct pair in the subset chosen. Bron and Kerbosch developed an algorithm finding every clique in 3^k steps, where k is a number of distinct K_3s in the network (Bron and Kerbosch, 576: 1973).

By starting with the largest of the 2^N (where N is the number of nodes) subsets and checking the subsets in decreasing number of nodes, it may not be necessary to check every subset. For example, if a K_5 is found, then the subsets of the nodes in the K_5 do not also need to be examined. However, if the network is completely disconnected, there are no arcs and this method requires checking all $2^N - 1$ subsets to find N K_1s.

It will be proven that the polynomial method offered in this section finds at least 75% of the cliques in a network. The innermost loop allows for groups of lower density than a pure clique. This step can be omitted only pure cliques in the network or faster performance are desired.

`Overlapping Cliques Heuristic:`

1. Let $G = (V, E)$ be an undirected network, with node set V and arc set E .

2. Let A be the adjacency matrix for G.

3. Let $k \geq 1$ be the number of nodes each individual does not have to be directly related to (as in k-plex). $k = 1$ is a pure clique

4. Let C be a set that tracks the set of nodes in a given clique.

5. For every node $i \in G = (V, E)$

6. For every node j such that $(i, j) \in A$

7. $C = \{i, j\}$

8. As long as there exists some node $x \in V \backslash C$ such that $(x, n) \in A \ \ \forall \ \ n \in C$

9. $C = C \cup \{x\}$

10. Output C

11. if $k > 1$

12. As long as there exists some node $\{x\} \in V \backslash C$ such that $(x, n) \in A$ for at least $(n - k)$ of the nodes in C

13. $C = C \cup \{x\}$

14. End

15. Next j

16. Next i

Each output C from the method produces a set of not necessarily unique k-plex on $|C|$ nodes with a clique in its core. This research considers only the cliques and k-plexes on three or more nodes. There are two main difficulties with this method: 1. though polynomial, it can be relatively slow (especially when seeking subgraphs less dense than cliques), 2. it is not guaranteed to find every clique. The MATLAB code currently runs on this algorithm in $\mathbb{O}(n^6)$. This is a theoretical maximum number of operations that may have to be run, and decreases significantly in a sparse network.

Theorems 3.2 through 3.6 show that the heuristic is guaranteed to find at least 75% of all cliques. The method does not find a given clique when each arc in the "missing" clique is used as the initial arc in a clique (in Step 7 of the heuristic) builds some other clique. A clique can be missing when only when the union of other cliques covers every arc of the missing clique. These theorems show that it requires at least three other cliques to cover the "missing" clique. Then the

structure of the networks for which the method does not find every clique is presented, along with an example of this rare case.

Theorem 3.2 *Two maximal overlapping cliques must be unique up to a node.*

Proof. Let K_i and K_j be maximal cliques. Suppose K_i has no node unique from K_j. Then $K_i \subseteq K_j$. But $K_i$cannot be a proper subset of K_j, or K_i is not maximal, violating the assumption. The only option is for $K_i = K_j$, in which case the two maximal cliques are not unique. ■

Corollary 3.3 *Two maximal cliques are unique up to an arc.*

Proof. By Theorem 3.2, if two maximal cliques are unique up to a node, then they are unique up to at least $|K| - 1$ arcs. The smallest clique of interest is size 3, so any clique has more than one unique arc. ■

Theorem 3.4 *Two maximal unique overlapping cliques, K_i and K_j, can share no more than $\binom{|K|}{2} - (|K| - 1)$, where $K = \min(K_i, K_j)$.*

Proof. A clique of size K has $\binom{|K|}{2}$ arcs (when excluding self-loops). Since we have shown that K_i and K_j must be unique up to a node, then there is some node, v, in $K = \min(K_i, K_j)$ that is not in the larger clique. Therefore, K has at least $\deg(v) = |K| - 1$ arcs not in the larger clique. If any of those arcs were in the larger clique, then the endpoint v would also have to be in the clique, contradicting the assumption that the cliques are unique. Therefore, K_i and K_j, can share no more than $\binom{|K|}{2} - (|K| - 1)$ ■

Theorem 3.5 *Let K, K_i, K_j be three unique maximal overlapping cliques such that $|K| \leq |K_i|$ and $|K| \leq |K_j|$. Then $K \cap K_1 \cap K_2$ can contain no more than $\binom{|K|}{2} - 2(|K| - 1) + 1$ arcs.*

Proof. Suppose K contains one node, $v_i \notin K_i$. By Theorem 3.2, we know at least one such node must exist. By Theorem 3.4, we know K and K_i share no more than $\binom{|K|}{2} - (|K| - 1)$ arcs (recalling $|K| \leq |K_i|$). A second clique unique from both K and K_1, K_2, can also share no more than $\binom{|K|}{2} - (|K| - 1)$ with K. If each of K_i and K_j share this many arcs with K, then many of

those arcs are the same. They overlap on at most $|K|-2$ nodes for a total of $\binom{|K|}{2} - 2(|K|-1)+1$ arcs. ■

Theorem 3.6 *The percent of undetected unique maximal cliques is bounded below by* 75%.

Proof. The method to detect cliques starts with each arc in the network, and builds a clique from it. Corollary 3.3 states each pair of cliques is unique up to an arc, and if a clique has an arc in no other clique, then the method builds that clique. However, if each arc of a clique K is also in another clique, then the method may build the clique that is not K at each of those arcs. Thus, it is necessary to find the smallest number of cliques that can cover every arc in K.

Given the three unique maximal cliques as described in Theorem 3, K, K_i, K_j, we have shown in Theorem 3 they overlap in at most $\binom{|K|}{2} - 2(|K|-1)+1$ arcs, and each pair K, K_i and K, K_j overlap in at most $\binom{|K|}{2} - (|K|-1)$ arcs. Thus, $(K \cap K_i) \setminus K_j$ overlap in at least the following number of arcs:

$$\left[\binom{|K|}{2} - (|K|-1)\right] - \left[\binom{|K|}{2} - 2(|K|-1)+1\right] = |K|-2$$

This means each of K_i and K_j share $|K|-2$ unique arcs with K. Thus $(K \cap K_i) \cup (K \cap K_j)$ has

$$2(|K|-2) + \left[\binom{|K|}{2} - 2(|K|-1)+1\right]$$

arcs. The first part of this sum is what K_i and K_j have unique in K, while the second part is what they share in K (from Theorem 3).

$$\begin{aligned} 2(|K|-2) + \left[\binom{|K|}{2} - 2(|K|-1)+1\right] &= \frac{4|K| - 8 + |K|(|K|-1) - 4|K| + 4 + 2}{2} \\ &= \frac{|K|^2 - |K| - 2}{2} \\ &= \frac{|K|^2 - |K|}{2} - 1 \end{aligned}$$

The clique K has the following number of arcs:

$$\begin{pmatrix} |K| \\ 2 \end{pmatrix} = \frac{|K|(|K|-1)}{2} = \frac{|K|^2-|K|}{2}$$

Since

$$\frac{|K|^2-|K|}{2} - 1 < \frac{|K|^2-|K|}{2} = \begin{pmatrix} |K| \\ 2 \end{pmatrix}$$

it takes at least three cliques $\neq K$ whose union would cover K. Therefore, an upper bound on the number of undetected cliques is 1 out of 4, and the percent of detected cliques is bounded below by 75% ■

This proof also implies that the undetected clique will be totally surrounded by other cliques. Such a topology may force the method to fail to detect all cliques. In fact, the upper bound is strict, as demonstrated in Figure 3. The method may fail to find the $K_3 = \{2, 3, 5\}$. When building off of arc $(2, 3)$, it may build the $K_3 = \{1, 2, 3\}$. Similarly, building off of $(2, 5)$ and $(3, 5)$, it may find $\{2, 5, 6\}$ and $\{3, 4, 5\}$, respectively, never detecting $\{2, 3, 5\}$.

3.4 Aggregation Techniques

At this point, appropriate subgroups for aggregation in the network are known, whether from previous knowledge of the network or from one of the methods described in this chapter. Assuming that the sets of nodes to be aggregated are known, it is now necessary to choose an appropriate node measure before beginning the aggregation analysis. As previously stated, most social network analysis does not have a weight to the arcs representing the relationship between individuals (or groups). If adequate information about the strength of the relationship is not known, then the methods introduced in this section simply assign each existing arc a unity weight. McAndrew suggests there are occasions when it may be preferable to assign each arc a unity weight, even when

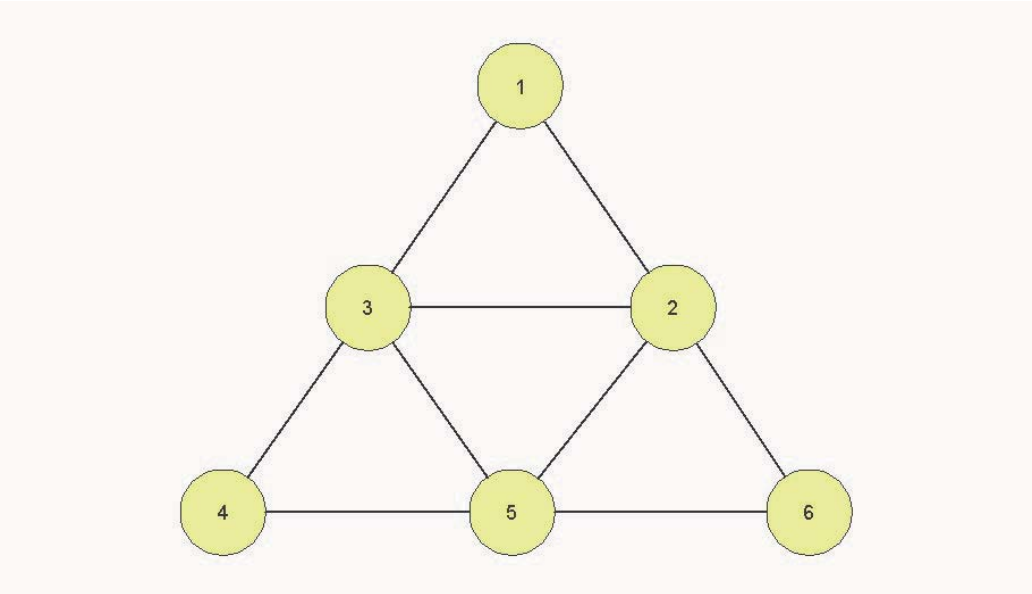

Figure 3. **Example of when the overlapping clique detection method fails to find a clique**

more information is available. He notes "discrepancies may occur due to more information being collected on ties between two individuals relative to other pairs" (McAndrew, 1999: 152).

The fact that a small number of individuals may be targets for intelligence collecting often means that there may be disproportionately more information about those few individuals. This might lead to fallaciously assigning the individuals under observation a higher positional weight than they merit. Consider a situation in which the weight on an arc is the number of phone conversations the individuals at each endpoint share. Suppose Alice and Bob are two suspected drug dealers, who know each other in multiple social contexts. Perhaps they are neighbors and attend the same school. Of their 30 conversations in one month, it is not necessarily known how many of those conversations are specifically related to drug dealing or any other illegal activity (McAndrew, 1999: 152). In each of these situations, it may be preferable to have only unity weights on the arcs, to avoid over or under-estimating weights due to imperfect information.

Alternatively, if the strength of relationships is known, then that information can be used in the analysis, providing more robust measures. These arc weights can be used to assign a measure to each node in the network. The two measures explored in this thesis, extensions of degree and closeness centrality, were chosen for their versatility and ease of use.

The degree centrality measure, a local property which assigns to each node the sum of the weights on arcs emanating out of the node, is appropriate when it is of interest to understand an individual's relationship with immediate neighbors. A person can only directly interact with immediate neighbors, and must rely on them to disseminate any information. A high measure may be indicative of an individual who communicates with many people - one who bridges the distance between scattered groups, for example. Alternatively, it may indicate an individual who has great influence, but directly interacts with only a few other individuals.

The closeness centrality measure, a global property which assigns to each node a weight equal to the sum of the lengths of shortest paths to every other node, considers the relative position an individual has to efficiently communicate with every individual in the network, not just the ones

with whom there is a direct relationship. Since the measures are based on length of the shortest path, it is important that a smaller weight on the arc represents a stronger level of relationship. If the arc weights instead has a positive correlation with the strength of the relationship of the two individuals the arc connects, then it is necessary to perform a transformation of that measure.

Any traditional transformation can be used. The one implemented in this thesis assigns to each arc the reciprocal of the weight. For example, if the weights on six nodes are $\{1, 2, 4, 4, 5.5, 7\}$, where the weight of 7 represents a stronger relationship than 5.5, then the transformation provides the following respective arc weights: $\{1, 1/2, 1/4, 1/4, 2/11, 1/7\}$. With a weight of $1/7$, that pair of individuals has the shortest path, and quickest communication where length relates to speed in the network. If any arc weight is not greater than one, then each weight can be scaled by multiplying by a factor which makes the smallest weight ≥ 1.

3.4.1 Overview of Aggregation Order

Arc weights are *relational*, representing a strength of relationship between the individuals connected by the arc, while node weights are *positional*, representing an individual's relative position in the network. Network data may be in the form of arc weights or node weights, so it is necessary to consider aggregating the network having weighted arcs or weighted nodes. Naturally, aggregating having weights on nodes or arcs provides different, but complementary, insights on the network; therefore, when possible, it is recommended to perform both. Since each can be appropriate, the definitions of aggregation provided in Sections 3.4.2-3.4.5 describe the aggregation process for either an arc weighted or node weighted network.

The strengths and weaknesses of aggregating weighted nodes or arcs are:

1. Assigning node measures before aggregating weighted individual nodes:
   Strength: This provides a smaller network with node measures. The measure is a function of an internal measure of the subgroup as well as its measure with other subgroups. The analyst can monitor all activity within the network.
   Weakness: It is difficult to distinguish how much of the measure on a node is from the weights within the aggregated node or from the weights to nodes outside of the aggregated node, as this distinction may get lost in the aggregation of the node measures.
2. Aggregate the network with arc weights before assigning node measures to the aggregated nodes
   Strength: This also provides a smaller network with node measures. In this instance, however,

the measure on the nodes is a function only of the weights between nodes in different groups. It is easier for the analyst to detect subtle intergroup activity
Weakness: Any weights between nodes within the set of nodes making up the aggregated node is lost in the process.

The strengths and weaknesses of the two methods underscore the benefit of aggregating both orders whenever possible to gain as much understand of the network structure and relationships as possible. This insight on aggregation order can now be added to the information of how to detect subgroups and which node measures are appropriate to begin the actual calculations associated with aggregating the network.

The remainder of this chapter develops the aggregation calculations for each of the node measures previously described, as well as an indication of when each might be appropriate. Each of the measures are given for when the aggregation allows overlapping subgroups or only non-overlapping subgroups. Faced with a disaggregated network of weighted arcs, it is possible to perform analysis is two orders, by first assigning to each node a measure and then aggregating, or by aggregating and then assigning to each aggregated group node a measure.

The subgroup detection methods for finding non-overlapping and overlapping subgroups, and the two node measures can be combined. Each of these four combinations is examined in the remainder of this thesis. Table 1 provides an indication of when each might be appropriate.

Table 1. **Summary Table of Four Aggregation Techniques**

	Non-Overlapping Subgroups	**Overlapping Subgroups**
Degree Centrality	- distinct subgroups - local property - additive measure	- liaison individuals membership in multiple groups - local property - additive measure
Closeness Centrality	- distinct subgroups - global property - speed or distance measure	- liaison individuals membership in multiple groups - global property - speed or distance measure

3.4.2 Non-Overlapping Subgroups, Degree Centrality Node Measure (NSDC)

The first subgroup detection method/node measure pair introduced is for the aggregation of non-overlapping subgroups and degree centrality node measure. Though the subgroup detection

methods assume the network is undirected, assigning the measure to a node does not require the arcs be undirected, and can be applied to any network. Aggregating individuals into non-overlapping subgroups is appropriate when the natural subgroups in the network are distinct, with no individual having membership in multiple subgroups. The degree centrality node measure is appropriate when a local property is of interest, as this measure encompasses only the individual and any immediate neighbors. Further, in the aggregation step, the weight of aggregated nodes or arcs is the sum of their components. Therefore, it is necessary when using this method to ensure the arc weights are additive. For example, number or frequency of phone calls in a month is an appropriate weight for this method, but speed of message traffic is not.

Equations (3.5) through (3.8) provide the definitions for network weights during aggregation. In Equations (3.5) and (3.6) each individual is assigned a degree centrality node measure and then the network is aggregated.

Let a_{ij} be the measure on the arc (i, j) and α_i be the measure on node i.

Define the measure of each individual node to be the sum of the weights on arcs out of that node:

$$\alpha_i = \sum_{j \in A(i)} a_{ij} \tag{3.5}$$

Then aggregating into subgroups, define the node measure of C_i to be the sum of individuals in the subgroup:

$$\begin{aligned} \alpha_{C_i} &= \sum_{i \in C_i} \alpha_i \\ \alpha_{C_i} &= \sum_{i \in C_i} \sum_{j} a_{ij} \end{aligned} \tag{3.6}$$

It is worth noting that the weight on an arc within a subgroup is accounted for twice in this definition for α_{C_i}. Now consider aggregating a network before assigning degree centrality node measures. In the aggregation step, the weights of arcs between two aggregated nodes C_i and C_j are added (in each direction, if the network is directed).

Define the weight of the new directed arc from C_i to C_j to be the sum of arcs connecting the subgroups:

$$a_{C_iC_j} = \sum_{\substack{i \in C_i \\ j \in C_j}} a_{ij} \tag{3.7}$$

Then in the measuring step, define the node measure on C_i, α_{C_i} to be the sum of the weights of arcs from C_i.

$$\begin{aligned} \alpha_{C_i} &= \sum_{C_j} a_{C_iC_j} \\ \alpha_{C_i} &= \sum_{\substack{i \in C_i \\ j \notin C_i}} a_{ij} \end{aligned} \tag{3.8}$$

By comparing the final sums of each of the two α_{C_i} from Equations (3.6) and (3.8), it can be seen that the order of aggregation affects the final subgroup positional node measure. Specifically, when assigning an arc measure precedes aggregation it can be seen that the final node measure on an aggregated subgroup is the sum of all arc weights incident to any individual in the subgroup. Alternatively, when aggregating a network with the arc weights, the node measure is a function only of the arc weights emanating from the set of nodes within the aggregated node to only those arcs outside of the aggregated node. Thus, it only gives an indication of the weight of relationship of one node to another, without regard to the aggregated node's internal weight. The information internal to the aggregated node is masked, and a subgroup's positional weight is all that remains. This shows a subgroup's position within the network related to other subgroups.

The rest of this subsection addresses the question of whether order matters in aggregation: Is α_{C_i} or α_{C_j} affected by whether C_i or C_j is aggregated first? It will be shown that the order of aggregation of sets of nodes into aggregated nodes representing subgroups does not affect the final node measure on the aggregated node for the aggregated defined in this NSDC method. Intuitively, it make sense that the order does not matter, since any individual node can be aggregated into at most one aggregated node, and the aggregation and measuring operations used in this study are all additive. Theorem 3.7 shows order does not matter for aggregation with weighted nodes, while Theorem 3.8 demonstrate the robustness of order when aggregating a network with weighted arcs.

Theorem 3.7 *Let each node in network N be assigned a degree centrality node measure and aggregated into non-overlapping subgroups (as in Equations (3.5) and (3.6)) The final node measure on the aggregated subgroups is not affected by the order of aggregation of the subgroups.*

Proof. Without loss of generality, by showing robustness of aggregation order on two generic aggregated subgroups, then it must be true for every pair of aggregated nodes, and therefore the entire network.

First focus on a subgraph of n nodes. Suppose each of the n nodes has already been assigned its degree centrality value as defined in Equations (3.5) and (3.6).

Let $x_1, x_2, ..., x_k$ $(k < n)$ be aggregated into the node C_1. By Equations (3.5) and (3.6),

$$\alpha_{C_1} = \sum_{x_k \in C_1} \alpha_{x_k} \tag{3.9}$$

Let $y_1, y_2, ..., y_l$ $(l + k = n$ and no $y_i = x_j)$ be aggregated into the node C_2. Then

$$\alpha_{C_2} = \sum_{y_l \in C_2} \alpha_{y_l} \tag{3.10}$$

Since no y is the same as any x, the value of α_{C_1} is independent of the calculation of α_{C_2}. Thus the order of aggregation of the two groups is irrelevant. ■

Theorem 3.8 *Let a network N be aggregated with weighted arcs such that the aggregated arcs have weight described in Equation (3.7). Then each aggregated node is assigned a degree centrality measure described in Equation (3.8). The final node measure on the aggregated subgroups is not affected by the order of aggregation of the subgroups.*

Proof. Again, let $x_1, x_2, ..., x_k$ $(k < n)$ be aggregated into the subgroup C_1 and let $y_1, y_2, ..., y_l$ $(l + k = n$ and no $y_i = x_j)$ be aggregated into the node C_2.

Aggregating C_1 first yields a graph on $l + 1$ nodes: C_1, $y_1, y_2, ..., y_l$. The weight on all arcs connecting any of the x_k to one another is now masked. Arcs connecting any of the y_l to each other do not affect the weights of the arcs between C_1 and any of the y_l. For this calculation, it is necessary only to consider arcs between C_1(ultimately the x_k) and y_l. For any y_l adjacent to

multiple x_i the weight on that arc becomes

$$a_{C_1 y_l} = \sum_{i:(x_i,y_l)\in A} a_{x_i y_l} \tag{3.11}$$

Note that if a y_l is adjacent to only one x_i, then this sum reduces to the original arc weight between x_i and y_l: $a_{C_1 y_l} = a_{x_i y_l}$

Once C_1 has been aggregated, the y_l can be aggregated to form the aggregated node C_2 and the arc weights between C_1 and any y_l sum to form $a_{C_1 C_2}$, the arc weight of the arc (C_1, C_2). Then the node measure assigned to C_1 becomes:

$$\begin{aligned} \alpha_{C_1} &= \sum_{y_l} a_{C_1 y_l} \\ &= \sum_{x_i \in C_1} \sum_{y_l \notin C_1} a_{C_1 y_l} \\ \alpha_{C_1} &= a_{C_1 C_2} \end{aligned} \tag{3.12}$$

(once the y_l have been aggregated into C_2). Similarly the node measure assigned to C_2 becomes

$$\begin{aligned} \alpha_{C_2} &= \sum_{y_l} a_{y_l C_1} \\ \alpha_{C_2} &= a_{C_2 C_1} \end{aligned} \tag{3.13}$$

Aggregating C_2 first produces a similar argument. The y_l are aggregated into the node C_2 and

$$\begin{aligned} \alpha_{C_2} &= \sum_{x_k} a_{C_2 x_k} \\ &= \sum_{y_l \in C_2} \sum_{x_k \notin C_2} a_{C_2 x_k} \\ \alpha_{C_2} &= a_{C_2 C_1} \end{aligned} \tag{3.14}$$

As in Equation (3.14), when the x_k are aggregated into the node C_1,

$$\begin{aligned} \alpha_{C_1} &= \sum_{x_k} a_{C_2 x_k} \\ \alpha_{C_1} &= a_{C_1 C_2} \end{aligned} \tag{3.15}$$

Thus it is shown that order of aggregation does affect the degree centrality node measures on the aggregated subgroups. ■

This subsection has provided the calculations for assigning degree centrality node measures for networks with distinct subgroups and additive measures. Table 2 provides a summary of the calculations as well as when this method is appropriate, what useful information comes out of it and examples of necessary arc and node data. The next section also uses the degree centrality node measure, considers a network structure allowing overlapping subgroups. Recall subgroups overlap when liaison individuals have membership in multiple groups.

3.4.3 Overlapping Subgroups, Degree Centrality Node Measure (OSDC)

The second subgroup detection method/node measure pair introduced considers the aggregation of overlapping subgroups, with degree centrality measure. Recall the subgroup detection methods assume the network is undirected. Aggregating individuals into overlapping subgroups is appropriate when one or a small number of individuals serve as liaisons between groups. These individuals are not cut-outs, as a person in multiple subgroups has a stronger relationship with each of the subgroups they connect than a cut-out would. Two examples are an Al Qaeda representative to Jema'ah Islamiyah, belonging to both groups, and a state law enforcement official on a federal task force.

Equations (3.16) through (3.19) provide the basic definitions for determining the positional weight of an aggregated node. Equations (3.16) and (3.17), assume a node measure has been assigned to each node equal to the sum of the weights on arcs out of the node. The aggregation step then assigns to each aggregated node the sum of the weights of the set of nodes aggregated into a node, say C_i.

Define the measure assigned to each individual node i to be the sum of weight on arcs out of i:

$$\alpha_i = \sum_j a_{i,j} \tag{3.16}$$

Table 2. **Summary for NSDC**

	aggregate weighted nodes	aggregate weighted arcs
appropriate when	- positional or relational data - additive data definition - distinct groups communicating through cut-outs - interest in within and inter group interactions	- relational data - additive data definition - liaison individuals have membership in multiple subgroups - interest in intergroup interactions only
examples	if starting with positional node data: - number of meetings an individual attends - number of phone calls individual initiates if starting with relational node data: same as next column	- number of emails the two individuals share - length of phone calls between individuals
calculations: - $a_{i,j}$ is the arc weight on arc (i,j) -α_i is the node measure for node i	if beginning with positional node data, skip to step 2, otherwise step 1 gives each node degree centrality positional measure 1. assign to each node, i, a degree centrality measure $\alpha_i = \sum_j a_{i,j}$ 2. aggregate into subgroups by assigning to each subgroup, C_i, a measure equal to the sum of the individuals in the subgroup: $\alpha_{C_i} = \sum_{i \in C_i} \alpha_i$	1. aggregate into subgroups by giving the aggregated arc a weight equal to the sum of arcs with exactly one endpoint in each subgroup: $a_{C_i,C_j} = \sum_{\substack{j \in C_j \\ i \in C_i}} a_{i,j}$ 2. assign to each subgroup a degree centrality measure $\alpha_{C_i} = \sum_{C_j} a_{C_i,C_j}$
input	-positional or relational additive data -weighted adjacency matrix	-relational additive data -weighted adjacency matrix
output	Step 1: individual node measures indicating individual position w.r.t the data definition Step 2: subgroup node measures indicating subgroup position and amount of activity within and between subgroups	Step 1: aggregated arc weights indicating total communication between two subgroups Step 2: subgroup node measures indicating subgroup position and intergroup activity

Then aggregate the set of nodes i into the node C_i, and define the subgroup's weight to be the sum of its member's weights:

$$\begin{aligned} \alpha_{C_i} &= \sum_{i \in C_i} \alpha_i \\ \alpha_{C_i} &= \sum_{i \in C_i} \left[\sum_j a_{i,j} \right] \end{aligned} \tag{3.17}$$

Note that α_i will contribute its weight into every aggregated node of which it is a member. Equations (3.18) and (3.19) consider the situation when the aggregation occurs on a network with weighted arcs.

Let C_i and C_j be two subgroups; $k \in C_i \cap C_j$; $j \in C_j \backslash C_i$; $i \in C_i \backslash C_j$

Define the arc weight between C_i and C_j to be the sum of arcs connecting them:

$$a_{C_i,C_j} = \sum_{\substack{i \in C_i \backslash C_j \\ k \in C_i \cap C_j}} a_{i,k} + \sum_{\substack{k \in C_i \cap C_j \\ j \in C_j \backslash C_i}} a_{k,j} + \sum_{\substack{i \in C_i \backslash C_j \\ j \in C_j \backslash C_i}} a_{i,j} \tag{3.18}$$

Thus the measure between two aggregated nodes with overlapping membership is defined to be the sum of three types of arcs: 1. those that connect individuals only within C_i to the liaisons, 2. those that connect the liaisons to the individuals only within C_j, and 3. those that connect individuals only in C_i directly with individuals only in C_j.

Then each aggregated subgroup node, C_i, is assigned a node measure, defined to be the sum of arcs out of the subgroup:

$$\alpha_{C_i} = \sum_{C_j} a_{C_i,C_j} \tag{3.19}$$

It is worth noting that the aggregated node measure presented here subsumes all weight on arcs which have both endpoints in the same subgroup and neither endpoint is in multiple subgroups. Though the information on weights within a subgroup may not be apparent in this aggregated network, if the purpose of the analysis being considered focuses primarily on interaction between groups, then this aggregation will be appropriate. The within-subgroup weights will be required to adequately depict the final positional aggregated measures in such an analysis.

Theorem 3.9 shows that the aggregation of order is robust for aggregating a network with node weights. Theorem 3.10 shows the same for when the network is aggregated with weighted arcs instead of nodes.

Theorem 3.9 *Let each node in network N be assigned a degree node measure and aggregated into overlapping subgroups (as in Equations (3.16) and (3.17)) The final node measure on the aggregated subgroups is not affected by the order of aggregation of the subgroups.*

Proof. Suppose the node set is to be aggregated into two nodes C_i and C_j. If robustness of order can be shown to be true for two non-specified aggregated nodes, then without loss of generality, it is true for any number, by considering them two at a time.

Let each node i be assigned its node measure α_i. It is assumed that the node measures have already been assigned to each node i in the network. Aggregate C_i first, as defined in Equation (3.17).

$$\alpha_{C_i} = \sum_{i \in C_i} \alpha_i \tag{3.20}$$

Aggregation of C_j without previous knowledge that at least one node $i \in C_i$ is the same as some $j \in C_j$, leads to the following inappropriate measure for α_{C_j} :

$$\alpha_{C_j} = \sum_{j \in C_j \setminus C_i} \alpha_j$$

This differs from expected the definition for α_{C_j}, since ignoring any node in multiple aggregation sets makes it impossible to add its weight to any other aggregated node after its initial aggregation into a node. However, previous knowledge of the sets to be aggregated yields the following appropriate node measure for the aggregated C_j

$$\alpha_{C_j} = \sum_{j \in C_j} \alpha_j$$

The measure for each aggregated node is precisely the sum of the measures of every node in the subgroup, whether the C_i or C_j is aggregated first. It has been shown that this is dependent upon having full knowledge of each set of nodes to be aggregated together. ■

It is dangerous, when faced with a large network, to start aggregating sets of nodes without first defining all the subgroups. If so, then the analyst has reverted to the case where the subgroups are non-overlapping and no individual is represented in more than one aggregated node. It is therefore necessary to complete the overlapping subgroup detection methodology before attempting to aggregate the network.

Theorem 3.10 *Let a network N be aggregated with weighted arcs such that the aggregated arcs have weight described in Equation (3.18). Then each aggregated node is assigned a degree centrality measure described in Equation (3.19). The final node measure on the aggregated subgroups is not affected by the order of aggregation of the subgroups*

Proof. When aggregating nodes, it is necessary to consider new arc weights between the aggregated node C and other nodes in the network

$$a_{C,j} = \sum_{\substack{i\in C \\ j\notin C \\ (i,j)\in A}} a_{i,j} + \sum_{\substack{i,k\in C \\ j\notin C' \\ (i,j)\in A}} a_{k,i} \tag{3.21}$$

Then when C' is aggregated, the new arc measure between C and C' is as follows:

$$\begin{aligned} a_{C,C'} &= \sum_{j\in C'} a_{C,j} \\ &= \sum_{j\in C'} \left[\sum_{\substack{i\in C, \\ j\in C' \\ (i,j)\in A'}} a_{i,j} + \sum_{\substack{i,k\in C \\ (i,j)\in A}} a_{k,i} \right] \end{aligned} \tag{3.22}$$

When the $\sum_{\substack{i\in C, \\ j\in C' \\ (i,j)\in A'}} a_{i,j}$ summand is broken into $\sum_{\substack{i\in C\cap C' \\ j\in C' \\ (i,j)\in A}} a_{i,j} + \sum_{\substack{k\in C\setminus C' \\ (k,j)\in A}} a_{k,j}$., then Equation (3.22) looks like Equation (3.16). This substitution gives the expected aggregated node measure to match the calculation for an aggregated node given in Equation (3.19). ■

Theorems 3.9 and 3.10 have shown that order of aggregation does not affect the final aggregated subgroup node measure. It is, however, necessary to have determined every subgroup before performing any aggregation. It has been shown that aggregating a network one subgroup at a time, without considering the fact that a node i may belong in more than one aggregating set, will not lead to consistent positional node measures on the aggregated nodes.

This subsection has provided the calculations for assigning degree centrality node measures for networks with overlapping subgroups and additive measures. Table 3 provides a summary of the calculations as well as when this method is appropriate, what useful information comes out of the aggregation proccess and examples of appropriate arc and node data.

3.4.4 Non-Overlapping Subgroups, Closeness Centrality Node Measure (NSCC)

The third subgroup detection method/node measure pair considers the aggregation of non-overlapping subgroups, with closeness centrality node measure. Aggregating individuals into non-overlapping subgroups is appropriate when the natural subgroups in the network are distinct. The closeness centrality node measure is appropriate when a global property is of interest, as this measure encompasses an individual's relationship with every other individual in the network. Since the closeness centrality is based on the length of shortest path, it is important that a smaller weight on an arc represents a stronger relationship.

In the aggregation step, the weight of aggregated nodes or arcs is the minimum of their components. If the measure is already on the nodes before aggregation, the aggregated node weight is equal to the minimum of the node measures to be aggregated. Alternatively, if the measure is on the arcs, the new arc between aggregated nodes will be the minimum of the arcs from any node in the first subgroup to any node in the second set.

Equations (3.23) through (3.26) provide the basic definitions for the weight of an aggregated node. Equations (3.23) and (3.24) consider the situation when the node measure has been assigned to each node (equal to the sum of the shortest paths to every other node) before aggregating the network into the predetermined subgroups. The aggregation step then assigns to each aggregated node the minimum of the weights of the set of nodes aggregated into a node, say C_i. As before, assume all sets to be aggregated together have been determined, and they are non-overlapping

Table 3. **Summary for OSDC**

	aggregate weighted nodes	aggregate weighted arcs
appropriate when	- positional or relational data - additive data definition - distinct groups communicating through cut-outs - interest in within and inter group interactions	- relational data - additive data definition - liaison individuals have membership in multiple subgroups - interest in between subgroup interactions only
examples	if starting with positional node data: - number of meetings an individual attends - number of phone calls individual initiates if starting with relational node data: same as next column	- number of emails the two individuals share - length of phone calls between individuals
calculations: - $a_{i,j}$ is the arc weight on arc (i,j) - α_i is the node measure for node i	if beginning with positional node data, skip to step 2, otherwise step 1 gives the node degree centrality positional measure 1. assign to each node, i, a degree centrality measure $\alpha_i = \sum_j a_{i,j}$ 2. aggregate into subgroups byassigning to each subgroup, C_i, a measure equal to the sum of the individuals in the subgroup: $\alpha_{C_i} = \sum_{i \in C_i} \alpha_i$	1. aggregate into subgroups by givingthe aggregated arc a weight equal tothe sum of arcs with exactly one endpoint in each subgroup: $a_{C_i,C_j} = \sum_{\substack{j \in C_j \setminus C_i \\ i \in C_i \setminus C_j}} a_{i,j} + \sum_{\substack{k \in C_i \cap C_k \\ i \in C_i \setminus C_j}} a_{i,k} + \sum_{\substack{k \in C_i \cap C_k \\ j \in C_j \setminus C_i}} a_{j,k}$ 2. assign to each subgroup a degree centrality measure $\alpha_{C_i} = \sum_{C_j} a_{C_i,C_j}$
input	-positional or relational additive data -weighted adjacency matrix	-relational additive data -weighted adjacency matrix
output	Step 1: individual node measures indicating individual position w.r.t the data definition Step 2: subgroup node measures indicating subgroup position and amount of activity within and between subgroups	Step 1: aggregated relational data indicating total communication between two subgroups Step 2: subgroup node measures indicating subgroup position and intergroup activity

Let d_{ij} be the distance of the shortest path from node i to node j in the network. Define the weight of node i to be the sum of shortest paths to every other individual:

$$\alpha_i = \sum_j d_{ij} \tag{3.23}$$

The positional weight of the aggregated subgroup C_i is defined to be the smallest of its members weights:

$$\alpha_{C_i} = \min_{i \in C_i} \sum_j d_{ij} \tag{3.24}$$

Equations (3.25) and (3.26) define the aggregated node measure when aggregating a network with the weights on the arcs. Define the arc measure between aggregated subgroups C_i and C_j to be the minimum of arcs connecting the subgroups:

$$a_{C_i C_j} = \min_{i \in C_i, j \in C_j} a_{ij} \tag{3.25}$$

Then the positional weight of the aggregated subgroup C_i is defined to be the sum of the shortest paths to every other subgroup:

$$\alpha_{C_i} = \sum_{C_j} d_{C_i C_j} \tag{3.26}$$

Again, it is worth noting the arc weights within an aggregated node do not add their weight in this calculation. Thus, all apparent knowledge of the internal structure of the aggregated node is lost when the aggregated network is analyzed. Of course, if the original data is stored, a subgroup of interest can be disaggregated to further explore individuals.

Theorems 3.11 and 3.12 show that the order of aggregating subgroups into aggregated nodes does not affect the aggregated node weight.

Theorem 3.11 *Order of aggregation does not affect the node measure for an aggregated node C_i when node measures are assigned to each individual before aggregation, for the aggregation as defined in Equations (3.27) and (3.28).*

Proof. Without loss of generality, showing robustness of aggregation order on two generic aggregated subgroups, shows it to be true for every pair of aggregated nodes, and therefore the entire network.

Suppose C_i is a predetermined set to be aggregated into a subgroup. Aggregate the appropriate set of nodes $i \in C_i$ into the subgroup. By the definition, C_i's measure is

$$\alpha_{C_i} = \min_{i \in C_i} \alpha_i \tag{3.27}$$

Then aggregation of the j nodes into C_j yields

$$\alpha_{C_j} = \min_{j \in C_j} \alpha_j \tag{3.28}$$

Since no i is the same as any j (recall the aggregated nodes do not overlap), the measures α_{C_i} and α_{C_j} are independent, and the order of aggregation does not affect the measures, and the theorem is shown to be true. ■

Theorem 3.12 *Order of aggregation does not affect the node measure for an aggregated node C_i when aggregation precedes assigning a node measure to each aggregated node, for the aggregation defined in Equations (3.25) and (3.26).*

Proof. First aggregate the previously determined nodes i into the aggregated node C_i. All (directed) arcs from $i \in C_i$ to any $j \in C_j \neq C_i$ become a single arc (C_i, j) in the aggregation step. The weight on that arc is now $a_{C_i,j} = \min_{i \in C_i} a_{i,j}$. Similarly, for all (directed) arcs from $j \notin C_i$ to $i \in C_i$ also become a single arc (j, C_i) in the aggregation step. The weight on that arc is now $a_{j,C_i} = \min_{i \in C_i} a_{j,i}$.

Aggregation of the set of nodes j into the subgroup C_j yields the following weighted arc from C_i to C_j :

$$\begin{aligned} a_{C_i,C_j} &= \min_{i \in C_j} a_{C_i,j} \\ a_{C_i,C_j} &= \min_{\substack{j \in C_j \\ i \in C_i}} a_{i,j} \end{aligned} \tag{3.29}$$

Similarly,

$$\begin{aligned} a_{C_i,C_j} &= \min_{j \in C_j} a_{C_i,j} \\ a_{C_i,C_j} &= \min_{\substack{j \in C_j \\ i \in C_i}} a_{i,j} \end{aligned} \tag{3.30}$$

From these arc measures, the node measure for any aggregated node C_i can be assigned as

$$\alpha_{C_i} = \sum_{C_j \neq C_i} d_{C_i,C_j} \tag{3.31}$$

Alternatively, consider aggregating C_j first. All (directed) arcs from $j \in C_j$ to any $i \notin C_j$ become a single arc (C_j, i) in the aggregation step. The weight on that arc is now $a_{C_j,i} = \min_{j \in C_j} a_{j,i}$. Similarly, for all (directed) arcs from $i \notin C_j$ to $j \in C_j$ also become a single arc (i, C_j) in the aggregation step. The weight on that arc is now $a_{i,C_j} = \min_{j \in C_j} a_{i,j}$.

Aggregation of the set of nodes i into the node C_i yields the following weighted arc from C_j to C_i :

$$\begin{aligned} a_{C_i,C_j} &= \min_{i \in C_i} a_{i,C_j} \\ a_{C_i,C_j} &= \min_{\substack{j \in C_j \\ i \in C_i}} a_{i,j} \end{aligned} \tag{3.32}$$

Similarly,

$$\begin{aligned} a_{C_i,C_j} &= \min_{j \in C_j} a_{C_i,j} \\ a_{C_i,C_j} &= \min_{\substack{j \in C_j \\ i \in C_i}} a_{i,j} \end{aligned} \tag{3.33}$$

From these arc measures, the node measure for any aggregated node C_i can be assigned as

$$\alpha_{C_i} = \sum_{C_j \neq C_i} d_{C_i,C_j} \tag{3.34}$$

All arc weights between the aggregated nodes C_i and C_j are the same whether C_i is aggregated before C_j or after. Therefore, order of aggregation under the NSCM method as defined does not affect the node measure for an aggregated node C_j when aggregation precedes assigning a node measure to each aggregated node. ■

The next question investigated is whether the node with minimum closeness centrality measure is in the aggregated subgroup with the smallest measure. A positive answer to this question implies that finding the individual of smallest measure in a large network can be found quickly. First the

analysis can be run quickly on the aggregated network to identify the subgroup of minimum measure. Then that subgroup can be disaggregated to find an individual of small measure. Alternatively, suppose information is known only about a subgroup, and not an individual. Then a positive answer to the question would help find an individual of small measure.

Suppose the person who can disseminate information most quickly is in the group with the smallest measure. Given only data on an aggregated network, an analyst can focus further intelligence resources only on individuals known to be in the aggregated node to learn more about those individuals and possibly others in this group that show such efficiency in communication. Even if the individual with smallest measure is not in the subgroup of smallest measure, the analysis still reveals some insight of a subgroup's relative position in the network. Theorem 3.13 and Remark 1 give results of whether the node with minimum measure is indeed in the aggregated node of minimum measure for aggregating a network with weights on nodes or arcs, respectively.

Theorem 3.13 *The node of minimum node measure is in the clique of minimum measure when node measures are assigned before individual nodes are aggregated.*

Proof. By contradiction: Suppose j is the node of smallest measure in the network, but it is aggregated into an aggregated node, C_j, that does not have the smallest measure. Instead, let some other aggregated node, say S, has smallest measure assigned to it in the aggregation step.

Let j be the node satisfying $\min_i \alpha_i$; $j \in C_j$; S be $\min_{C_i} \alpha_{C_i}$ Then $\alpha_{C_j} > \alpha_S$. Since α_j is the smallest in the network, it is also the smallest in any subset of the network, specifically in C_j. By the definition of the aggregation step $\alpha_{C_j} = \alpha_j$.

$\min_{i \in S} \alpha_i = \alpha_S < \alpha_{C_j} = \alpha_j = \min_i \alpha_i \leq \min_{i \in S} \alpha_i$, which is impossible. $\Longrightarrow\Longleftarrow$

It is therefore true that the node of minimum node measure is in the clique of minimum measure when node measures are assigned before individual nodes are aggregated, and property 3 is shown to be true. ■

Remark 1 *It is not in general true that the individual with the smallest closeness centrality measure will be in the aggregated node with the smallest measure when aggregation precedes measure, as measuring and aggregation operation have been defined.*

In Figure 4, node A has a smaller measure than node B, but C_A (the K_3 containing A) has a greater measure than C_B (the K_3 containing B) after aggregation. The weight of M on an arc is a number sufficiently large that it will not be used in any shortest path.

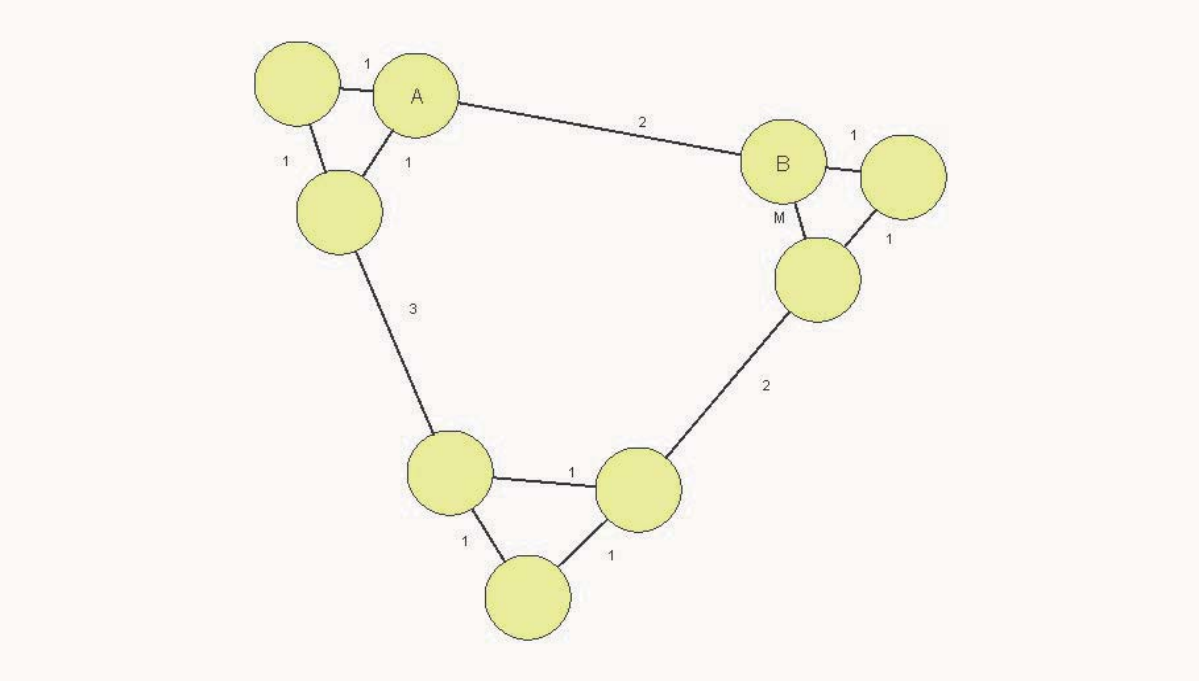

Figure 4. **Example of the individual of minimum measure not in the aggregated subgroup of minimum measure**

Remark 1 shows that when aggregating a network with weighted arcs and closeness centrality node measure as defined in this section, the individual with minimum measure is not guaranteed to be in the aggregated subgroup of minimum measure. However, if the analyst is interested only in a subgroup's relative position in the network, and unconcerned with an individual, then aggregation of the network with weighted nodes is appropriate. For example, when trying to determine communication time between terrorist cells in the network (where speed of communications is related to path length), then only the weights between subgroups are of interest.

When aggregation preceded assigning a node measure, many strict requirements must be made on the network structure in order to assure the individual of smallest measure is in the subgroup of smallest measure. Equations (3.35) through (3.38) demonstrate the necessary constraints on network

structure. There are many parts of the network structure that can cause the individual of smallest measure to be outside of the subgroup of smallest measure. One such structure occurs when the weights between members in a subgroup are substantially larger than the weights between members in different subgroups. This is unlikely in a social network, however, since members generally have stronger relationships with other individuals in their own group. Another structure occurs when the subgroup of smallest measure, say C_i, has many leaves, individuals of degree 1. The path distance to an overabundant number of individuals of degree 1 in C_i from another subgroup, say C_j, inflates the individual weights of C_j members.

Consider the node measure of a generic node i. Let i be in the set of nodes to be aggregated into the node C; $d_{C\max}$ be the length of the longest shortest path in C; $\Delta = \max_C d_{C\max}$; $p_{C,K}$ be the set of nodes on the shortest path from C to K, not including C and K.

In closeness centrality, recall an individual's node weight is the sum of shortest paths to every other node. This weight, $d_{i,j}$ can be broken into the distance to nodes either in the same subgroup, or to those in other subgroups, as seen in Equation (3.35).

$$\begin{aligned} \alpha_i &= \sum_j d_{i,j} \\ \alpha_i &= \sum_{j\in C} d_{i,j} + \sum_{j\notin C} d_{i,j} \end{aligned} \tag{3.35}$$

The shortest path from i to any other node in the subgroup is bounded above by $d_{C\max}$. Since $i \in C$, its subgroup has $|C|-1$ other nodes in it. Therefore the sum of shortest paths from i to every other node in the same subgroup is bounded above by the calculation seen in Equation (3.36).

$$\sum_{j\in C} d_{i,j} \leq (|C|-1)d_{C\max} \tag{3.36}$$

The shortest path from i to individuals in other subgroups consists of two parts: 1. arcs within i's subgroup and 2. arcs outside of i's subgroup to the final node. Equation (3.37) shows the calculation for the upper bound of the shortest paths from i to nodes outside of i's subgroup, C. The first term represents the portion of the path necessary to get out of the subgroup, C, and the

second represents the portion of the path outside of C. This second term finds an upper bound which considers three terms: 1. the distance on paths from C to K, the subgroup of the endpoint of the path, 2. the longest shortest path in the subgroup of the endpoint of the path, and 3. any path between subgroups between C and K.

$$\sum_{j \notin C} d_{i,j} \leq \sum_{j \notin C} d_{C\,\max} + \sum_{K \neq C} |K| \left(d_{C,K} + d_{K\,\max} + \sum_{K' \in p_{C,K}} d_{K'_{\max}} \right) \tag{3.37}$$

Equations (3.36) and (3.37) are substituted into Equation (3.35). The equations are then simplified and given an upper bound in the next series of equations:

$$\begin{aligned}
\alpha_i &\leq (|C|-1)d_{C_{\max}} + \sum_{j \notin C} d_{C_{\max}} + \sum_{K \neq C} |K| d_{C,K} + \sum_{K \neq C} |K| d_{K_{\max}} + \sum_{K \neq C} \sum_{K' \in p_{C,K}} |K| d_{K'_{\max}} \\
&= (|C|-1)d_{C_{\max}} + (|G|-|C|)d_{C_{\max}} + \sum_{K \neq C} |K| d_{C,K} + \sum_{K \neq C} |K| d_{K_{\max}} + \sum_{K \neq C} \sum_{K' \in p_{C,K}} |K| d_{K'_{\max}} \\
&\leq (|C|-1)\Delta + (|G|-|C|)\Delta + \sum_{K \neq C} |K|\Delta + \sum_{K \neq C} |K| d_{C,K} + \sum_{K \neq C} \sum_{K' \in p_{C,K}} |K|\Delta \\
&= (2|G|-|C|-1)\Delta + \sum_{K \neq C} \sum_{K' \in p_{C,K}} |K|\Delta + \sum_{K \neq C} |K| d_{C,K}
\end{aligned} \tag{3.38}$$

Note if (C, K) is an arc in the aggregated network, then $p_{C,K} = \phi$ and $\sum_{K' \in p_{C,K}} d_{K'_{\max}} = 0$

It can be seen that the measure on an individual node is dependent on each of the following:

1. the size of every set of nodes to be aggregated
2. the length of the longest shortest path in every set of nodes
3. the shortest paths between the aggregated nodes

It therefore would require some complicated, strict requirements to ensure the individual node of smallest measure is in the aggregated node of smallest measure. Though it has become apparent that the individual with the smallest sum of shortest paths is not necessarily within the group of smallest measure, analysis of the aggregated network does provide information on what subgroups have the smallest closeness centrality measure. Aggregation of the network using this global measure offers information on how quickly or efficiently individuals or subgroups can disseminate information or materiel throughout the network (depending on the defined arc data).

Table 4. **Summary for NSCC**

	aggregate weighted nodes	aggregate weighted arcs
appropriate when	- positional or relational data - node weights representing speed or distance - distinct groups communicating through cut-outs - interest in within and inter group interactions	- relational data - arc weights representing speedor distance - liaison individuals have membership in multiple subgroups - interest in intergroup interactions only
examples	if starting with positional node data: - ability of an individual to influence everyone else in the network - speed of sending a message to every other individual if starting with relational node data: same as next column	- time for an email to pass between individuals - cost of transmitting goods between two locations
calculations: - $a_{i,j}$ is the arc weight on arc (i,j) -α_i is the node measure for node i - $d_{i,j}$ is the shortest path from i to j	if beginning with positional node data, skip to step 2, otherwise step 1 gives the node closeness centrality positional measure 1. assign to each node, i, a closeness centrality measure $\alpha_i = \sum_j d_{i,j}$ 2. aggregate into subgroups by assigning to each subgroup, C_i, a measure equal to the min of the individuals in the subgroup: $\alpha_{C_i} = \min_{i \in C_i} \alpha_i$	1. aggregate into subgroups bygiving the aggregated arc a weightequal to the min of arcs withexactly one endpoint in eachsubgroup: $a_{C_i,C_j} = \min_{\substack{i \in C_i \\ j \in C_j}} a_{i,j}$ 2. assign to each subgroup a closeness centrality measure $\alpha_{C_i} = \sum_{C_j} d_{C_i,C_j}$
input	-positional or relational data -weighted adjacency matrix	-relational additive data -weighted adjacency matrix
output	Step 1: individual node measures indicating individual position w.r.t the data definition Step 2: subgroup node measures indicating subgroup position and amount of activity within and between subgroups	Step 1: aggregated relational data indicating minimum communication between two subgroups Step 2: subgroup node measuresindicating minimal speed ofinteractions between subgroups

Table 4 provides summary information for the NSCC method. The next section this measure for a network structure in which liaison individuals have membership in multiple subgroups.

3.4.5 Overlapping Subgroups, Closeness Centrality Node Measure (OSCC)

The fourth subgroup detection method/node measure pair considers the aggregation of overlapping subgroups, with closeness centrality measure. Equations (3.39) through (3.42) provide the basic definitions for the weight of an aggregated node. Equations (3.39) and (3.40) assume the node measure has been assigned to each node, i, equal to the sum of the shortest path from i to every other node in the network. The aggregation step then assigns to each aggregated node the minimum of the weights of the set of nodes aggregated into a node, say C_i.

Define the node measure for a node to be the sum of shortest paths to all other individuals:

$$\alpha_i = \sum_j d_{ij} \tag{3.39}$$

The aggregated subgroup has measure defined to be the minimum of its members weights:

$$\begin{aligned} \alpha_{C_i} &= \min_{i \in C_i} \sum_j d_{ij} \\ \alpha_{C_i} &= \min_{i \in C_i} \alpha_i \end{aligned} \tag{3.40}$$

Any node i in multiple aggregated nodes will be considered when calculating the minimum measure for every aggregated node in which it is a member. Alternatively, when aggregating the network with weighted arcs, the calculations are slightly more complex.

Define the arc between aggregated subgroups C_i and C_j to be the minimum of arcs connecting the two subgroups:

$$a_{C_i,C_j} = \min \left\{ \begin{array}{ccc} \min\limits_{\substack{i \in C_i \setminus C_j \\ k \in C_i \cap C_j}} a_{i,k}, & \min\limits_{\substack{i \in C_i \setminus C_j \\ j \in C_j \setminus C_i}} a_{i,j}, & \min\limits_{\substack{k \in C_i \cap C_j \\ j \in C_j \setminus C_i}} a_{k,j} \end{array} \right\} \tag{3.41}$$

The positional measure on the aggregated subgroup C_i is defined to be the subgroup's sum of shortest paths to every other subgroup:

$$\alpha_{C_i} = \sum_{C_j} d_{C_i C_j} \tag{3.42}$$

Theorems 3.14 and 3.15 examine whether order of aggregation affects the final measure on the aggregated nodes. These theorems show robustness of order of aggregation on networks that are aggregated with weights on the nodes and the arcs, respectively.

Theorem 3.14 *Order of aggregation does not affect the node measure for an aggregated node C when node measures are assigned to each individual before aggregation, when sets of nodes to be aggregated are known before any aggregation begins, and the node measures for each individual node can be stored.*

Proof. First assign each node i its node measure as defined in Equation 3.39:

$$\alpha_i = \sum_j d_{ij}$$

Suppose there are at least two sets of nodes to be aggregated into C_i and C_j. Aggregate first all nodes $i \in C_i$. Then $\alpha_{C_i} = \min_{i \in C_i} \alpha_i$.

If any $j \in C_j \neq C_i$ is adjacent to any node $i \in C_i$, it is important to store the node measure information for those individual nodes i in aggregated nodes that have already been aggregated.

When aggregating the individual nodes into C_j, it is necessary to consider not only the individual nodes on the remaining network, but any liaison node i that was aggregated into C_i that also belongs in C_j. Assuming all sets to be aggregated together are already known, let $\alpha_{C_j} = \min_{j \in C_j} \alpha_{j'}$.

Suppose instead that the individual nodes j are aggregated into C_j before C_i.

$$\alpha_{C_j} = \min_{j \in C_j} \alpha_j$$

It must be noted that if the sets of nodes to be aggregated together for every aggregated node are not known before any aggregation begins, then it is not guaranteed that order of aggregation will be robust. This implies it is necessary to have completely determined the sets of nodes to be aggregated together into possibly overlapping subgroups before beginning the aggregation step in certain types of analysis. ■

Theorem 3.15 *Order of aggregation does not affect the node measure for an aggregated node C when aggregating a network with weights on the arcs, when the subgroups are known before any aggregation begins, and the node measures for each individual node can be stored. Again, this is true only when*

the sets of individual nodes will be aggregated together are known and have been stored before any aggregation begins.

Proof. Aggregate C_i first. Then any arc between a node $i \in C_i$ and any $j \notin C_i$ becomes

$$a_{C_i,j} = \min_{i \in C_i} a_{i,j}$$

When aggregating C_j, if the arc measures between $i \in C_i$ and $k \in C_i \cap C_j$ have been stored, then it is possible to assign the following arc measure:

$$a_{C_i,C_j} = \min \left\{ \begin{array}{ccc} \min\limits_{\substack{i \in C_i \setminus C_j \\ k \in C_i \cap C_j}} a_{i,k}, & \min\limits_{\substack{i \in C_i \setminus C_j \\ j \in C_j \setminus C_i}} a_{i,j}, & \min\limits_{\substack{k \in C_i \cap C_j \\ j \in C_j \setminus C_i}} a_{k,j} \end{array} \right\}$$

Aggregation of C_j first gives the following measure for the aggregated arc between any $j \in C_j$ and any $i \notin C_j$

$$a_{i,C_j} = \min_{j \in C_j} a_{i,j}$$

When aggregating C_i, if the arc measures between $j \in C_j$ and $k \in C_i \cap C_j$ are known, then it is possible to assign the following arc measure:

$$a_{C_i,C_j} = \min \left\{ \begin{array}{ccc} \min\limits_{\substack{i \in C_i \setminus C_j \\ k \in C_i \cap C_j}} a_{i,k}, & \min\limits_{\substack{i \in C_i \setminus C_j \\ j \in C_j \setminus C_i}} a_{i,j}, & \min\limits_{\substack{k \in C_i \cap C_j \\ j \in C_j \setminus C_i}} a_{k,j} \end{array} \right\}$$

Thus if the individual arc information is stored and the sets to be aggregated are already known, then robustness of order does occur. ■

Alternatively, if the information is not stored or the sets are not known, then the order of aggregation does affect the final arc measure. Therefore, if the sets are not known before aggregation begins, it is important to carefully select which nodes to aggregate, since this reverts to the case in which an individual node cannot be in more than one aggregated node.

Table 5 provides a summary information for the OSCC method. All the methods demonstrated require perfect knowledge of the network's topology and measures. It was mentioned in Chapter 2 that the latter problem can be appeased by using fuzzy measures, if relationships are known to exist, but the level of the relationship is uncertain. (See Appendix A for more information and examples.)

Table 5. **Summary for OSCC**

	aggregate weighted nodes	aggregate weighted arcs
appropriate when	- positional or relational data - node weights representing speed or distance - distinct groups communicating through cut-outs - interest in within and inter group interactions	- relational data - arc weights representing speed or distance - liaison individuals have membership in multiple subgroups - interest in between group interactions only
examples	if starting with positional node data: - ability of an individual to influence everyone else in the network - speed of sending a message to every other individual if starting with relational node data: same as next column	- time for an email to pass between individuals - cost of transmitting goods between two locations
calculations: - $a_{i,j}$ is the arc weight on arc (i,j) -α_i is the node measure for node i - $d_{i,j}$ is the shortest path from i to j	if beginning with positional node data, skip to step 2, otherwise step 1 gives the node closeness centrality positional measure 1. assign to each node, i, a closeness centrality measure $\alpha_i = \sum_j d_{i,j}$ 2. aggregate into subgroups by assigning to each subgroup, C_i, a measure equal to the min of the individuals in the subgroup: $\alpha_{C_i} = \min_{i \in C_i} \alpha_i$	1. aggregate into subgroups by giving the aggregated arc a weightequal to the min of arcs with exactlyone endpoint in each subgroup: $a_{C_i,C_j} = \min\{\min_{\substack{i \in C_i \setminus C_j \\ j \in C_j \setminus C_i}} a_{i,j}, \min_{\substack{i \in C_i \setminus C_j \\ k \in C_i \cap C_j}} a_{i,k}, \min_{\substack{k \in C_i \cap C_j \\ j \in C_j \setminus C_i}} a_{k,j}\}$ 2. assign to each subgroup a closeness centrality measure $\alpha_{C_i} = \sum_{C_j} d_{C_i,C_j}$
input	-positional or relational data -weighted adjacency matrix	-relational additive data -weighted adjacency matrix
output	Step 1: individual node measures indicating individual position w.r.t the data definition Step 2: subgroup positional data indicating subgroup's position and amount of activity within and between subgroups	Step 1: aggregated relational data indicating minimum communication between two subgroups Step 2: subgroup positional data indicating minimal speed ofinteractions between subgroups

However, if it is uncertain whether a relationship exists or not, different techniques must be used to work around the missing information.

3.5 Summary

This chapter has provided the necessary information for the two steps in aggregating a network: 1. determine appropriate subgroups for aggregation, and 2. select an appropriate node measure for the calculations in aggregation. The subgroups detection methods, detailed in Section 3.3, allow the subgroups to be distinct, connected only through cut-outs, or overlapping, with liaison individuals having membership in multiple subgroups. The node measures utilized in this thesis are an extended degree centrality and closeness centrality measures. These two measures encompass a variety of data possibilities, such as speed, distance, and counts. Then calculations for each combination of subgroup structures and node measures are described in Section 3.4

Chapter 4 demonstrates the four techniques with notional networks, partially generated using the method in Appendix B. Further, open source information on Jema'ah Islamiyah is analyzed in Chapter 5 using appropriate techniques to gain insight to the social network. Then Chapter 6 offers an exploration of how imperfect information (in the form of missing arcs) affects subgroup detection.

Chapter 4 - Demonstration of Methodology

4.1 Introduction

Chapter 4 demonstrates the methodology detailed in Chapter 3 through four notional scenarios. These scenarios are tailored to each of the four techniques introduced in Sections 3.4.2 - 3.4.5. The networks for each of the four scenarios are partially generated from the method presented in Appendix B. The aggregation process requires identification of subgroups which provides structural information about the network. The nature of the subgroups offers information on whether the network is split into distinct cells with cut-outs providing the link between subgroups, or whether subgroups have merged and some liaison individuals are full members of several groups. The density of subgroups can give insight to the cohesiveness of each of the subgroups, suggesting which groups are most easily infiltrated or otherwise exploited. Individuals that analysis shows not to be a subgroup member may also be targets of influence. These are individuals that are perhaps new to the network, whose views can still be changed. Alternatively, it may be an individual who was once a trusted member, and therefore knows a great deal of information, but is no longer in a position of direct authority. The individual's discontent at falling to the fringe of the network may make such an individual a potential person of interest for possible influencing.

The first two scenarios examine a social network with distinct subgroups, while the latter two examine a second social network allowing individual membership in multiple subgroups. For each of the scenarios, the arc weights are randomly assigned using a uniform distribution, in which individuals within the same subgroup are assumed to have a stronger relationship than any relationship between individuals in different subgroups. Details on the values used for the distributions can be found in each section.

As stressed in Chapter 3, the aggregation step can be accomplished with weighted nodes or weighted arcs, revealing different insight into network activity. The methodology detailed in Chapter 3 use aggregation of weighted nodes to show total activity in the network, both within and between subgroups, while aggregation of weighted arcs reveals more subtle insight to the interactions only

between subgroups. Each of the four scenarios in this chapter mirrors the calculations in Chapter 3, and performs aggregation both with weighted nodes and weighted arcs. All calculations were performed using a script coded in MATLAB on a 2 ghz Dell Dimension XPST800 Pentium 4 Mobile.

4.2 Notional Network of Distinct Subgroups

Figure 5 shows the network structure used as the notional example for the demonstration of techniques in this scenario. A close look at the network appears to show several distinct subgroups: {1,2,3,4,5,6,7}, {9,10,11,12,13,14,15,16},{30,31,32,33,34,35},{36,37,38,39,40}. Only one of these subgroups is actually a clique; subgroup detection will find all cliques first and then extend them to k-plexes. At an initial review, these node sets seem to be good candidates for subgroups, but a more rigorous demonstration of subgroup detection is in Section 4.2.1.1. Furthermore, though most of the subgroups communicate only through cut-outs, it is assumed node 32 does not practice good OPSEC by ignoring the rules against direct communication between subgroups.

In the initial look at the network, it seems no single subgroup is too far removed from the others, and each can receive information or materiel from several other subgroups. Even if a small number of cut-outs are removed from the network, information and goods can still flow, though perhaps with less efficiency. Thus, this network shows good connectivity. There are two individuals (represented by nodes 15 and 16) that are on the periphery of the network. They do not appear to belong in a subgroup, nor do they act as cut-outs to any known subgroup. It is possible that they connect this network to another and information about them is incomplete. Alternatively, it is possible that they are new members or fringe members to the network. Either way, while these individuals do not play a central part in the aggregation analysis, gaining more information about them may be worthwhile.

4.2.1 Subgroup detection for network with distinct subgroups

Application of the techniques for finding cliques non-overlapping cliques in Section 3.3 yields the following cliques on three or more nodes: {1,2,3,4,5}, {9,10,11,12,13}, {21,22,23}, {21,22,24} {30,31,32,34}, and {36,37,38,39,40}. The performance time for the clique detection was 0.06 seconds.

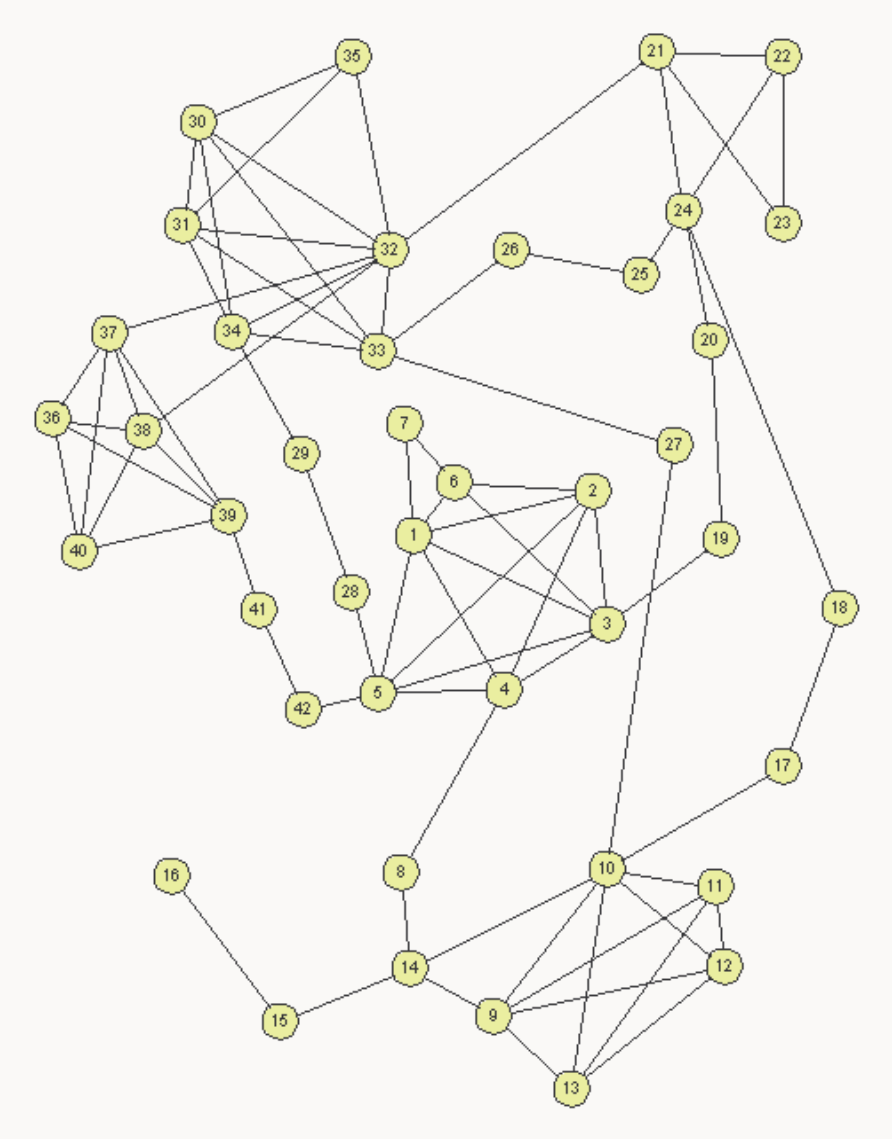

Figure 5. **Disaggregated network structure for NSDC**

Each of these cliques is part of a subgroup identified visually in the introduction to Figure 5. However, most of the subgroups clearly contain more members than those in just the pure cliques. It therefore appears promising to relax the pure clique constraint and extend the cliques found to k-plexes. The value of k is dependent on the density of the subgroups present, and how the operational setting suggests the density of a cell may be. $k = 1$ requires every individual in the group to be adjacent to all but one individual in the group (themselves). Thus $k = 1$ finds only the pure cliques again. As k increases, the density of the subgroup decreases, and the number of individuals in the subgroup increases. It is important at each increase in k to ensure no members outside of the subgroup are falsely entering. This may be difficult in some situations, requiring a review and analysis of the groups formed. In this notional scenario, however, the subgroups are distinct, and it is clear which individuals are in a particular subgroup, which individuals are cut-outs who pass information and goods from one group to another, and which are peripheral members.

Table 6. **k-plexes for network of distinct subgroups**

k	k-plexes	computational time
2	$\{1,2,3,4,5\}$ $\{9,10,11,12,13\}$ $\{21,22,23,24\}$ $\{30,31,32,33,34\}$ $\{36,37,38,39,40\}$	0.20 sec
3	$\{1,2,3,4,5,6\}$ $\{9,10,11,12,13\}$ $\{21,22,23,24\}$ $\{30,31,32,33,34,35\}$ $\{36,37,38,39,40\}$	0.20 sec
4	$\{1,2,3,4,5,6\}$ $\{9,10,11,12,13,14\}$ $\{18,21,22,23,24\}$ $\{30,31,32,33,34,35\}$ $\{32,36,37,38,39,40\}$	0.21 sec

The k-plexes for $k = 2$ to $k = 4$ are summarized in Table 6. By $k = 4$, the k-plexes start inserting non-subgroup members into the k-plex. Consider the 4-plex {32,36,37,38,39,40}. An examination of Figure 5 shows that nodes {36,37,38,39,40} clearly form a K_5, a clique on five nodes, while node 32 is in another subgroup. Although the individual represented by node 32 does communicate directly with two members in the K_5, node 32 is not actually a member of the

subgroup. The appropriateness of other 4-plexes is not so clear. Consider, for example, the 4-plex {9,10,11,12,13,14}. It is not immediately obvious whether node 14 functions as a cut-out, merely shuttling information between the two subgroups, or if 14 is actually a member of the group {9,10,11,12,13}, who simply does not deal directly with several members of the group. In either case, node 14 is not as close to {9,10,11,12,13} as they are to each other.

This process shows that determining which individuals belong together in subgroups is not always clear, even when the subgroups are non-overlapping. The techniques offer options, but there are no definitive rules for every situation, and the analyst must make the decision on which is most appropriate for the problem at hand. Any other information known about a network can be used when determining subgroups. All further analysis of this network structure uses the 3-plexes to represent subgroups. Figure 6 shows the aggregated network structure.

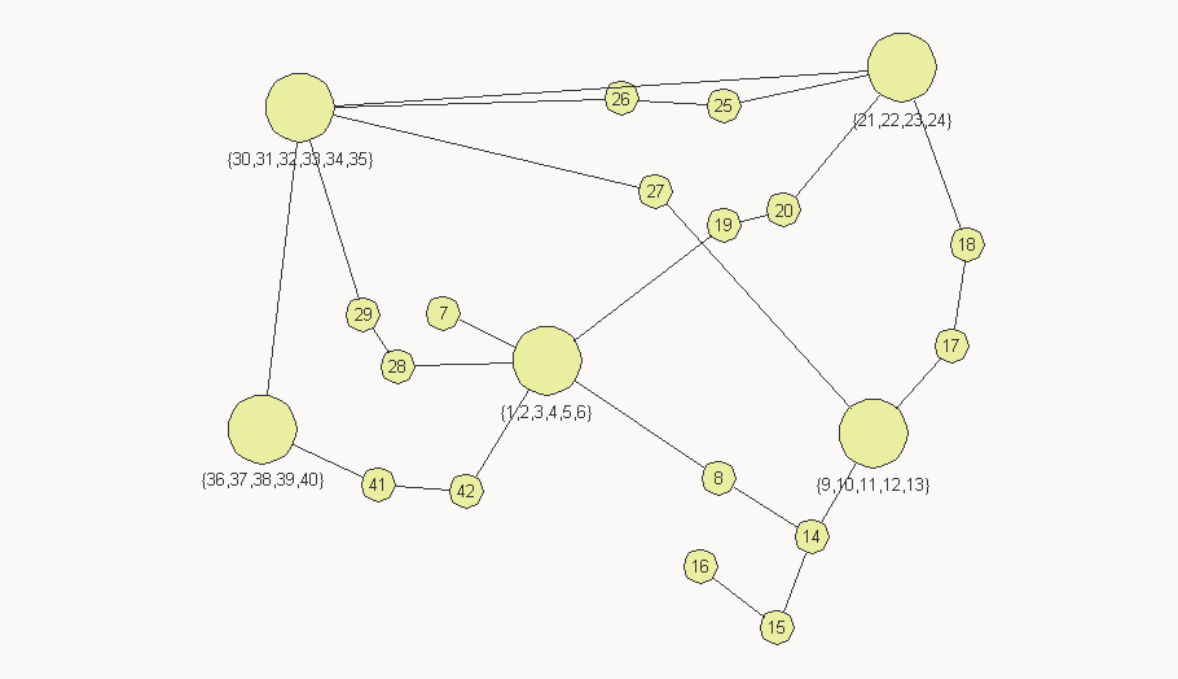

Figure 6. **Structure of aggregated 3-plex network for NSDC**

4.2.2 Scenario 1: Non-overlapping subgroups, Degree centrality (NSDC)

Sleeper cells are starting to become active and interact with one another. Intelligence data has been gathered on length of communications. It is assumed in this scenario that a higher weight on an arc represents a greater relationship. The cells are distinct groups, connected mostly through cut-outs. Though the bulk of the communications between subgroups occurs only through cut-outs, some members do communicate directly with individuals in another cell. It is not known specifically why some individuals are communicating directly, but two possibilities are that they share family ties or attended training together and have since kept in touch.

The additive nature of the data makes the degree centrality node measure appropriate. Since for this scenario, the separate groups are just starting to come together and are still distinct, an individual is assumed to have membership in only one aggregated subgroup. Each arc has been assigned a notional weight, corresponding to the length of the endpoints' shared communications. An arc within a subgroup was assigned an arc weight randomly from a uniform distribution on [6, 10]. Arcs in which both endpoints are not in the same subgroup,were assigned a weight randomly from a uniform distribution on [1, 4]. Weighting the arcs in such a manner represents the situation in this scenario that relationships within subgroups are stronger than relationships between individuals in different subgroups. Figure 7 shows these weights. Table 7 summarizes of the calculations and interpretations of the aggregation for NSDC.

Table 7. **Summary of Calculations for NSDC**

Aggregation Order	Calculations	Interpretation
Assign node measures then aggregate	$\alpha_i = \sum_j a_{ij}$	Individual's total length of communication
	$\alpha_{C_i} = \sum_{C_j} a_{C_i C_j}$	Total subgroup communication within and between subgroups
Aggregate then assign node measures	$a_{C_i C_j} = \sum_{\substack{i \in C_i \\ j \in C_j}} a_{ij}$	Total length of communication between subgroups
	$\alpha_{C_i} = \sum_{C_j} a_{C_i, C_j}$	Total subgroup communication between subgroups only

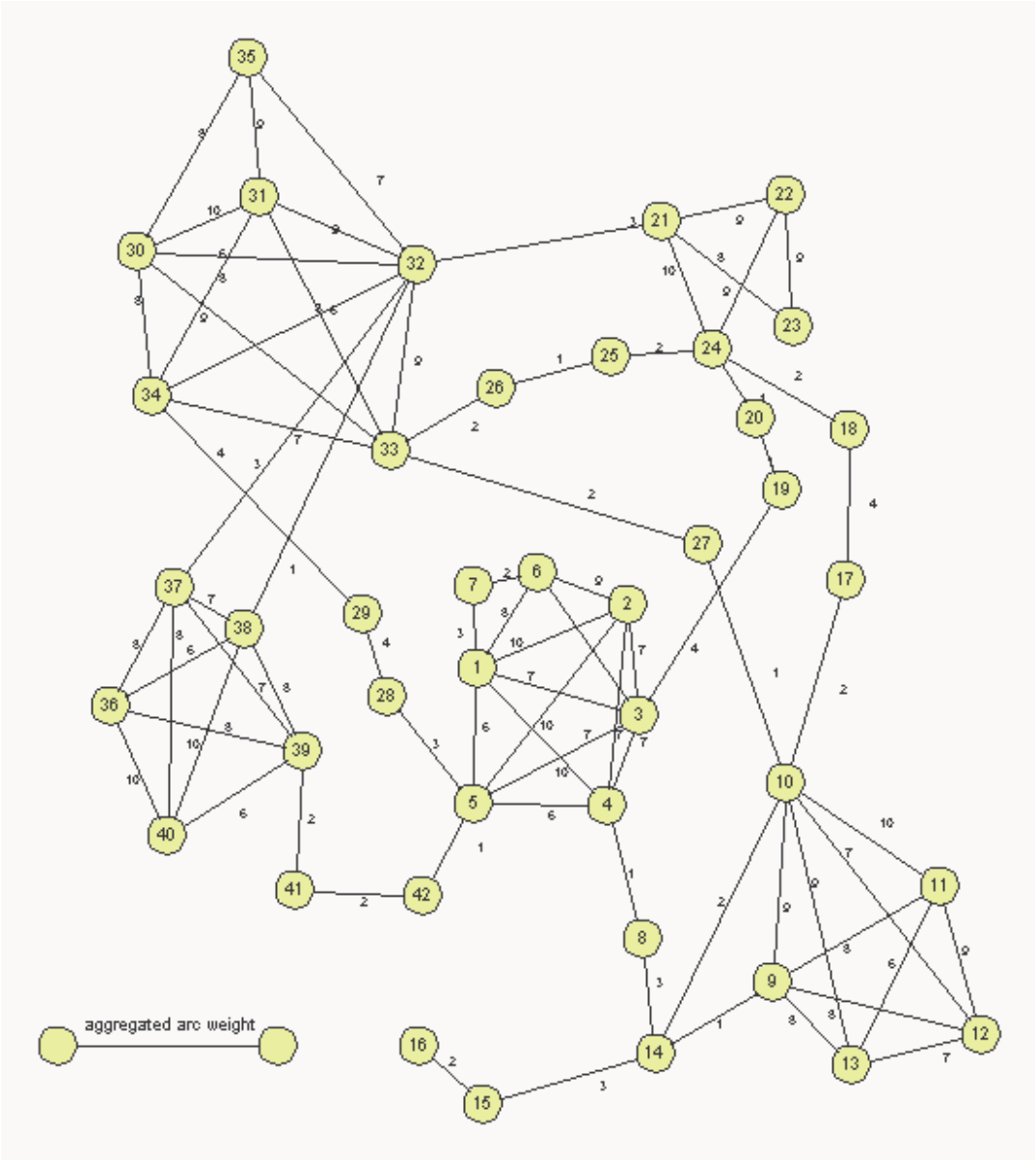

Figure 7. **Disaggregated network with weighted arcs for NSDC**

4.2.2.1 Apply degree centrality node measures before aggregation for NSDC. Recall in this method that a node, i, is assigned a weight, α_i, equal to the weight on arcs out of i. Figure 8 shows the network in which every individual has been assigned the appropriate degree centrality node weight. The weights in the network represent each individual's position with respect to length of communication. Since individuals within groups communicate more with others in their own subgroups, and have longer conversations than the cut-outs do, individuals in subgroups have higher weights. In general, individuals in groups have a higher positional weight with respect to lengths of phone calls in which they are engaged. Individuals within the same subgroup have approximately the same measure. Similarly, the cut-outs are not distinguishable from one another in their positional weights.

The discrepancy between subgroups and cut-outs is even clearer in Figure 9, which shows the weight of each subgroup or individual in the network. Recall the weight of a subgroup, C, is equal to the sum of the weights of the individuals in that subgroup. At this level of the analysis, the measures on the nodes represent the total length of all communications a subgroup has. Any communications between individuals in the same subgroup has been double counted. It was noted in Chapter 3 that any arc weight connecting nodes in a subgroup are accounted for twice in the degree centrality measure on the aggregated subgroups.

4.2.2.2 Aggregate before applying degree centrality node measures for NSDC. Returning to Figure 7 on page 76, the network is first aggregated into appropriate subgroups and then each subgroup or remaining individual is assigned a positional node weight. Recall that when aggregating a network with weighted arcs, the arc connecting two aggregated subgroup has weight equal to the sum of the arcs with exactly one endpoint in each subgroup.

These arc weights for this scenario can be seen in Figure 10. These weights are relatively small, since they do not measure communications between members of the same subgroup. Thus if only concerned with intergroup relationships, these weights provide an indication of the level of those communications.

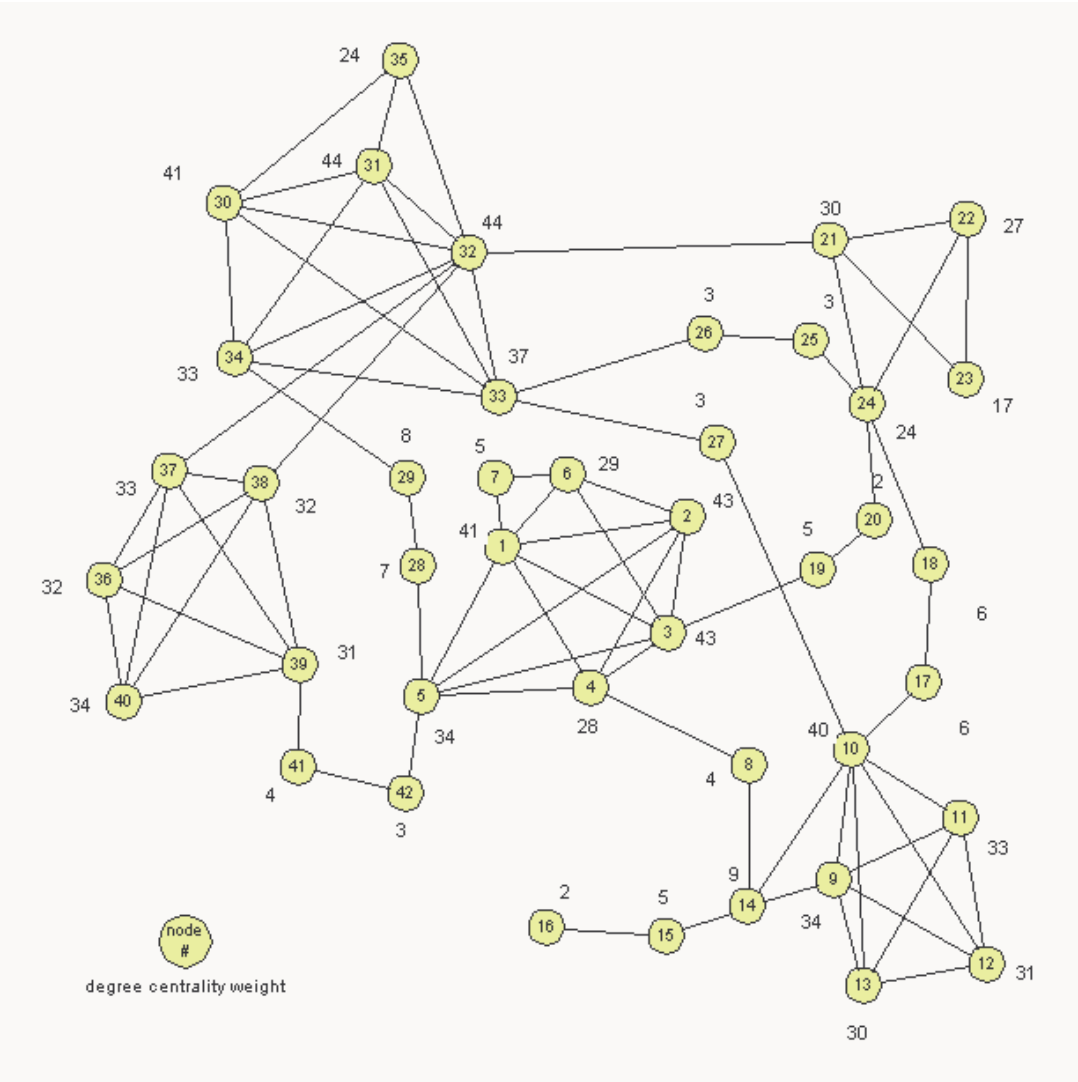

Figure 8. **Individual positional degree centrality measures for NSDC**

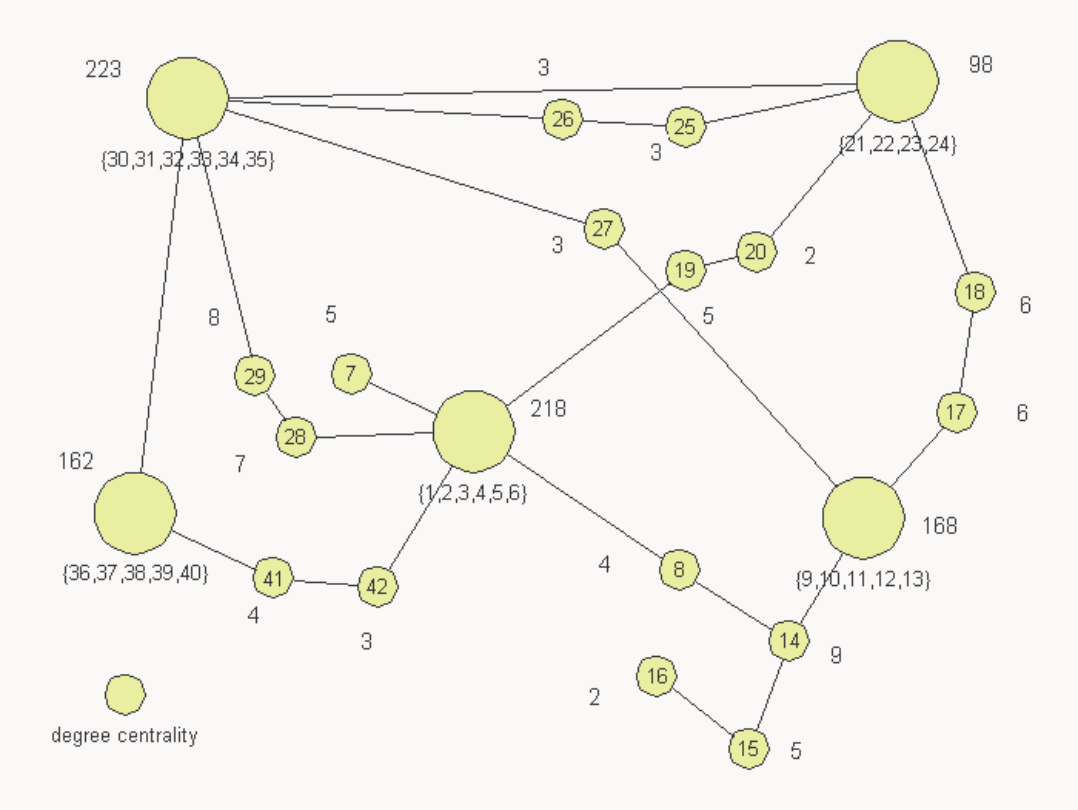

Figure 9. **Node weighted 3-plex aggregated network for NSDC**

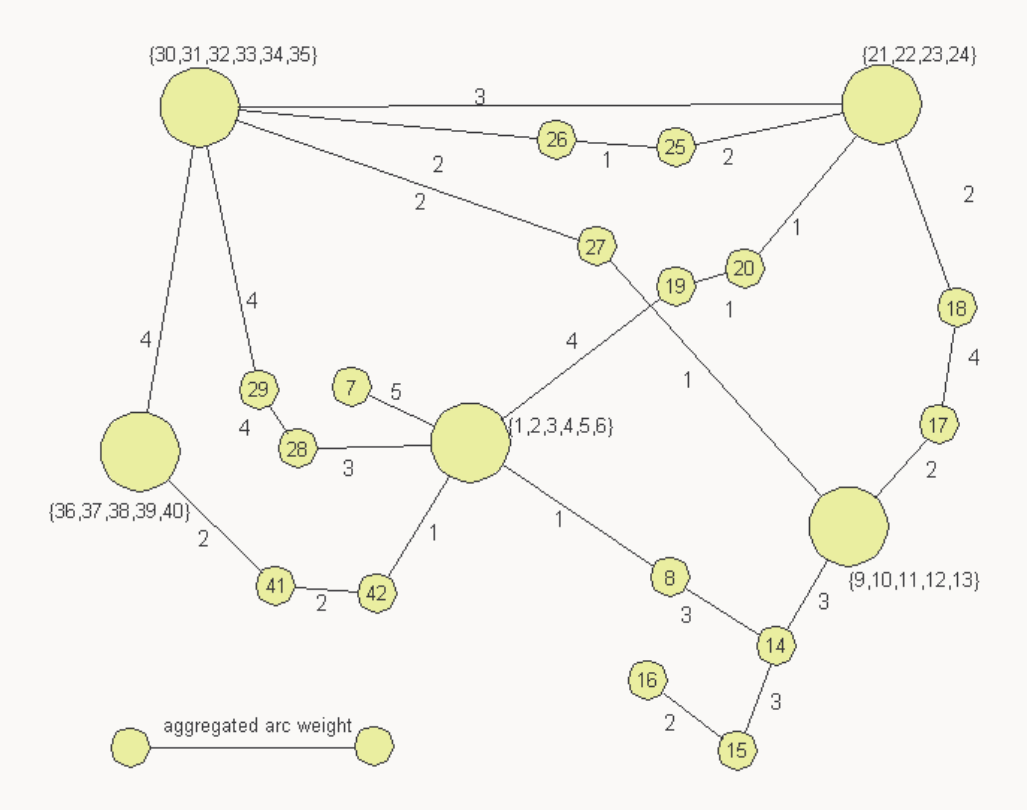

Figure 10. **3-plex aggregated network with weighted arcs for NSDC**

The next step of the analysis is to assign each subgroup a degree centrality node measure to the subgroups. Figure 11 shows these measures for each of the aggregated subgroups. These weights represent the positional weight a subgroup has among the subgroups with respect to length of communication. Since these weights do not include communications within the subgroup (as those in Figure 9 on page 79), they do not overwhelm the weights on the individuals.

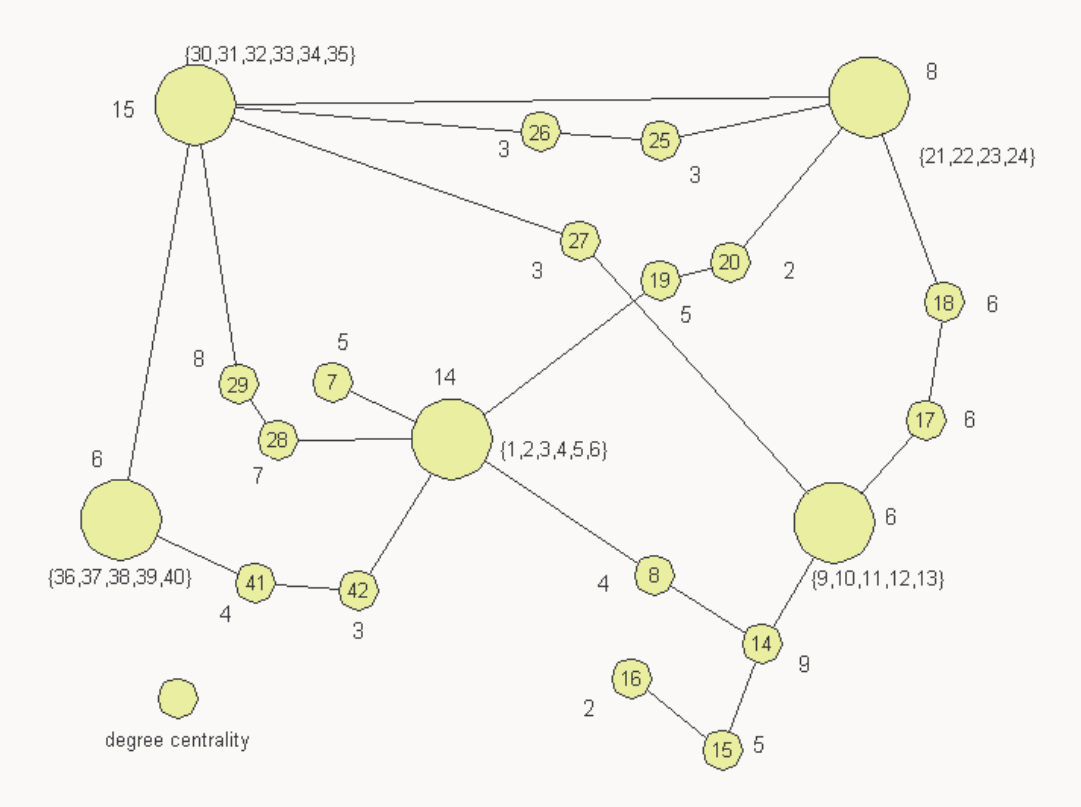

Figure 11. **3-plex aggregated network of intergroup communications for NSDC**

When considering only intergroup interactions, the subgroup {30,31,32,33,34,35} has the greatest positional weight; they communicate more with other subgroups or individuals than any other subgroup. However, recall that node 32 in this subgroup has some direct communications with other groups, a property not exhibited by any other individual in the group. Node 32's neglect in following good OPSEC makes that individual a potential target for exploitation. Though the direct phone calls to individuals in other groups may not necessarily contain important content, they give

an indication of where the groups are located or how accessible they are. Furthermore, any drastic change in 32's unofficial communications, either increased or decreased, may be of interest as it may portend a changing level activity for subgroup {30,31,32,33,34,35}.

The cut-outs hold a great deal of power in this network, as they control the flow of information and materiel through the network. Their measured communications are all fairly small, indicating that they are used only when necessary. Though it is tempting to think of these cut-outs as promising targets for severing the network, it is also worth considering them as targets for exploitation. Knowledge of the content of their messages may give indication of what multi-subgroup activities are planned. Any information that is essential for the network must pass through these cut-outs.

Overall, the analysis of this network in this scenario demonstrates the discrepancy between cut-outs and group members. Since degree centrality is a local property, a node's location is irrelevant. The cut-outs, though necessary to pass messages along, are not used more than necessary and therefore have much lower measures in this scenario. However, this implies the contents of communications involving the cut-outs may be more worthwhile for intelligence resources.

Subgroup {30,31,32,33,34,35,36} has the highest weight, followed closely by {1,2,3,4,5,6}. This is because the measure considers only direct communications, and each is a large subgroup, so individuals communicate directly with many others. In addition, node 32's direct communication to individuals in other subgroups increases the total length of conversation of the subgroup.

This scenario has considered this social network with an additive weight to learn more about the local properties of individuals and subgroups. The next scenario continues analysis on the same network structure, but transforms the weights for application of the closeness centrality node measure.

4.2.3 Scenario 2: Non-overlapping subgroups, Closeness centrality (NSCC)

It is suspected that some separated groups are in the planning stages of a simultaneous attack on multiple targets. Individuals who are organizing attacks or other events have positioned themselves and the people who are loyal to them such that the organizers can disseminate orders and information

quickly. However, the groups are distant enough that no single person acts as a liaison between groups. Examining communication with everyone else in the network requires using the global property of closeness centrality, defined as the sum of shortest paths. Since the measure is based on shortest path for this scenario, it is important that a smaller weight on the arc represent a higher level of communication. Two techniques for transforming the arc weights are offered in Section 3.4.4. Since each arc has weight ≥ 1 (see Figure 7 on page 76), the transformation is as follows: if the arc has weight $a_{i,j}$ then the new weight for this scenario is $1/a_{ij}$. These weights can be seen on the network in Figure 12. Table 8 shows a summary of the calculations and interpretations of the aggregation for NSCC. All the aggregation and node measure calculations for the analysis in this section were performed in 35.75 seconds of processing time. All shortest path calculations were performed using Dijkstra's shortest path algorithm.

Table 8. **Summary of Calculations for NSCC**

Aggregation Order	Calculations	Interpretation
Assign node measures then aggregate	$\alpha_i = \sum_j d_{i,j}$	Individual's total time of communication
	$\alpha_{C_i} = \min_{i \in C_i} \alpha_i$	Minimum subgroup communication within and between subgroups
Aggregate then assign node measures	$a_{C_i,C_j} = \min_{\substack{i \in C_i \\ j \in C_j}} a_{i,j}$	Minimum time of communication between subgroups
	$\alpha_{C_i} = \sum_{C_j} d_{C_i,C_j}$	Total subgroup communication time between subgroups only

4.2.3.1 Apply closeness centrality node measures before aggregation. The first aggregation method assigns to each individual a closeness centrality node measure. Then in aggregation, each subgroup is assigned the weight of the minimum weighted node in that subgroup. Figure 13 shows the positional weights of each individual with respect to the relative speed of communicate with every other individual in the network. Recall from Figure 12 that the weights on arcs within a subgroup are smaller (uniformly distributed on $\left[\frac{1}{10}, \frac{1}{6}\right]$) than the arcs connecting subgroups or cut-outs (which are uniformly distributed on $\left[\frac{1}{4}, 1\right]$). This weighting supports the assumption that

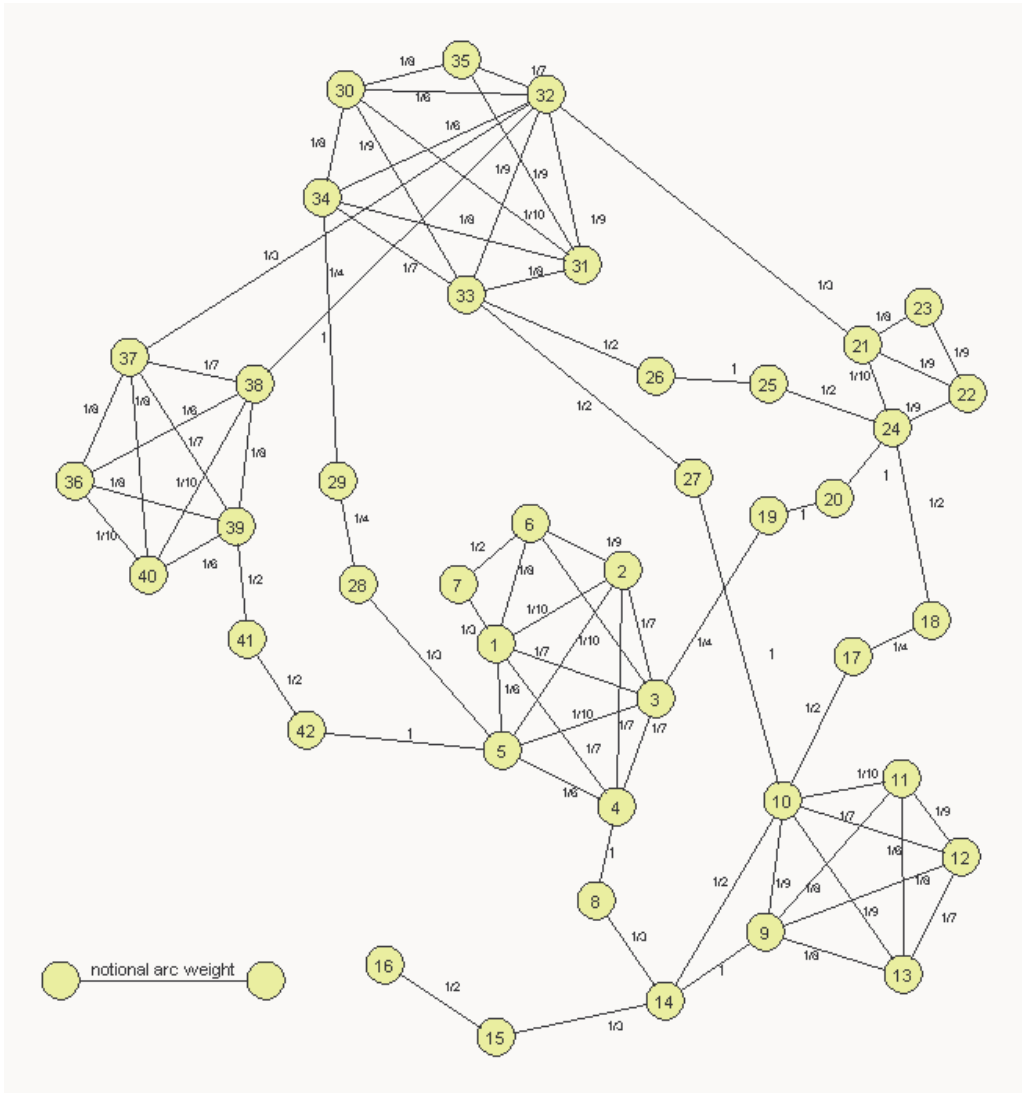

Figure 12. **Disaggregated network with tranformed arc weights for NSCC**

members are closer and therefore communicate more quickly and efficiently with other individuals within their own subgroup than they do with members outside of their subgroup.

Since closeness centrality is a global property, the entire network structure affects which individuals have high or low node measures. The closeness centrality measures assigned to individuals can be seen in Figure 13, where the closeness centrality weights have been rounded to the third decimal. While most of the cut-outs actually have a fairly *high* measure, node 27, with a weight of 52.861, has a substantially lower weight than any of the other cut-outs.

It was expected that nodes 1 through 6 would have the lowest node measures since their subgroup appears to hold the most central position. The subgroup {1,2,3,4,5,6} is the only one that can communicate with every other subgroup, without relying on a third subgroup as an intermediary. However, even being in that position in the network, the individuals in subgroup {1,2,3,4,5,6} have weights greater than those for individuals in subgroup {30,31,32,33,34,35}. This is due to node 32's ability to interact directly with individuals in two of the other four subgroups without having to incur the extra time spent passing information or materiel through cut-outs.

In the aggregation step individuals are aggregated into their appropriate subgroups and the weight on a subgroup node is equal to the minimum of the member's weights. For example,

$$\begin{aligned} \alpha_{\{1,2,3,4,5,6\}} &= \min\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6\} \\ \alpha_{\{1,2,3,4,5,6\}} &= \min\{54.967, 53.387, 53.800, 53.968, 51.404, 57.086\} \\ \alpha_{\{1,2,3,4,5,6\}} &= 51.404 \end{aligned}$$

This definition of a subgroup's measure of communication speed means that its ability to disseminate information or materiel to the rest of the network rests upon the individual in the network who can do so the quickest. Of course the method would be as effective with any other additive measure in the same circumstances. Figure 14 shows the weighted aggregated network.

The individual with the smallest node measure, node 32, makes subgroup {30,31,32,33,34,35} the subgroup of smallest measure. The depiction of the aggregated network in Figure 14 shows

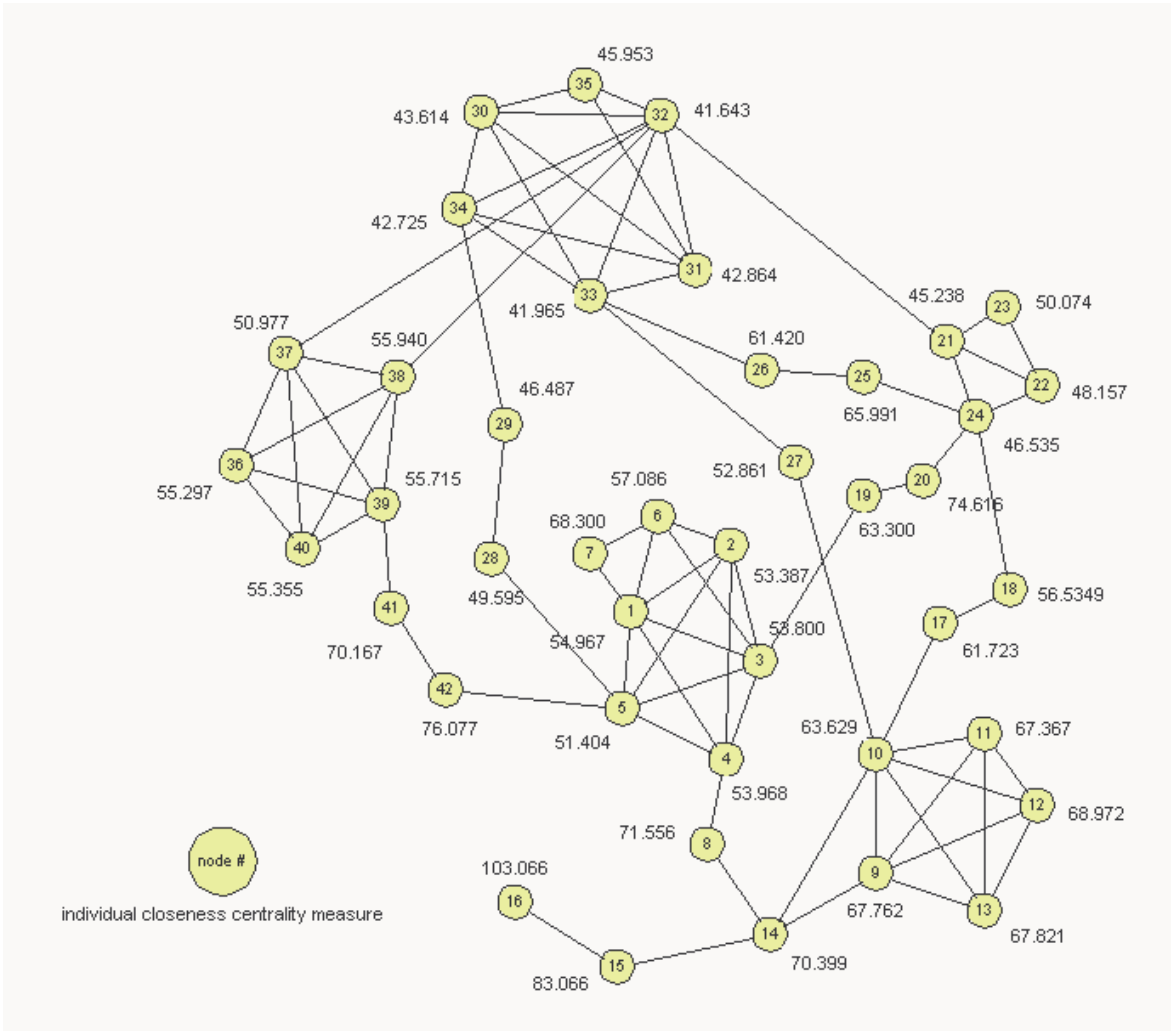

Figure 13. **Individual positional closeness centrality weights for NSCC**

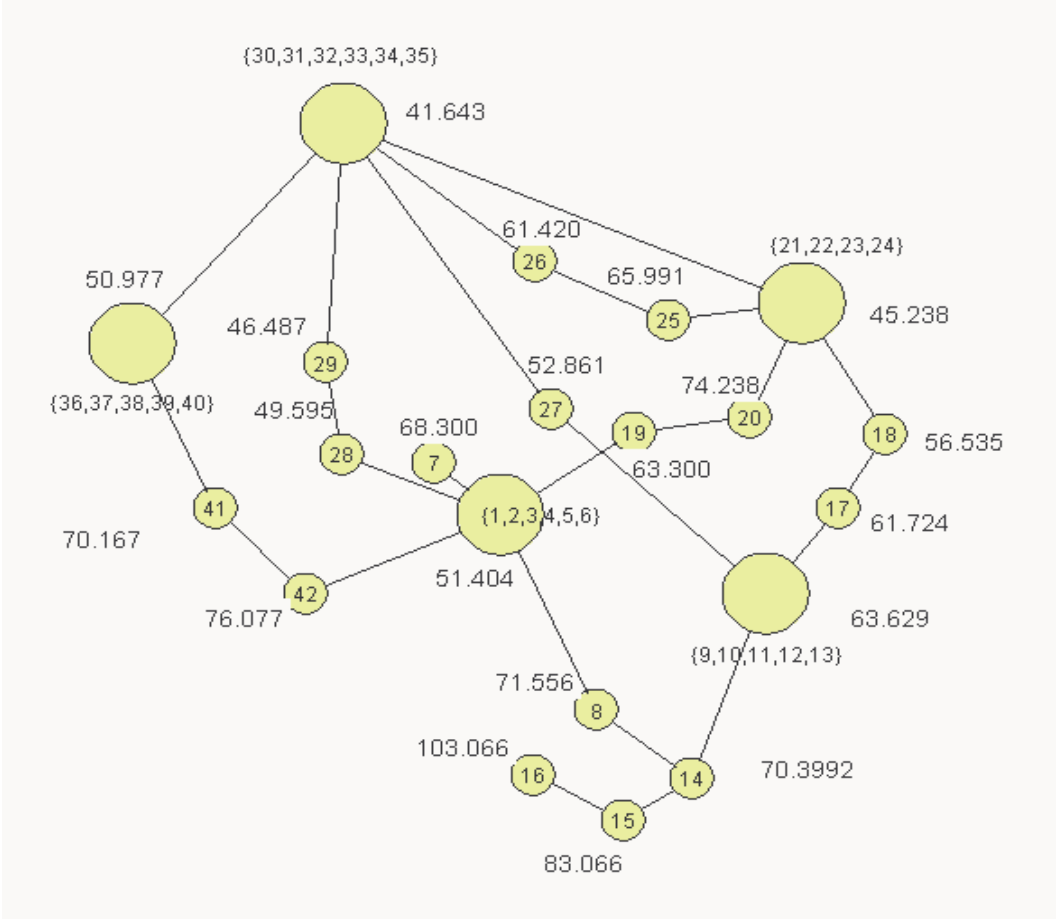

Figure 14. **Closeness centrality measures for 3-plex aggregated weighted nodes for NSCC**

that while subgroup {1,2,3,4,5,6} lies in a central position in a network, it also becomes clear why subgroup {30,31,32,33,34,35} has such a low weight. The latter's use of few cut-outs makes for stronger closeness centrality measure. Subgroup {30,31,32,33,34,35}uses no cut-outs to two of the other subgroups, one cut-out to subgroup {9,10,11,12,13}, and two cut-outs to {1,2,3,4,5,6}. Subgroup {1,2,3,4,5,6}, however, must pass all information or goods to every other subgroup through two cut-outs. Since the weights on paths utilizing cut-outs are greater than weighted arcs within subgroups, using cut-outs to disseminate information to the other subgroups increases the weight for subgroup {1,2,3,4,5,6}.

4.2.3.2 Aggregate before assigning closeness centrality node measures. The second aggregation technique aggregates individuals into subgroups with weighted arcs, and then assigns to each subgroup a closeness centrality node measure. In the aggregation step, the arc between two aggregated subgroups is defined to be the minimum weighted arc among the arcs with exactly one endpoint in each of the two subgroups. The aggregated network with weighted arcs is shown in Figure 15. Following the same notional scenario, these weights are defined to show the minimum speed with which subgroups can exchange information or materiel with each other. Therefore, a smaller weight implies quicker or more efficient interactions.

Aggregation of the network in this manner does not take into account the speed of communications within a subgroup. In many practical situations, the within-subgroup communications may not be important, since each of the subgroups has only one or a small number of individuals who interact with the cut-outs and act as representatives for their subgroup. The aggregation step then assigns to each subgroup a closeness centrality node measure. Figure 16 shows the subgroups' weights for their positional measures in the network. Again, subgroup {30,31,32,33,34,35} has the lowest weight, since it can communicate directly with two other subgroups. Its shortest paths to every other subgroup requires the use of very few cut-outs.

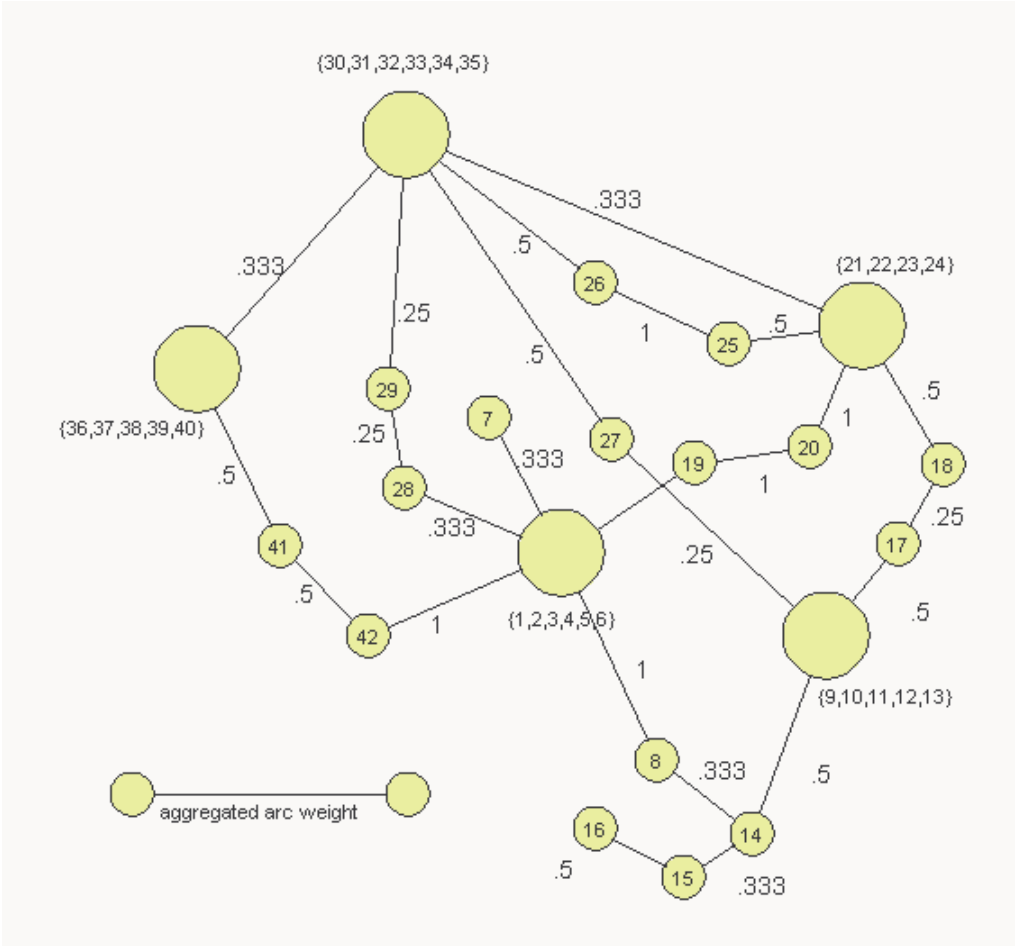

Figure 15. **Weighted arcs in 3-plex aggregated network for NSCC**

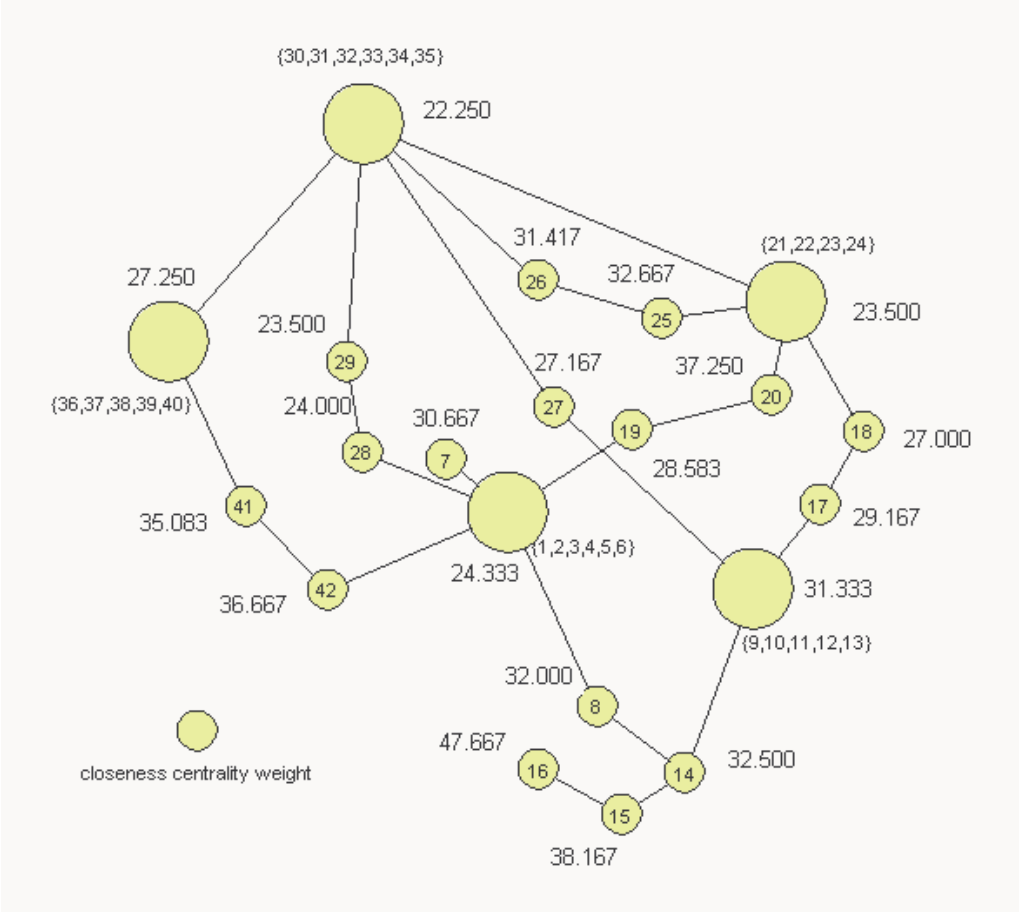

Figure 16. **3-plex aggregated closeness centrality weights for NSCC**

The first order of analysis for aggregation, in which each individual is assigned a closeness centrality measure before aggregation results in the weights shown in Figure 14, gives the following list of subgroups, in increasing order of closeness centrality measure:

1. $\{30, 31, 32, 33, 34, 35\}$
2. $\{21, 22, 23, 24\}$
3. $\{36, 37, 38, 39, 40\}$
4. $\{1, 2, 3, 4, 5, 6\}$
5. $\{9, 10, 11, 12, 13\}$

In the second order of analysis for aggregation, in which individuals are aggregated and subgroups are assigned a closeness centrality measure, shown in Figure 16, the increasing order of subgroup closeness centrality measures through NSCC is slightly different, as {1,2,3,4,5} shows itself to have a smaller weight than {36,37,38,39,40}:

1. $\{30, 31, 32, 33, 34, 35\}$
2. $\{21, 22, 23, 24\}$
3. $\{1, 2, 3, 4, 5, 6\}$
4. $\{36, 37, 38, 39, 40\}$
5. $\{9, 10, 11, 12, 13\}$

The reason the order changes stems from whether the distances within subgroups are included in the subgroup node measures. In the former, within-subgroup weights are included and an individual (the one with minimum weight) represents the subgroup. In the latter, however, subgroup weight is determined solely by length of paths between subgroups. It has been shown that the analysis differs whether the aggregation is performed with weighted arcs or weighted nodes. Rather than choose which method to use, it is suggested that both be performed to understand the most about the network at hand.

Analyzing this illustrative network structure with a global property closeness centrality shows the extent to which node 32's direct relationships with members of other subgroups speeds up network communications. Though {1,2,3,4,5,6} appears to hold a central position, disseminating information or materiel to the rest of the network requires the use of many cut-outs, driving up

the total speed. This implies that proper use of the cut-outs may slow down communications. Consider a scenario in which a governing council of an organization is careful to use cut-outs then the leadership forms the slowest link in the chain, if the lower level members choose to use direct connections instead of the cut-outs.

The closeness centrality measure, unlike the degree centrality measure, considers a node's position in the network. Therefore, in this second scenario, the cut-outs, as gate keepers, become more important, as they control the flow of communications by being on many of the shortest paths.

These first two scenarios have examined a notional network with distinct subgroups through a local and global property, degree and closeness centrality, respectively. In both, node 32 has distinguished himself as being more communicative than anyone else. This has raised the degree centrality measure both for him and his subgroup, simply because he communicates directly with many people. It has also improved the closeness centrality measure for him and his subgroup, as he can avoid the extra time incurred by using the cut-outs. There is no official line of communication from subgroup {30,31,32,33,34,35} to {36,37,38,39,40} (as evidenced by the lack of cut-outs), but node 32 circumvents that by communicating directly with 37 and 38. The rest of the network structure shows that node 5 is supposed to control communications to the latter subgroup.

The next two scenarios use the same node measures to analyze a network in which some individuals serve as liaison members between subgroups. Without cut-outs to control the flow of information, the liaisons, as group members, hold what seems to be a position of great power. The analysis in the two scenarios shows that to be true.

4.3 Notional Network of Overlapping Subgroups

Figure 17 shows the structure for a second notional network with overlapping subgroups. Nodes {4,5,11,15} are the liaison individuals that connect subgroups, though not all subgroups have such intimate connections. Note that arc (7,8) holds the entire network together; communications throughout the network can be severed quite easily by removing that single arc. However, arc (7,8) is the

only weak place in terms of connectivity. The rest of the network is held together quite strongly by either multiple liaison nodes such as 4 and 5, or by redundant arcs such as (10,20) and (13,17).

4.3.1 Subgroup detection for network of overlapping subgroups

It is clear that the subgroups in this scenario share members. Therefore, the technique introduced in Section 3.3.2 for detecting overlapping subgroups is appropriate. This method recommends first finding all overlapping cliques and then, if necessary, extending those cliques to k-plexes. Though the heuristic for finding overlapping cliques is only guaranteed to find 75%, this network does not have the structure for which the method fails. Therefore, it is expected that all cliques will be detected. The overlapping cliques found in 0.18 seconds using a MATLAB script on the 2 ghz Dell Dimension XPST800 Pentium 4M:

$\{1, 2, 3, 4, 5\}$
$\{4, 5, 6, 7\}$
$\{8, 9, 10\}$
$\{8, 10, 11\}$
$\{11, 12, 14, 15\}$
$\{12, 13, 14, 15\}$
$\{15, 16, 18, 19, 20\}$
$\{16, 17, 18, 19, 20\}$

Observe cliques {11,12,14,15} and {12,13,14,15}, which overlap in three nodes and have only one node different. The structure of these two cliques shows that one missing arc, $(11, 13)$, keeps this set of five nodes from being a K_5. This is a situation in which it is appropriate to consider extending cliques to k-plexes, and examining them as candidates for subgroups. $k = 1$ finds only the pure cliques which have already been determined, so the k-plex analysis starts with $k = 2$. The k-plexes for $k = 2$ to $k = 4$ are shown in Table 9.

The list of 3-plexes shows {1,2,4,5,6,7}, which is inappropriate in missing node 3. The list of 4-plexes in the table shows {7,8,9,10,11}; however, a review of Figure 19 shows this to be an inappropriate subgroup. For this notional example, further analysis of this network structure uses the 2-plexes to aggregate into subgroups. Figure 18 shows the structure of the aggregated network with five aggregated subgroups and one remaining disaggregated individual, node 21.

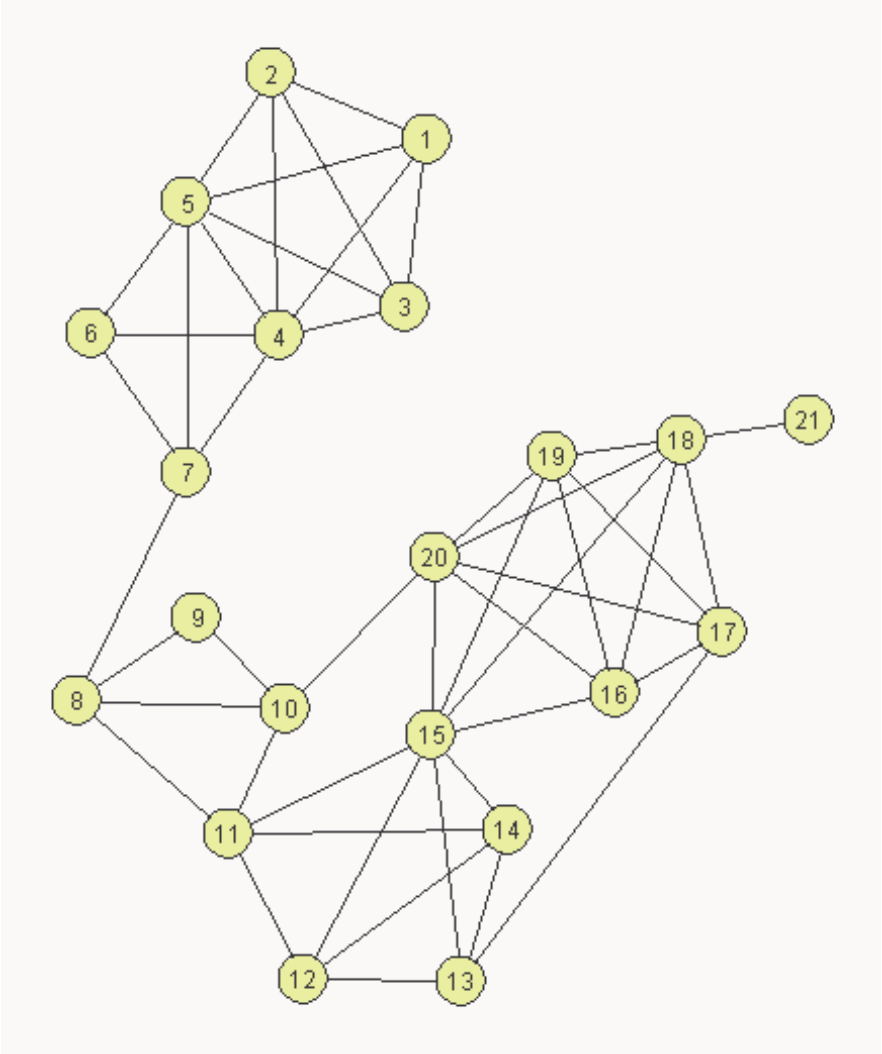

Figure 17. **Disaggregated Network Structure for Overlapping Subgroups**

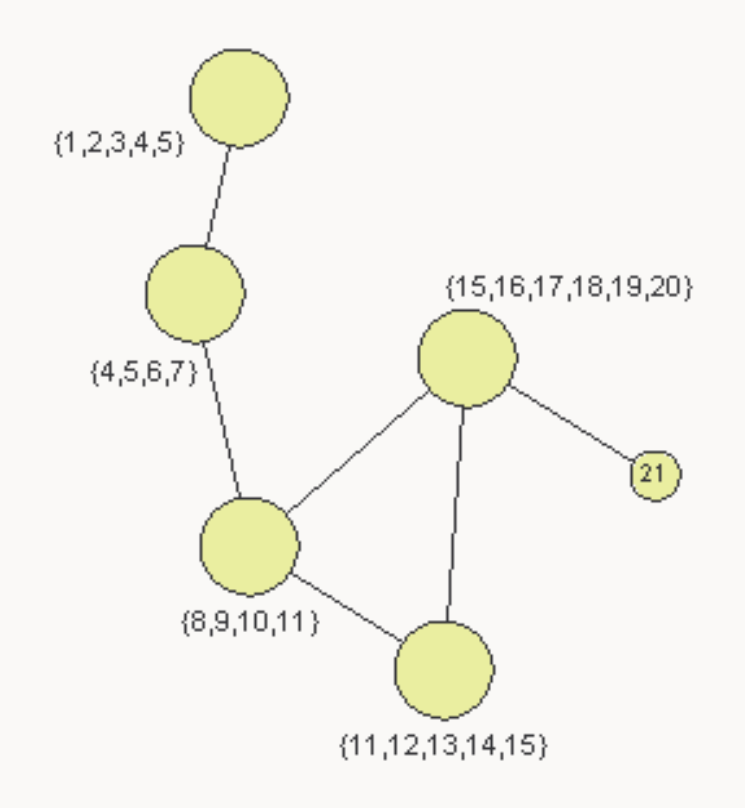

Figure 18. **Structure of 2-plex aggregated network of overlapping subgroups**

Table 9. **k-plexes for network of overlapping subgroups**

k	k-plexes	computational time
2	$\{1,2,3,4,5\}$ $\{4,5,6,7\}$ $\{8,9,10,11\}$ $\{11,12,13,14,15\}$ $\{15,16,17,18,19,20\}$	0.30 sec
3	$\{1,2,3,4,5\}$ $\{1,2,4,5,6,7\}$ $\{8,9,10,11\}$ $\{11,12,13,14,15\}$ $\{15,16,17,18,19,20\}$	0.36 sec
4	$\{1,2,3,4,5,6,7\}$ $\{7,8,9,10,11\}$ $\{11,12,13,14,15\}$ $\{15,16,17,18,19,20\}$	0.42 sec

4.3.2 Scenario 3: Overlapping subgroups, Degree centrality (OSDC)

Terrorist cells have now become connected to the rest of the network. Some individuals within subgroups act as liaisons between groups, implying these individuals may need to be modeled as members of multiple aggregated subgroups. The intelligence data gathered is assumed to be numbers of conversations, in which a greater arc weight is assumed to represent a stronger relationship. The structure of the network and the arc data can be seen in Figure 19. All arcs in this illustration are assigned a weight randomly from a uniform distribution on $[1,5]$. Any arc incident to an apparent liaison individual ($4, 5, 11$, or 15), is randomly assigned a weight from a uniform distribution on $[1,6]$. This allows the liaison individuals to communicate more than individuals in only one subgroup. The additive nature of the data represented by the arc weights makes using the degree centrality node measure appropriate. Table 10 shows a summary of the calculations and interpretations of the aggregation for OSDC.

4.3.2.1 Assign degree centrality node measures and then aggregate into subgroups.

The analysis in this section first assigns to each individual in the network a degree centrality node measure. The individuals are then aggregated into the 2-plex subgroups. The measure on the aggregated subgroup node is the sum of the weights of the individual members of that subgroup.

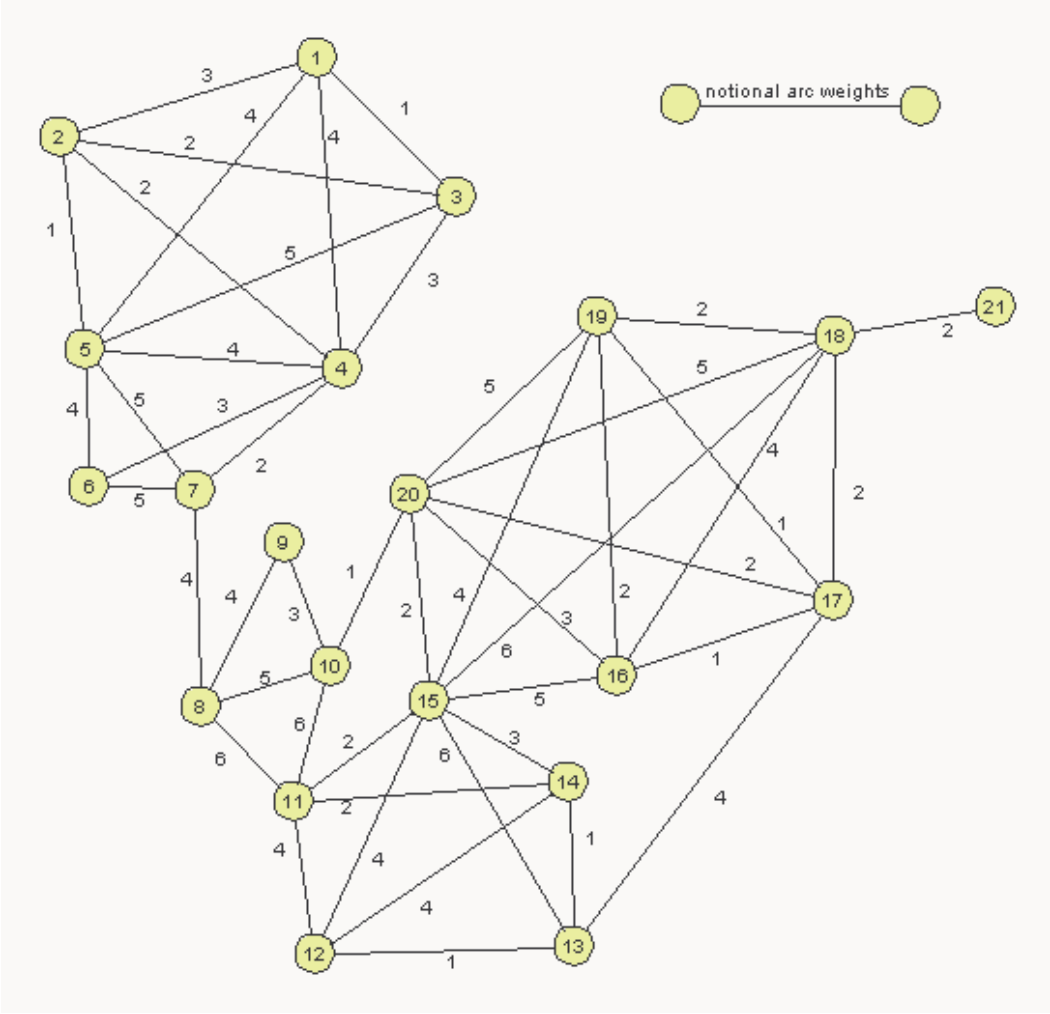

Figure 19. **Arc weighted disaggregated network for OSDC**

Table 10. **Summary of Calculations for OSDC**

Aggregation Order	Calculations	Interpretation
Assign node measures then aggregate	$\alpha_i = \sum_j a_{i,j}$	Individual's total communication
	$\alpha_{C_i} = \sum_{i \in C_i} \alpha_i$	Total subgroup communication within and between subgroups
Aggregate then assign node measures	$a_{C_i,C_j} = \sum_{\substack{j \in C_j \setminus C_i \\ i \in C_i \setminus C_j}} a_{i,j} + \sum_{\substack{k \in C_i \cap C_k \\ i \in C_i \setminus C_j}} a_{i,k} + \sum_{\substack{k \in C_i \cap C_k \\ j \in C_j \setminus C_i}} a_{j,k}$	Total communication between subgroups
	$\alpha_{C_i} = \sum_{C_j} a_{C_i,C_j}$	Total subgroup communication between subgroups only

Determining all weights for nodes and arcs in the remainder of the analysis in this section took 0.761 seconds of computer processing time on the 2ghz Dell Dimension XPST800 Pentium 4M.

Figure 20 shows the first step of the aggregation, in which each individual has been assigned a degree centrality measure. Each individual node is given a weight equal to the sum of weight on arcs incident to the node. This is a local property, representing the number of conversations in which the individual has been a participant. Not surprisingly, the highest weight is held by node 15, who is able to communicate directly with so many other individuals. Similarly, node 21, who has direct communication with only one individual has a low weight. Perhaps 21 is new to the network and has not yet developed many direct relationships with any other network members, or perhaps 21 is not well trusted by the rest of the network. Conversely, node 21 may be the actual organizational head and only rarely communicates through node 18. Without knowledge of the content of their conversations, it is difficult to know the relationship shared by nodes 18 and 21.

The second step of the aggregation technique aggregates individuals into appropriate subgroups. Each subgroup's weight is equal to the sum of the individual weights in the subgroup. Those individuals who are members of more than one subgroup (11, for example) contribute their individual weight to each subgroup to which they belong.

Figure 21 shows the weighted network after the aggregation step. The weight on a node represents the total number of conversations or meetings in which any individual in the subgroup

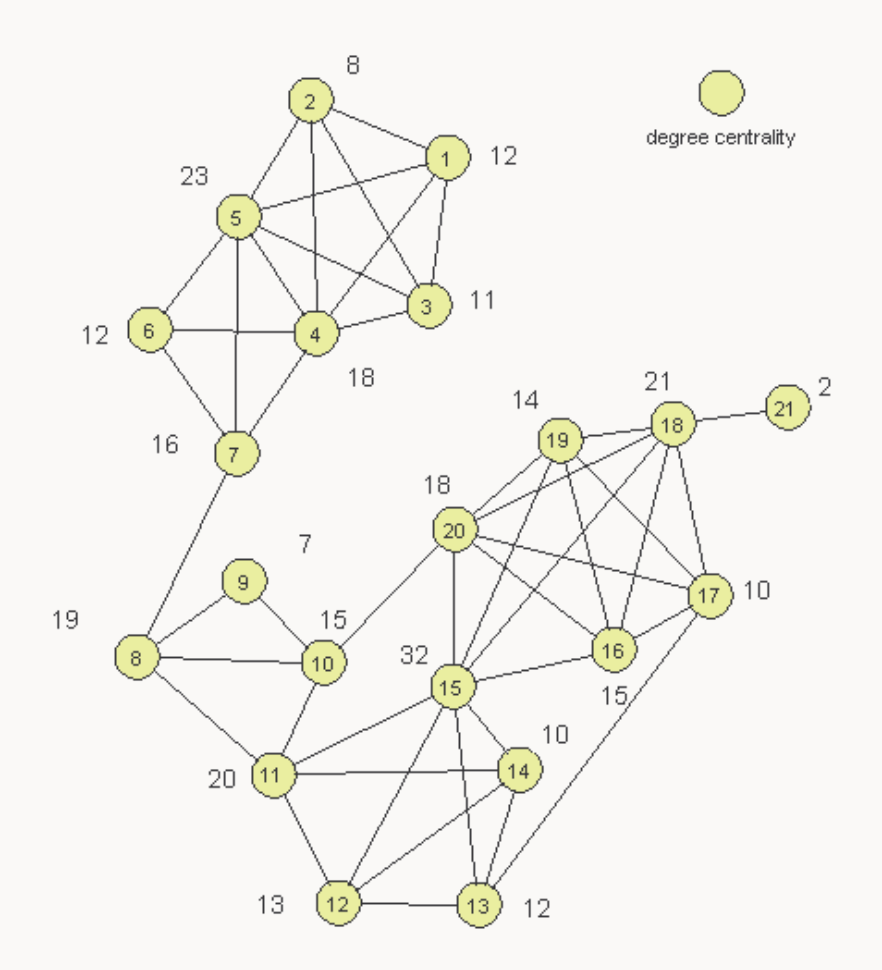

Figure 20. **Individual positional weights for OSDC**

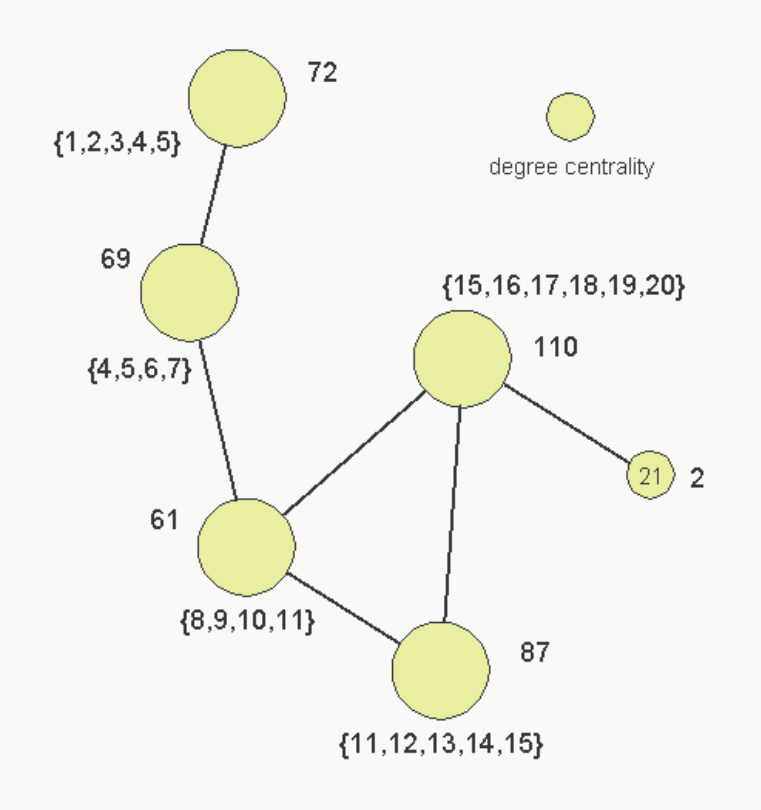

Figure 21. **Subgroup positional weights for OSDC**

has participated. This does include some double counting, but still offers a relative measure among the subgroups of the amount of communications subgroup members are having. The subgroup {15,16,17,18,19,20} has the highest measure mainly because the subgroup is large and is fairly dense. Position of the subgroup in the network is not a consideration in this analysis, since degree centrality is a purely local property.

4.3.2.2 Aggregate individuals into subgroups and assign degree centrality node measures. The second aggregation technique demonstrated for OSDC performs the aggregation steps in the opposite order: first individuals are aggregated into predetermined subgroups, and then each subgroup is assigned a degree centrality node measure. The aggregation occurs with weighted arcs: the arc connecting two aggregated subgroup nodes is given a weight equal to the sum of all arcs with exactly one endpoint in each of the two subgroups. The arc weights between aggregated subgroup nodes count the number of phone calls or meetings the subgroups share.

Since members can be in multiple subgroups, this arc weight is actually the sum of the weights on arc in three sets. Let C and C' be two subgroups. Then the three sets are:

1. any arc with one endpoint in $C \backslash C'$ and one in $C' \backslash C$
2. any arc with one endpoint in $C \backslash C'$ and one in $C \cap C'$
3. any arc with one endpoint in $C \cap C'$ and one endpoint in C'

Consider Figure 19 on page 97, and let {11,12,13,14,15} be subgroup C and let {15, 16, 17, 18, 19, 20} be subgroup C' . Arc $(13, 17)$ has one endpoint in $C \backslash C'$ and one in $C' \backslash C$, but neither endpoint in both, so $(13, 17)$ belongs in the first set. Arc $(14, 15)$ has one endpoint (node 14) in $C \backslash C'$ and one endpoint (node 15) in $C \cap C'$, so $(14, 15)$ belongs in the second set. Arc $(15, 20)$ has one endpoint (node 15) in $C \cap C'$ and one endpoint (node 20) in $C' \backslash C$. Each of these three arcs will add their weight to the final weight on arc (C, C'). It is worth noting that the arc weights in the aggregated network do not count conversations or meetings between individuals wholly within one subgroup. This includes arc weights on arcs such as $(19, 20)$ and $(4, 5)$. The arc weights are shown in Figure 22.

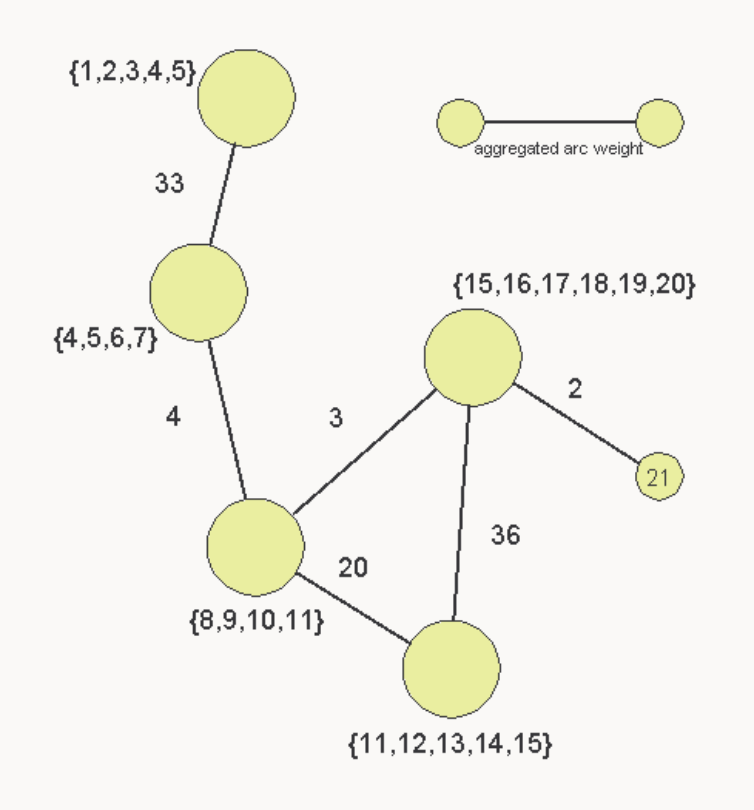

Figure 22. **Arc weights for 2-plex aggregated network in OSDC**

The precarious single arc $(7, 8)$ representing the relationship between nodes 7 and 8 still shows itself to be crucial for network connectivity, but its weight is low. This implies that the two sets of subgroups do not interact often.

The arc connecting subgroups {8,9,10,11} and {15,16,17,18,19,20}, while also of small weight is not as crucial as the $(7, 8)$ arc. Removal of arc $(10, 20)$ would make communications slower between the two subgroups, but does not disconnect them completely. Any information or materiel the two subgroups need to share could still pass through subgroup {11,12,13,14,15}.

The second part of the aggregation places a node measure on each subgroup or remaining individual equal to the sum of the weights on arcs incident to the subgroup. This yields the network in Figure 23. These subgroup node measures count the number of phone conversations and meetings an individual in a subgroup is a participant only with individuals in other subgroups. Any conversations within a subgroup conversations, except those held with the liaison nodes, are not counted in these node weights. Again, in this notional scenario, more communication is assumed to imply a stronger relationship.

Unlike the first aggregation technique, whose final aggregated node measures can be seen in Figure 21 on page 100, here subgroup {11,12,13,14,15} has the greater measure. In Figure 21, {15,16,17,18,19,20} had the higher measure since the number of phone conversations and meetings held between individuals within a subgroup distorted the subgroup's node measure.

This scenario has considered a social network with an additive weight to learn more about the local properties of individuals and subgroups when individuals can have overlapping membership. The next scenario continues analysis on the same network structure, but transforms the arc weight for appropriate use of the closeness centrality node measures. Since the entire network structure affects these measures, the limiting arc (7,8) that controls the flow of information will be seen to greatly affect the network measures.

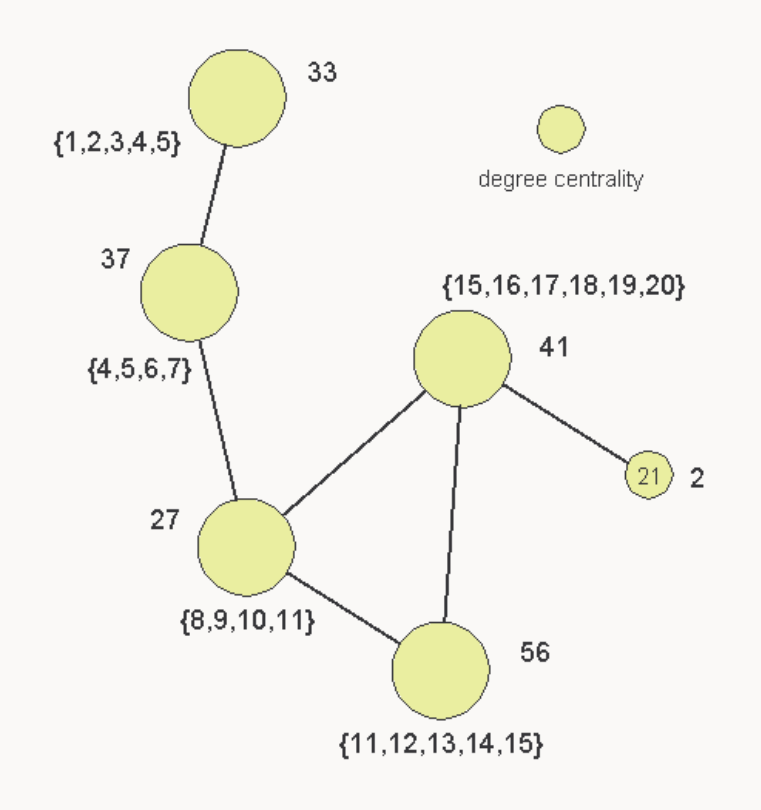

Figure 23. **Positional 2-plex aggregated subgroup weights for OSDC**

4.3.3 Scenario 4: Overlapping subgroups, Closeness centrality (OSCC)

Several groups are in close communication as they plan a coordinated simultaneous attack. The individuals who are organizing the attacks have positioned themselves and the people who are loyal to them such that the organizers can disseminate orders and information quickly. Some individuals are known to be liaisons between subgroups but have membership ties with each of the subgroups, and therefore belong in more than one aggregated subgroup. The data gathered reflects speed with which an individual can pass a message to immediate neighbors. This data is assumed to be the critical measure for this notional analysis. A smaller measure with this data represents a stronger relationship.

Since this is a shortest path analysis, the arc weights must be transformed such that a smaller weight represents faster communication. The network for this scenario has the structure of Figure 17 on page 94, but an arc's weight is the reciprocal of the weight shown in Figure 19 on page 97. Specifically, for a pair of individual nodes i and j with weight $a_{i,j}$ in Figure 19, for this scenario, the same arc has weight $1/a_{ij}$. These weights can be seen in Figure 24. Table 11 shows a summary of the calculations and interpretations of the aggregation for OSCC.

Table 11. **Summary of Calculations for OSCC**

Aggregation Order	Calculations	Interpretation
Assign node measures then aggregate	$\alpha_i = \sum_j d_{i,j}$	Individual's total communication
	$\alpha_{C_i} = \min_{i \in C_i} \alpha_i$	Minimum subgroup communication within andbetween subgroups
Aggregate then assign node measures	$a_{C_i,C_j} = \min\{\min_{\substack{i \in C_i \setminus C_j \\ j \in C_j \setminus C_i}} a_{i,j}, \min_{\substack{i \in C_i \setminus C_j \\ k \in C_i \cap C_j}} a_{i,k}, \min_{\substack{k \in C_i \cap C_j \\ j \in C_j \setminus C_i}} a_{k,j}\}$	Minimum speed of communication between subgroups
	$\alpha_{C_i} = \sum_{C_j} d_{C_i,C_j}$	Total subgroup speed of communication between subgroups only

4.3.3.1 Apply closeness centrality node measures before aggregation. In the first order of analysis for this scenario, each individual is assigned a closeness centrality measure.

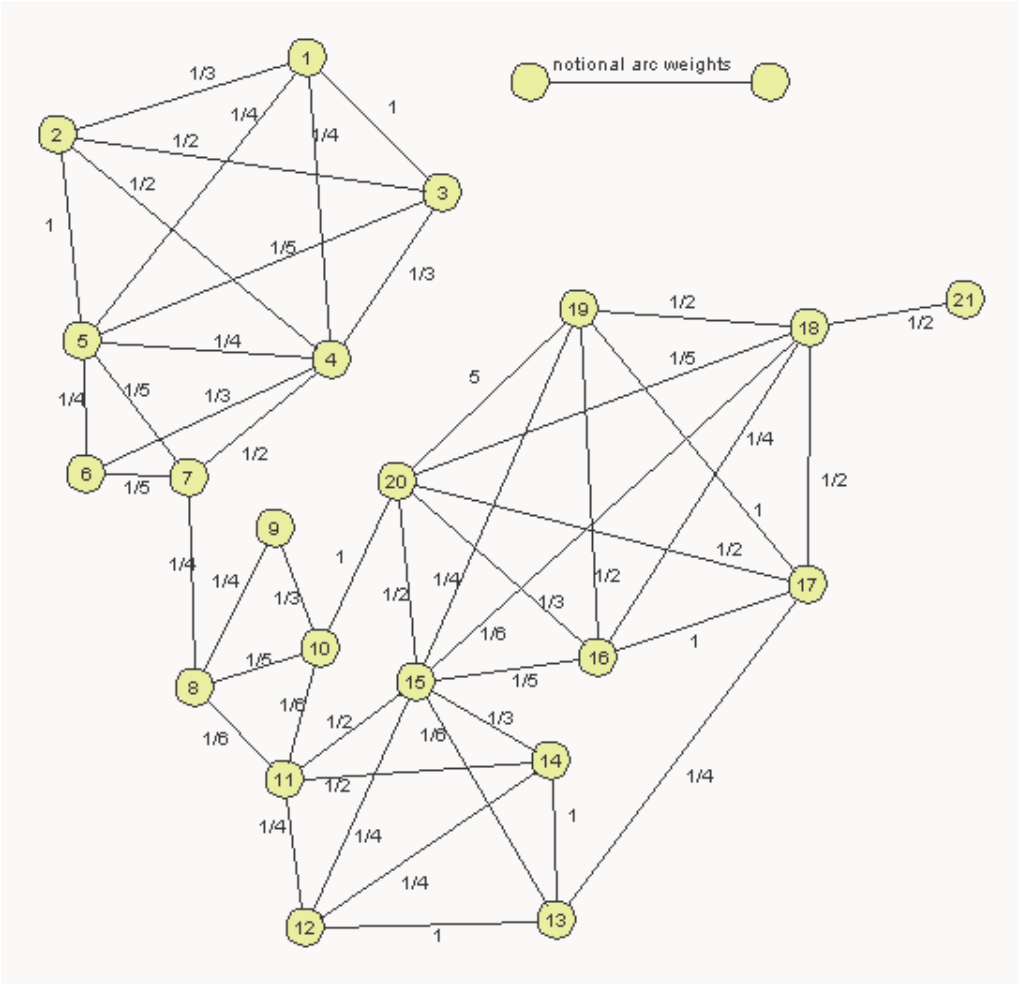

Figure 24. **Notional arc weights for OSCC**

Then the network is aggregated into the subgroups previously determined, and each subgroup or remaining individual is assigned a weight equal to the minimum of the individual node weights of its membership. This allows the analyst to consider one individual, the fastest communicator, as representing the entire network.

Figure 25 first shows each individual node's closeness centrality measure. Not surprisingly, the highest measure in the network is at node 21, due to its position at the fringe of the network. Furthermore, almost every subgroup's smallest weighted node is an individual who is the liaison between groups. The exception is subgroup $\{4, 5, 6, 7\}$, whose smallest weighted member is node 7. Though node 7 is not in multiple groups, it does serve as the liaison from subgroups $\{1, 2, 3, 4, 5\}$ and $\{4, 5, 6, 7\}$ to the rest of the network. For that reason, it is reasonable that node 7 would have the smallest measure of any of the nodes in $\{1, 2, 3, 4, 5\}$ and $\{4, 5, 6, 7\}$.

The aggregation step takes the node-weighted network in Figure 25, aggregates individuals into subgroups, and assigns to each subgroup a weight equal to the minimum of the weights of individuals in the subgroup. The weighted subgroups and remaining individual can be seen in Figure 26. Since the individual of smallest weight, represented by node 11, is in both subgroups {8,9,10,11} and {11,12,13,14,15}, node 11's weight is incorporated in both aggregated subgroup nodes.

4.3.3.2 Aggregate before assigning closeness centrality node measures.

The second order of aggregation analysis for this scenario first aggregates the individuals into their subgroups, and then assigns to each subgroup or remaining individual a closeness centrality measure. This technique ignores arc weights between individuals wholly contained within the same subgroup, and provides information focused on intergroup relationships. This methodology is appropriate when unconcerned with the details within subgroups, but seeks to find information on the speed with subgroups can pass information or goods to the rest of the network. This aggregation technique allows the analyst to detect subtle changes in subgroup interactions, eliminating the noise of within-subgroup interactions.

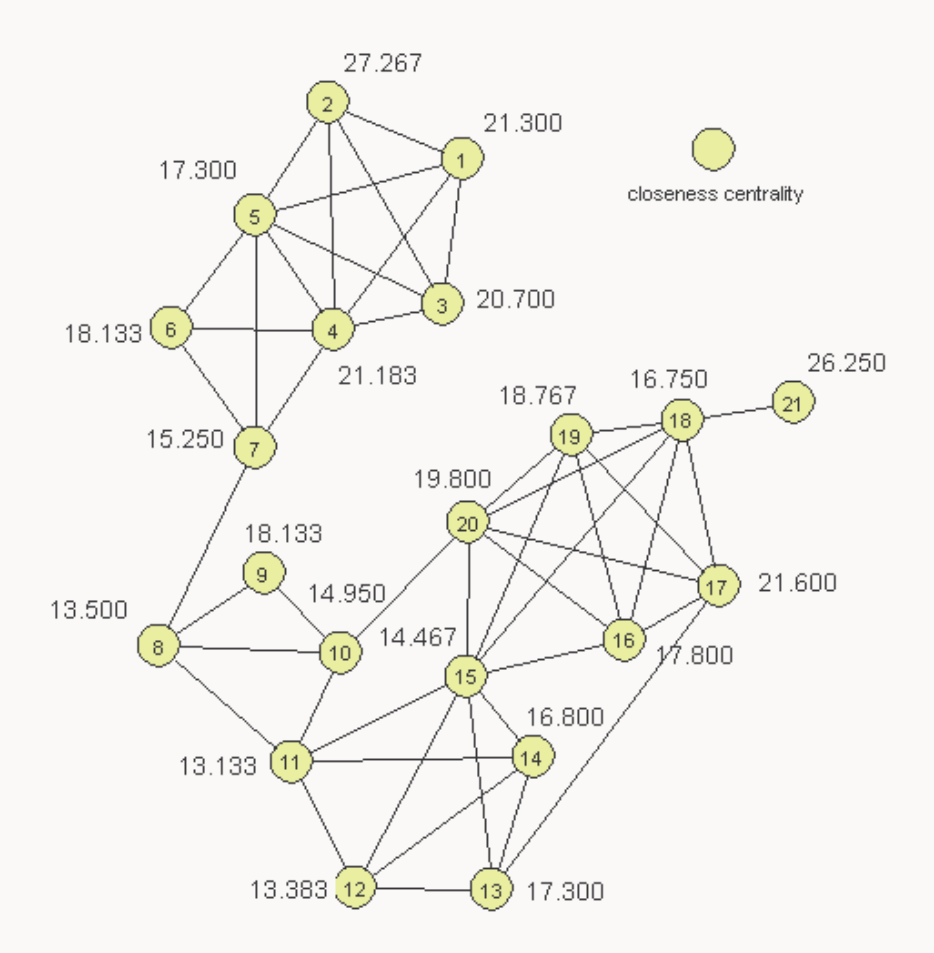

Figure 25. **Individual positional weights for OSCC**

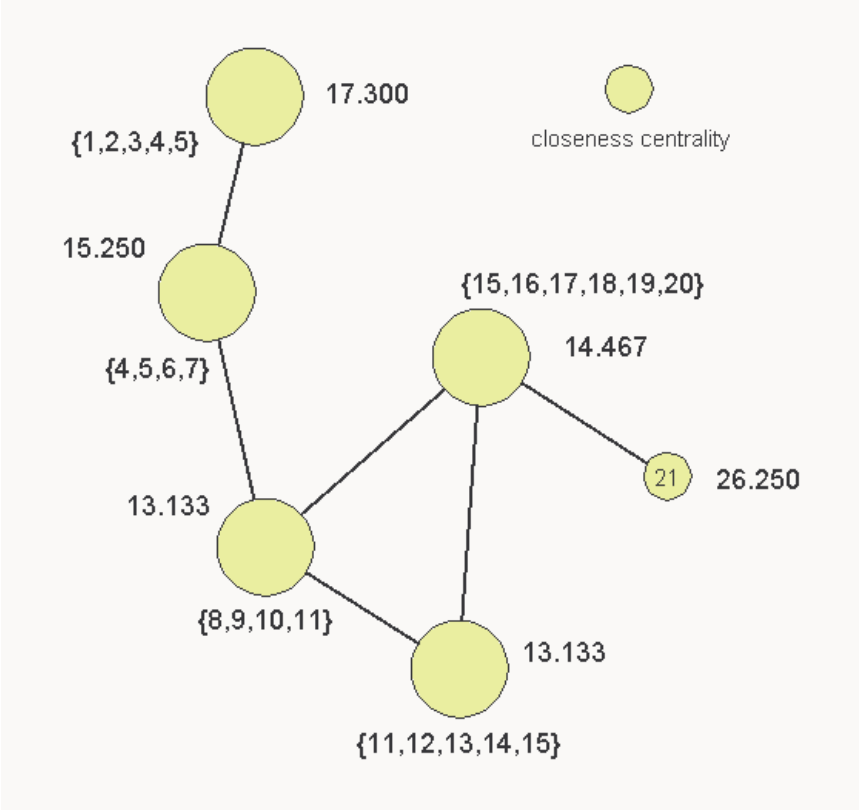

Figure 26. **Aggregated 2-plex subgroup positional weights for for OSCC**

The arc between two aggregated subgroups is defined to be the minimum of any arcs connecting those two subgroups. The arcs connecting two subgroups can be broken down into three sets:

1. any arc with one endpoint in $C \backslash C'$ and one in $C' \backslash C$
2. any arc with one endpoint in $C \backslash C'$ and one in $C \cap C'$
3. any arc with one endpoint in $C \cap C'$ and one endpoint in C'

Consider Figure 17 on page 94, and let {11,12,13,14,15} be subgroup C and let {15,16,17,18,19,20} be subgroup C' . Arc $(13, 17)$ has one endpoint in $C \backslash C'$ and one in $C' \backslash C$, but neither endpoint in both, so $(13, 17)$ belongs in the first set. Arc $(14, 15)$ has one endpoint (node 14) in $C \backslash C'$ and one endpoint (node 15) in $C \cap C'$, so $(14, 15)$ belongs in the second set. Arc $(15, 20)$ has one endpoint (node 15) in $C \cap C'$ and one endpoint (node 20) in $C' \backslash C$. The min of each of these three arcs will become final weight on arc (C, C'). It is worth noting that the arc weights in the aggregated network do not consider arc weights between individuals wholly within one subgroup. This includes arc weights on arcs such as $(19, 20)$ and $(4, 5)$. The weighted arcs are provided in Figure 27. These weights show only interactions of the entire subgroup; this assumes that one member of the group represents the entire group.

It is worth taking a moment to consider what these arc weights really represent. Consider the 1/5 weight connecting subgroups {1,2,3,4,5}. That weight is the minimum of any pair of weights in which one endpoint is in {1,2,3,4,5} and the other is in {4,5,6,7}, excluding the arc $(4, 5)$. Recalling the weighted arcs in the disaggregated network in Figure 24 on page 106, this minimum measure is on two arcs $(3, 5)$ and $(5, 7)$. Therefore, the information or goods can get from subgroup {1,2,3,4,5} to {4,5,6,7} (or vice versa), but relies on node 5, one of the liaisons between the two groups. It is the same situation with every other pair of subgroups, even in the pairs of subgroups that have a connection that avoids the cut-out (specifically arcs $(13, 17)$ and $(10, 20)$). This implies that when considering the minimum speed of information or goods that reach the entire network, the direct relationships do not decrease the speed. Moving goods or information relies on the liaison individuals who have membership in more than one subgroup in this scenario. If the liaison individuals are not

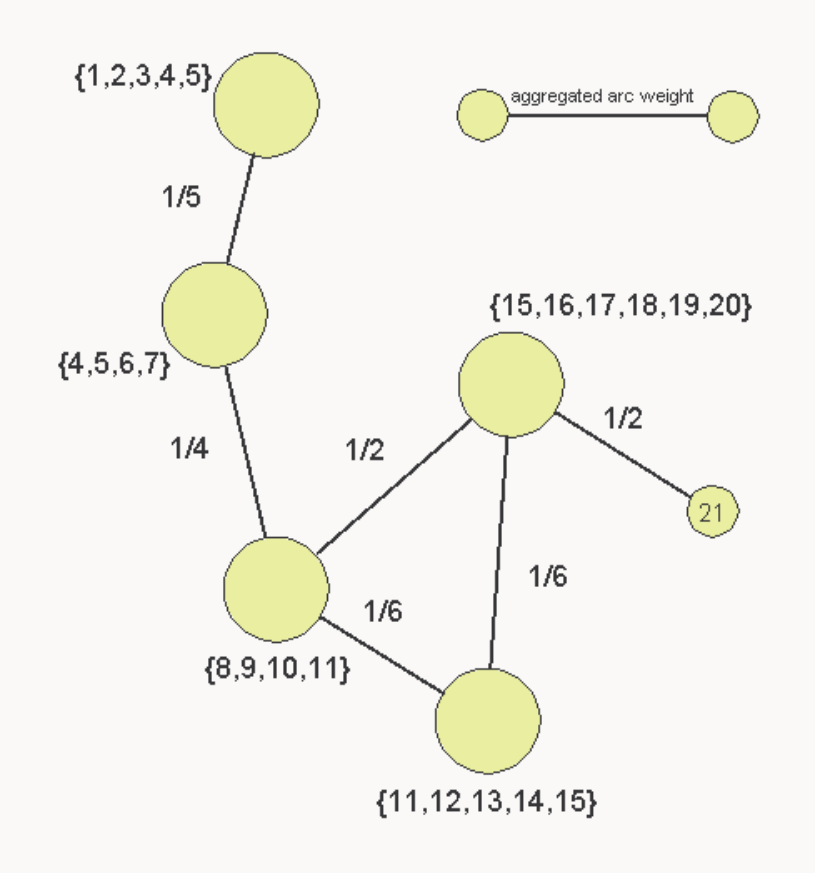

Figure 27. **2-plex aggregated arc weights for OSCC**

efficient at performing this function, it is likely arcs (10,20) and (13,17) would become more heavily utilized to avoid the inefficient liaisons.

The aggregation step now assigns to each subgroup or remaining individual a weight equal to the sum of shortest paths to the rest of the network. Figure 28 shows each subgroup's positional weight. These weights represent the speed with which a subgroup can pass information or goods to the rest of the network. It assumes that each subgroup works as a whole; i.e. in subgroup {8,9,10,11}, as soon as the subgroup receives a package from {4,5,6,7} it immediately passes it on to subgroup {15,16,17,18,19,20}, ignoring the fact that individual 8 actually receives the package and 11 has to send it on to the next subgroup. If critical to the analysis, a node delay might be added to the aggregated node. In the case of subgroup {8,9,10,11} having to send a package to {11,12,13,14,15}, this is not an issue, since 11 is a member of both subgroups. Again, it becomes clear these individuals in multiple nodes are very important.

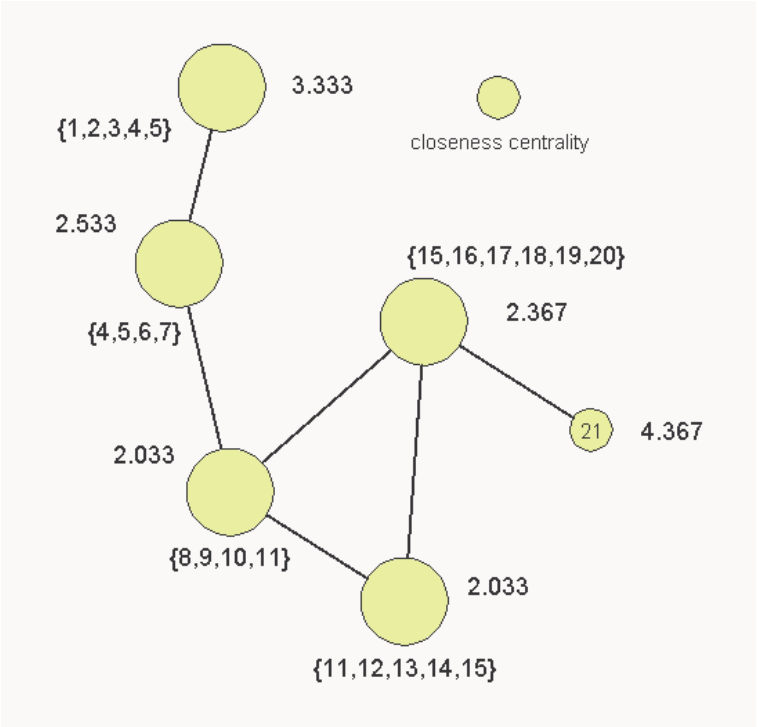

Figure 28. **2-plex aggregated subgroup weights for OSCC**

Figures 28 and 26 both show an aggregated social network with positional subgroup weights. Figure 28 considers only the speed of communications between groups, while Figure 26 also takes into account speed within the subgroup. It is interesting to note that in both figures, each of the six subgroups holds the same relative position in the order of increasing node weights:

1. {8,9,10,11}
2. (tie) {11,12,13,14,15}
3. {15,16,17,18,19,20}
4. {4,5,6,7}
5. {1,2,3,4,5}
6. {21}

4.4 Summary

The analysis in this chapter has demonstrated the four techniques introduced in Chapter 3 with four notional networks. Each of these notional networks have complete information and weighted arcs. These scenarios are informative to exhibit the information obtained from the two node measures and the subtle insight gained from the different orders of aggregation. The definition of arc weights in the notional examples are but a small number of appropriate arc weights. Any additive weight can be used with these techniques analysis, such as distance, time, length, frequency, cost, speed, etc.

This chapter demonstrates the techniques to find appropriate subgroups and aggregate networks introduced and detailed in Chapter 3. The aggregation techniques allow the analyst to consider a local property through the extended degree centrality measure, and a global property through the extended closeness centrality measure. The four scenarios demonstrate the techniques on notional networks, designed specifically to be appropriate for each of the subgroup detection/node measure pairs introduced in Chapter 3. The use of the degree and/or closeness centrality measures is scenario dependent, but using both can provide more information about the network interactions and structure.

It is tempting to aggregate leaves (nodes of degree 1) immediately into the node to which they are adjacent. However, the analysis in this chapter has shown that leaves can offer the analyst a great deal of information, as well as potentially good targets for exploitation, in some situations.

Any aggregation analysis can be performed with positional node weights or relational arc weights. All of the analysis in this thesis assume the analyst begins with weights on the arcs, representing the strength of relationship between the two endpoints. However, in the course of the analysis, nodes become weighted as well; thus, if the analyst begins with weighted nodes, representing the positional weight of an individual or subgroup in the network, he can pick up the aggregation process there.

Each of the four techniques detailed in Chapter 3 and explored through examples in this chapter, demonstrate first aggregation after each node has been assigned a node measure, and second aggregation with weighted arcs, after which each subgroup is assigned a node measure. The first includes within group interactions, providing the analyst with information about activity in the network as a whole. The second explores only between group interactions, allowing the analyst to detect more subtle changes in the network. Neither of the two orders of the process of aggregation and assigning a node measure are preferable in every situation. Occasionally, the scenario may make only one order applicable. In general, however, it is recommended that the analyst perform both to obtain greater information about the network.

While this chapter examined notional networks of complete information, the next chapter examines a real-world social network of incomplete information through open source data on Jema'ah Islamiyah, and exhibit how to approach a not-so-perfect social network.

Chapter 5 - Analysis of Jema'ah Islamiyah

5.1 Introduction

Jema'ah Islamiyah (JI) is a terrorist group, allegedly connected to Al Qaeda. This analysis to characterize the structure and activity of JI is approached in the two steps Chapter 3 recommends: 1. determine appropriate subgroups, and 2. aggregate using degree and closeness centrality measures to learn the level of within and between subgroup interactions. First, section 5.2 seeks appropriate subgroups for aggregation. Then, Sections 5.3 and 5.4 demonstrate the application of the degree centrality and closeness centrality node measures, respectively, to the aggregated network. This allows investigation of both a local and a global property for this network to obtain as much information as possible.

Figure 29 shows an undirected, unweighted network of individuals in JI from *open source* information as of April, 2003. As an open source example, it is recognized that the information is possibly both incomplete and inaccurate. Much of the information is obtained from individuals who have been apprehended, and is therefore quite biased and incomplete. It is not presented here as an operational analysis of JI. It is offered to illustrate the approaches developed in this research on an open source, real world network. The arcs represent the existence of a relationship between individuals. Since the strength of the relationship between individuals are unknown, each arc is assigned a unity weight.

It is worth noting that the structure of the network may be an artifact of the the method by which data is collected. Many of the individuals in the network have been apprehended, so more information is known about them and their relationships with other individuals. It is important to remember that this may cause some bias in the network structure, and therefore the insight gained. The data is collected in such a manner that the individuals are not ordered sequentially, as they had been the previous notional examples. The names associated with each of the individual numbered nodes can be found in Appendix C.

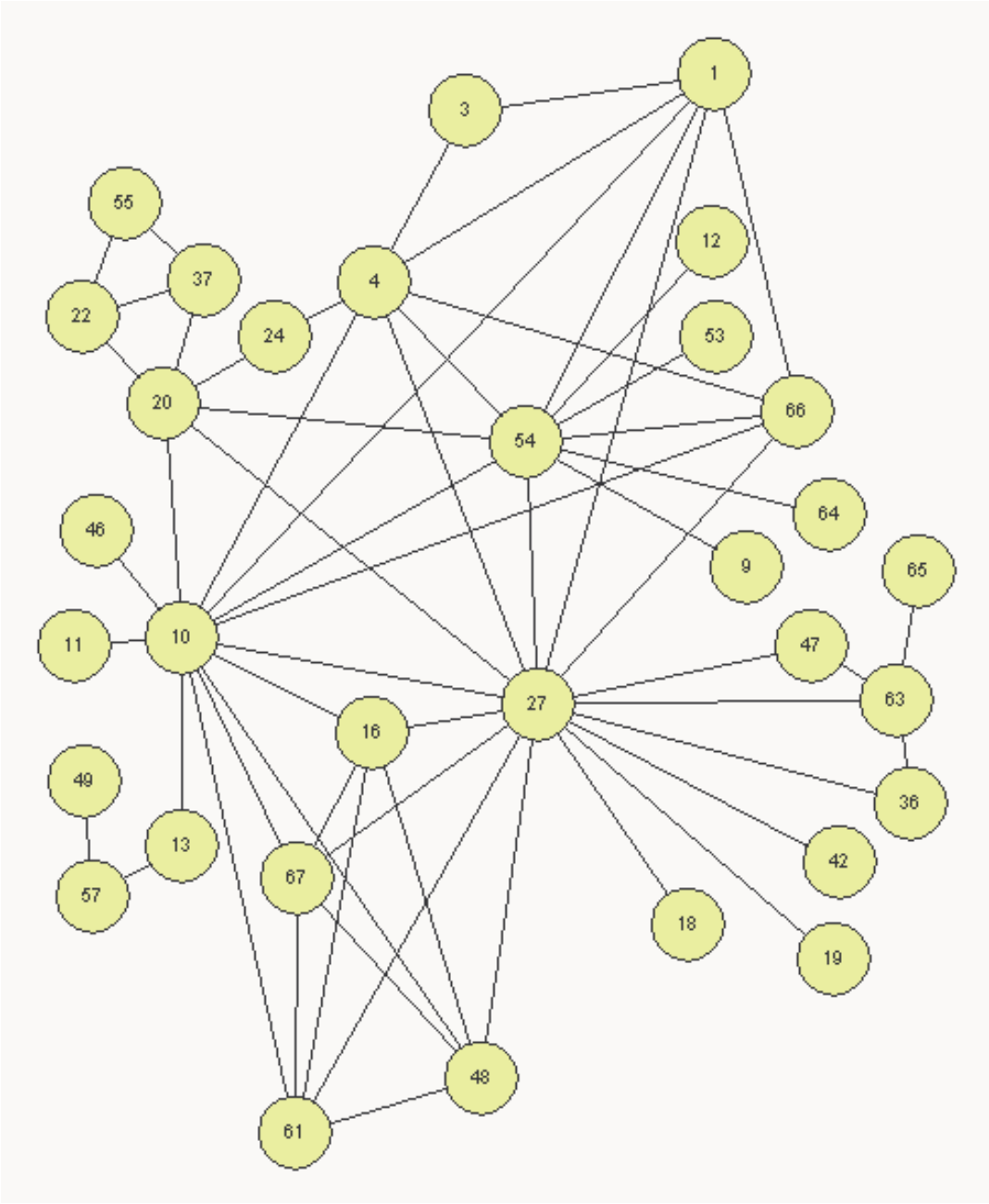

Figure 29. **Disaggregated Open Source Network Structure for JI**

5.2 Subgroup Detection in JI Network

An initial review of the network in Figure 29 shows there are no obvious distinct subgroups. All cliques in the network on three or more nodes are:

$\{1, 3, 4\}$
$\{1, 4, 10, 27, 54, 66\}$
$\{10, 16, 27, 54, 66\}$
$\{10, 16, 27, 48, 61, 67\}$
$\{10, 20, 27, 54\}$
$\{20, 22, 37\}$
$\{22, 37, 55\}$
$\{27, 36, 63\}$
$\{27, 47, 63\}$

It is tempting to think that since both 22 and 66 are in two cliques, they are in some way equal in importance. However, 22 is in two K_3s, while 66 is in a K_4 and a K_5. In fact, nodes 10, 27, and 54 are also in both the K_4 and the K_5, implying that 10's power to control communications between the two cliques is not as strong as if he were the only liaison between the two cliques.

The aggregation of the individuals into cliques does not greatly reduce the size of the network. This suggests representing subgroups with k-plexes (for some empirically determined k). Several values of k and the associated k-plexes are shown in Table 12.

Note that as k increases, the subgroups contain more individuals and become less dense. It is also interesting to note that the subgroups are simply overlapping more, rather than encompassing nodes not already in a clique. Since increased relaxation of the clique to k-plexes serves only to decrease the distinctiveness of subgroups, the rest of the analysis of this network uses 2-plexes. Figure 30 shows the aggregated network used in the remainder of the analysis of JI, and the aggregated subgroups are :

$\{1, 3, 4, 10\}$
$\{1, 4, 10, 27, 54, 66\}$
$\{1, 10, 20, 27, 54\}$
$\{10, 16, 27, 48, 61, 67\}$
$\{20, 22, 37, 55\}$
$\{27, 36, 47, 63\}$

Nodes 1, 4, 10, 27, and 54 are liaison individuals in multiple subgroups, and therefore merit consideration. Individuals in multiple groups interact with a greater number of people and are

Table 12. k**-plexes for JI**

k	k-plex
2	$\{1, 3, 4, 10\}$
	$\{1, 4, 10, 27, 54, 66\}$
	$\{1, 10, 20, 27, 54\}$
	$\{10, 16, 27, 48, 61, 67\}$
	$\{20, 22, 37, 55\}$
	$\{27, 36, 47, 63\}$
3	$\{1, 3, 4, 10, 27\}$
	$\{1, 4, 10, 20, 27, 54\}$
	$\{1, 4, 10, 27, 54, 66\}$
	$\{10, 16, 27, 48, 61, 67\}$
	$\{20, 22, 37, 55\}$
	$\{27, 36, 47, 63\}$
4	$\{1, 3, 4, 10, 27, 54\}$
	$\{1, 4, 10, 20, 27, 54, 66\}$
	$\{1, 4, 27, 36, 47, 63\}$
	$\{10, 16, 27, 48, 61, 67\}$
	$\{10, 20, 22, 27, 37, 55\}$
6	$\{1, 3, 4, 10, 20, 27, 54, 66\}$
	$\{1, 4, 10, 16, 27, 48, 54, 61, 66, 67\}$
	$\{1, 4, 10, 27, 36, 47, 54, 63\}$
	$\{1, 10, 20, 22, 27, 37, 54, 55\}$
10	$\{1, 3, 4, 10, 16, 20, 22, 27, 37, 54, 55, 66\}$
	$\{1, 3, 4, 10, 16, 20, 27, 36, 47, 54, 63, 66\}$
	$\{1, 3, 4, 10, 16, 20, 27, 48, 54, 61, 66, 67\}$

involved in more activities than anyone else in the network. Changes in their behavior may portend a change in overall network activity or imminent action.

It is important to consider the structure of the aggregated network in Figure 30. The network gives the appearance of more connectedness than truly exists. Consider, for example, nodes 18 and 19, each of which have degree 3. In Figure 29, these same nodes have only degree 1, and are adjacent to node 27. Since node 27 is a member of three aggregated subgroups, all of its neighbors (including 18 and 19) are also adjacent to each of the subgroup nodes into which 27 is aggregated. Care must be taken to not become overwhelmed by the number of arcs in the aggregated network.

5.3 Application of degree centrality node measure to JI

This analysis first considers the application of the local property, degree centrality, to the open source network. This measure assigns to each individual or subgroup a weight equal to the sum of the weight on arcs incident to the node which represents that individual or subgroup. Since the

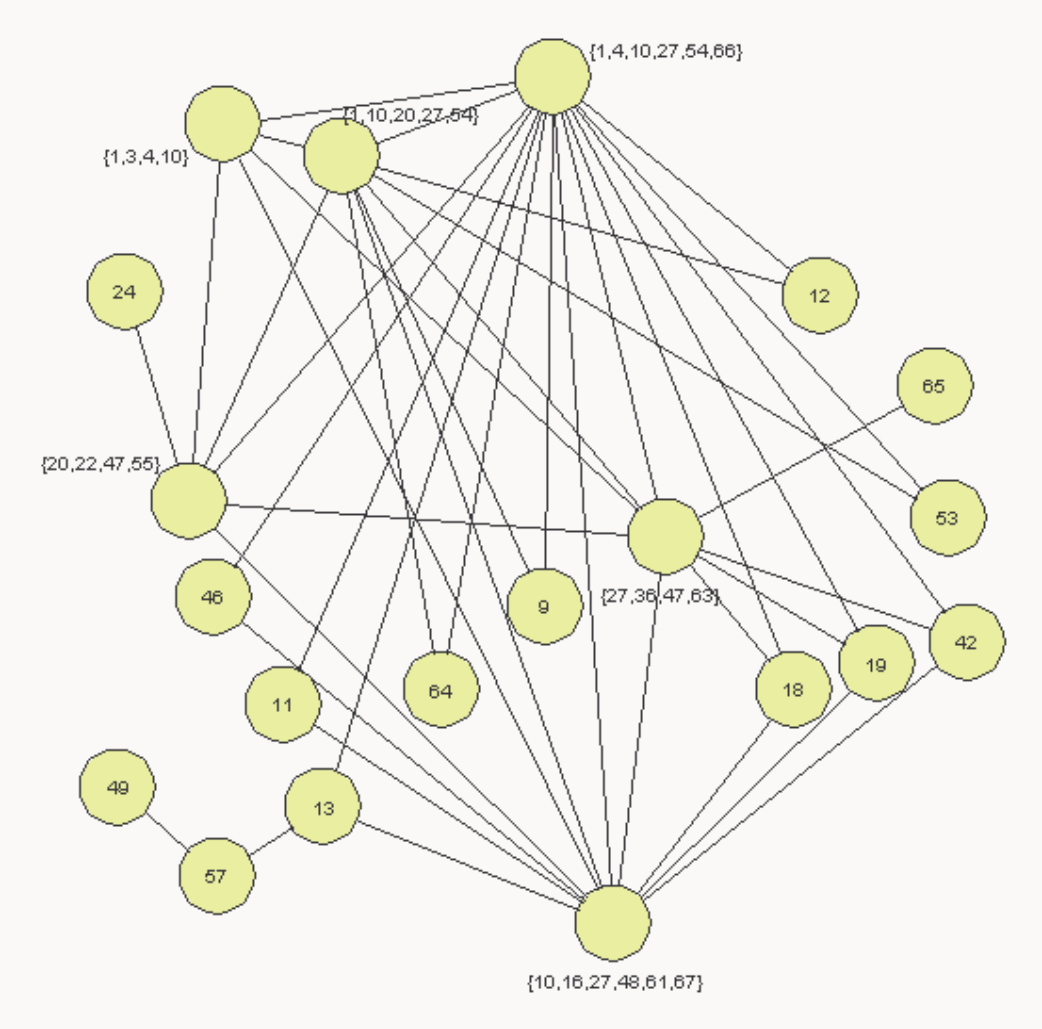

Figure 30. **2-plex aggregated structure for open source JI**

network is undirected and unweighted, this measure will equal the degree of the node. All the degree centrality calculations were performed in 0.6710 seconds of computer time on a 2 ghz Dell Dimension XPST800 Pentium 4M.

5.3.1 Assign degree centrality node measures before aggregating

Each individual in the network is first assigned a degree centrality measure equal to the node's degree. These weights are shown in Figure 31. Not surprisingly, the individuals in multiple large subgroups have the highest measures. These are nodes 10 (Ali Gufron/Muklas), 27 (Hambali), and 54 (Imam Samudra). The degree centrality measure is a local property, and considers only an individual's relationship with immediate neighbors. This is why node 4 (Abu Bakar Bashir), though somewhat on the edge of the network, can still have a relatively high measure for this network of 7.

The aggregation step now replaces individuals in the network with the appropriate subgroups, previously determined, listed on page 117. The weight on each remaining individual does not change in the aggregation step, but each aggregated subgroup is assigned a weight equal to the sum of the individual weights of its members. The weighted nodes in Figure 32 measures the relative strength of the relationships of members of each subgroup. This means, for example, that node 27 (Hambali) assigns his weight to subgroups {1,4,10,27,54,66}, {1,10,20,27,54}, {27,36,47,63}, and {10,16,27,48,61,67}.

It was noted that nodes 10 (Ali Gufron/Muklas), 27 (Hambali), and 54 (Imam Samudra) have the highest individual weights; not surprisingly, the groups of which they are members, {1,4,10,27,54, 66} and {1,10,20,27,54}, also have the highest weights, 57 and 51, respectively. It was shown in Chapter 3 that theoretically this is not necessarily the case, in practice, it is likely to occur, as it has in this open source JI example. Since every arc is accounted for in each endpoint, the weights are large, but they provide the analyst with information about the relative strength of the relationships of the subgroups.

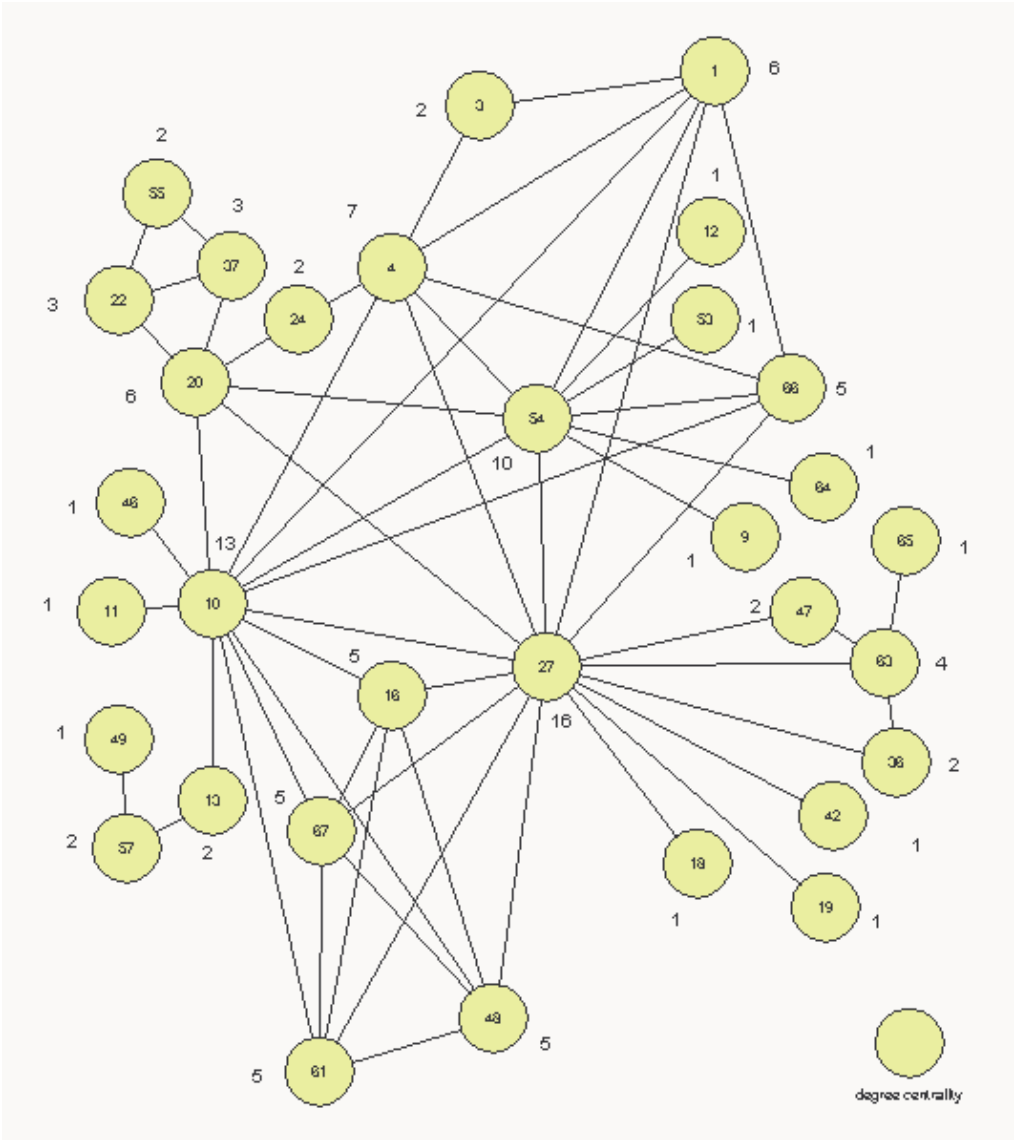

Figure 31. **Weighted individual nodes under OSDC for open source JI**

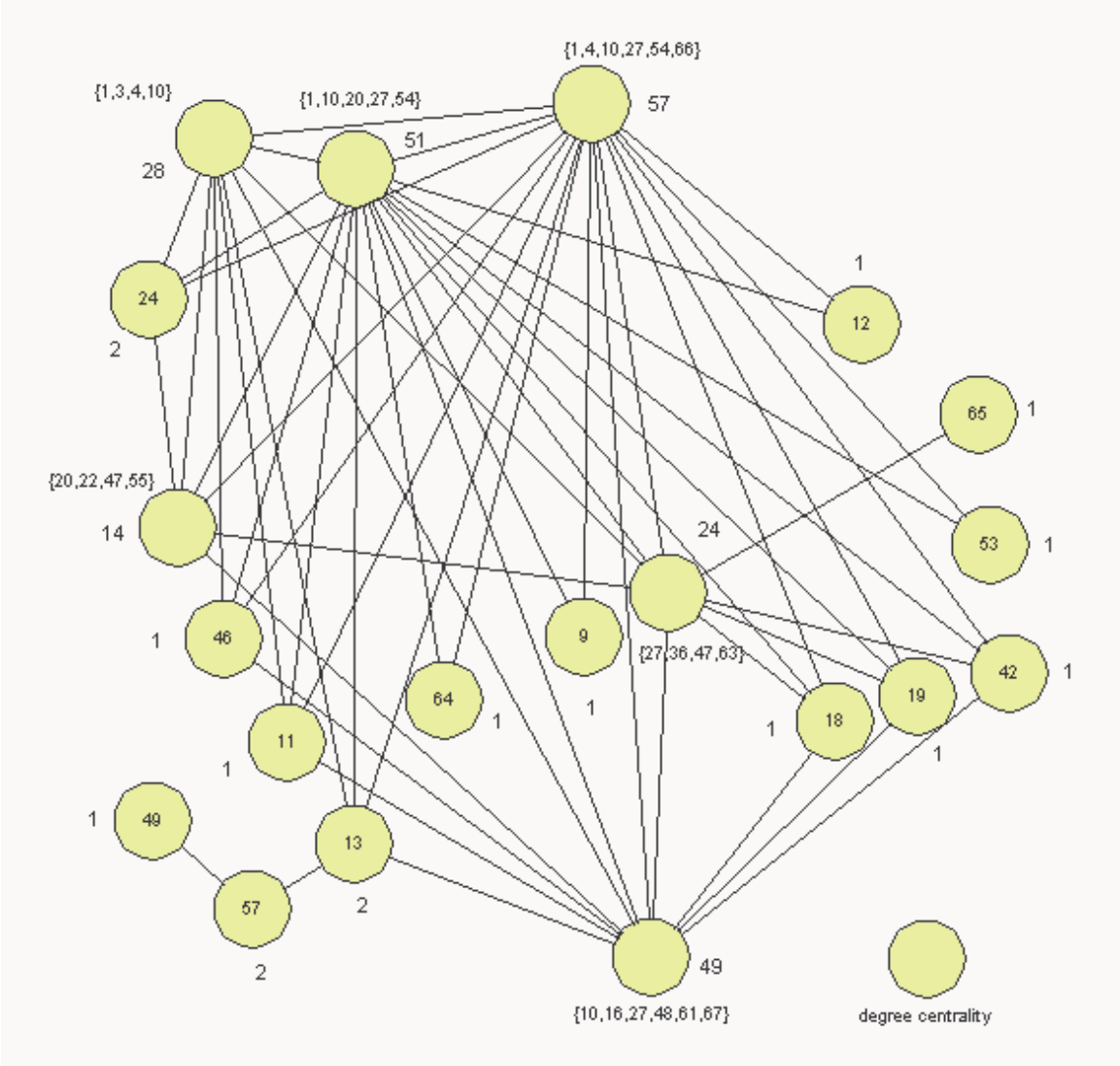

Figure 32. **2-plex aggregated network of within and between subgroup OSDC weights for open source JI**

5.3.2 Aggregate before assigning degree centrality node measures

The second method of aggregation explores only intergroup relationships. Any edges that exist wholly within a subgroup are not counted in the calculations. The network with aggregated arc weights is shown in Figure 33. Recall the weight of an arc connecting two subgroups (or individuals) is assigned a weight equal to the sum of the arcs in which exactly one endpoint is in each subgroup. For ease of examining the network in Figure 33, any unity weighted arc is left unlabeled.

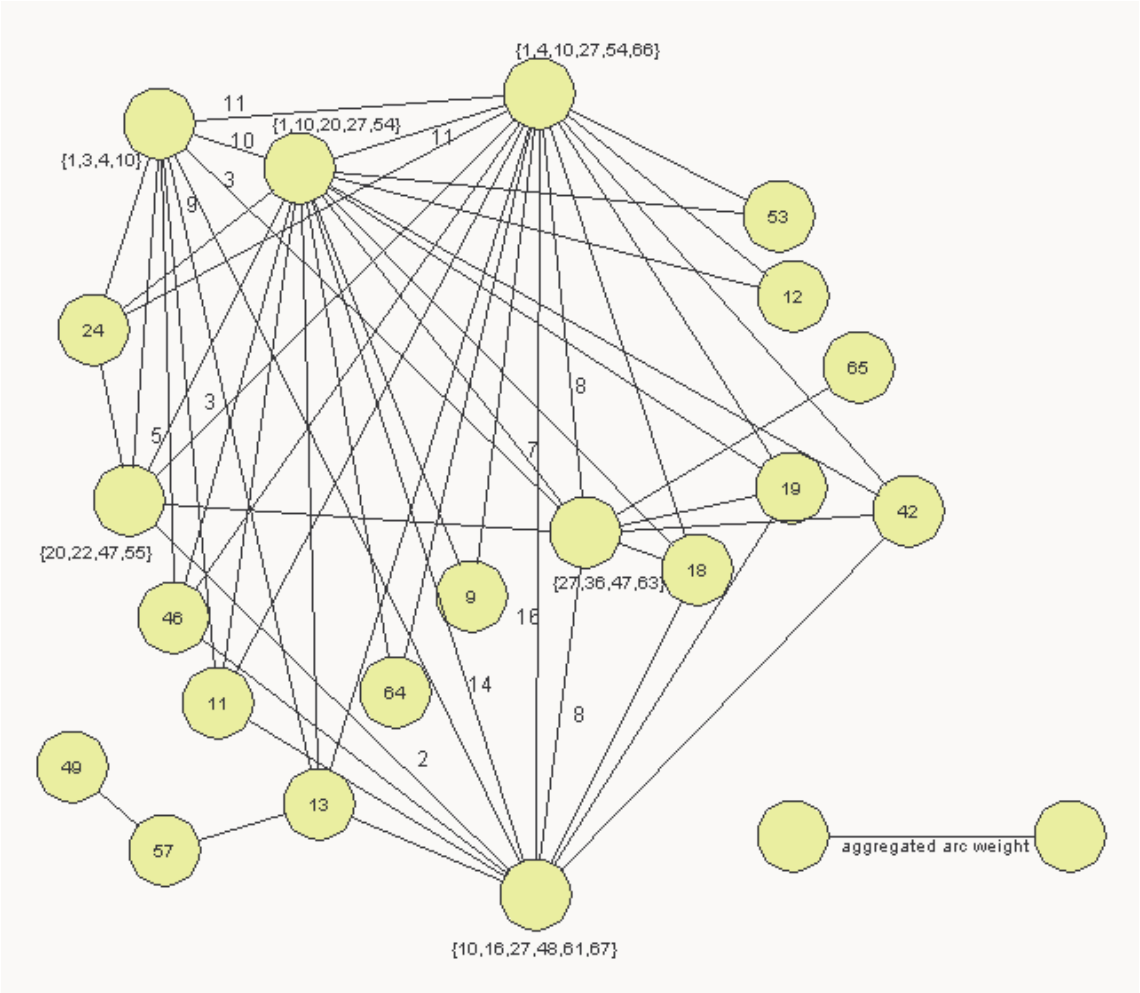

Figure 33. **2-plex aggregated arc weights for OSDC for open source JI**

Naturally the highest weighted arcs are those between the aggregated subgroups, as they not only have multiple members, but many have multiple members in common. Node 27 (Hambali),

which had the highest individual measure in Figure 31, is a member of the three subgroups with the highest arcs connecting the trio, {1,4,10,27,54,66}, {1,10,20,27,54}, and {10,16,27,48,61,67}. The weighted subgroup nodes in Figure 34 show interactions between adjacent subgroups.

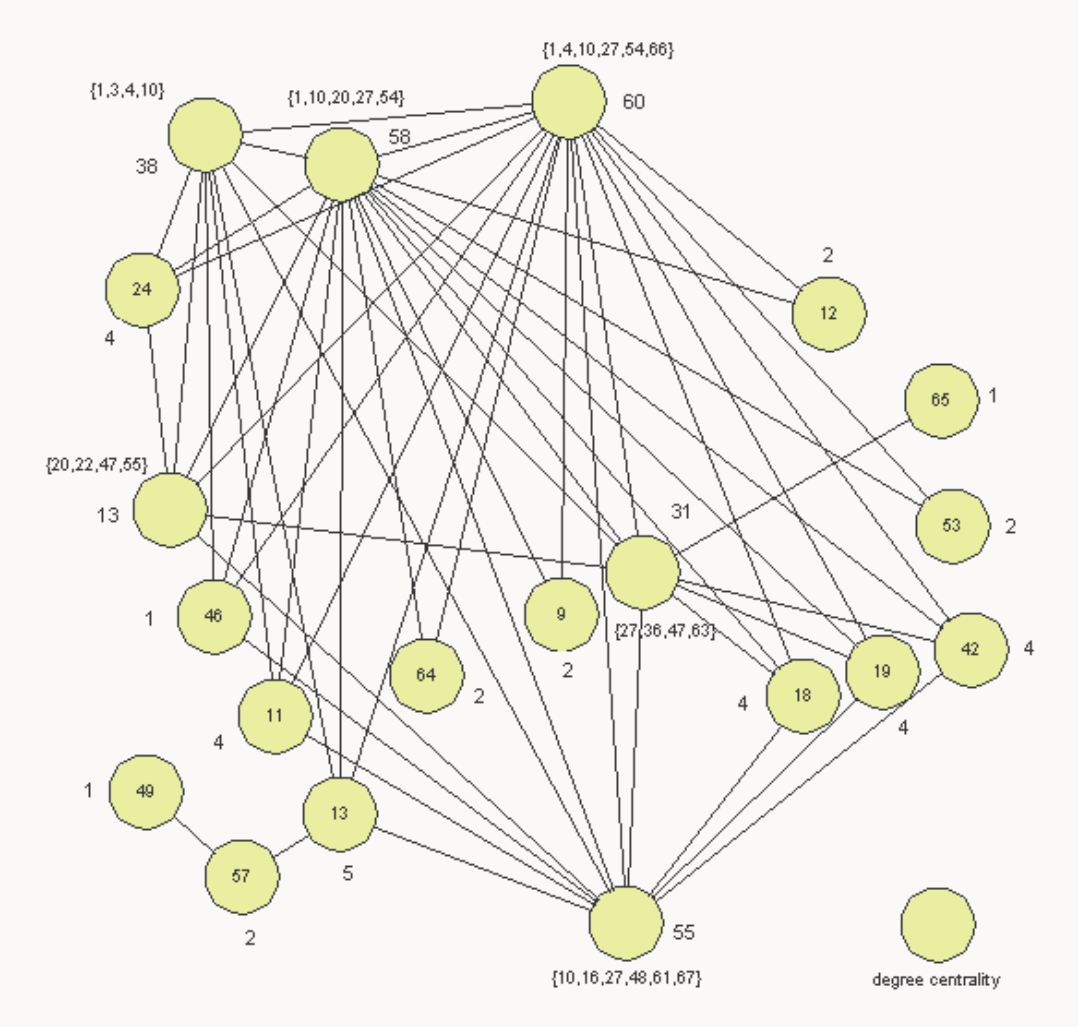

Figure 34. **2-plex aggregated between subgroup OSDC weights for open source JI**

Leaf nodes that appear unimportant in the disaggregated network in Figure 29 such as 18,19, or 42, seem much more integrated into the network in the aggregated network. These three nodes now have a measure of 4, but only because each is adjacent to 27, who is in 4 aggregated subgroup nodes. This method can therefore be used to identify an individual who is not himself an inner circle member, but is connected to such an individual. Node 42, for example, has a relationship

with the clearly important node 27, but node 42 is on the fringe of the disaggregated network. Node 42 may possibly be a susceptible target.

Both of the degree centrality aggregation methods produce an aggregated network with positional node measures as a final product. Note that in this example, unlike some of the earlier notional examples, the five largest subgroups order the same weight in both techniques. The subgroups, in order of decreasing weight, are:

1. $\{1, 4, 10, 27, 54, 66\}$
2. $\{1, 10, 20, 27, 54\}$
3. $\{10, 16, 27, 48, 61, 67\}$
4. $\{1, 3, 4, 10\}$
5. $\{27, 36, 47, 63\}$

Naturally the liaison individuals have proven to have the highest degree centrality measure, since they communicate directly with the most number of other people. The individuals with low weight should not be discounted, however, since they also can be viable intelligence targets. Consider nodes 18, 19, and 42, who are adjacent to 27. Any of these three could provide information on 27. Similarly, nodes 9, 12, 53 and 64 are adjacent to 54.

This section examined the large component of JI in an open source network using a local property, degree centrality. This property examines an individual or group only in the context of its immediate neighborhood, without consideration of its place in the network. The next section re-examines the network with a global property using the closeness centrality node measure. This considers the cumulative sum of total length of all paths an individual or subgroup must use to disseminate information to every other individual or subgroup in the network. The entire network structure affects these measures, so individuals who have positioned themselves to give orders quickly have the lowest weight. It is expected that the liaison individuals identified earlier as highly communicative (1, 4,10, 27, and 54) can also disseminate information quickly and have low closeness centrality measures. The subgroups of these individuals will also have small weight. Using the

global property, those on the fringe of the network will have high weights, as they are not as well integrated into the network.

5.4 Application of closeness centrality node measure to JI

In the aggregation, each subgroup is assigned a weight equal to the minimum of the weights of its members. Individuals who are not aggregated keep their individual weight in the aggregation step. Alternatively, in the second order of analysis in section 5.4.2, the individuals are aggregated into subgroup first. The methodology introduced in Section 3.4.5 assigns to each arc connecting a pair of subgroups a weight equal to the minimum among arcs with exactly one endpoint in each subgroup. However, since each arc in this network is unity weighted, each arc connecting aggregated subgroups also has unity weight. Each subgroup or remaining individual is then assigned a weight equal to the sum of shortest paths from it to every other node. All closeness centrality calculations were accomplished in 289.1060 seconds of computational time on the 2 ghz Dell Dimenstion XPS T 800 Pentium 4M. Computationally, these calculations have taken longer since the aggregated network has many more nodes than any of the notional examples.

5.4.1 Assign closeness centrality node measures before aggregating

Figure 35 shows the closeness centrality individual weights. Naturally, the largest weights are on individuals at the fringe of the network, such as node 49, who is on a path of length three from a denser part of the network. The five smallest weighted nodes in the network, in increasing order of weight are 27, 10, 54, 20, 1. These are also exactly the members of one of the subgroups. It is therefore expected that this subgroup will also have the least weight after aggregation.

It is interesting to consider the great disparity that exists on individual weights within a subgroup. Consider, for example, the subgroup $\{20, 22, 37, 55\}$. The node that forms the liaison to the rest of the network, 20, has the smallest measure. The largest weight in the subgroup, on node 55, which is only a path of length 2 away from node 20, has a weight of 117, which is almost double the weight of node 20, 61. Another subgroup on four nodes, $\{27, 36, 47, 63\}$ does not have such a great difference of weights within the subgroup, due to its structure and how it connects to the rest

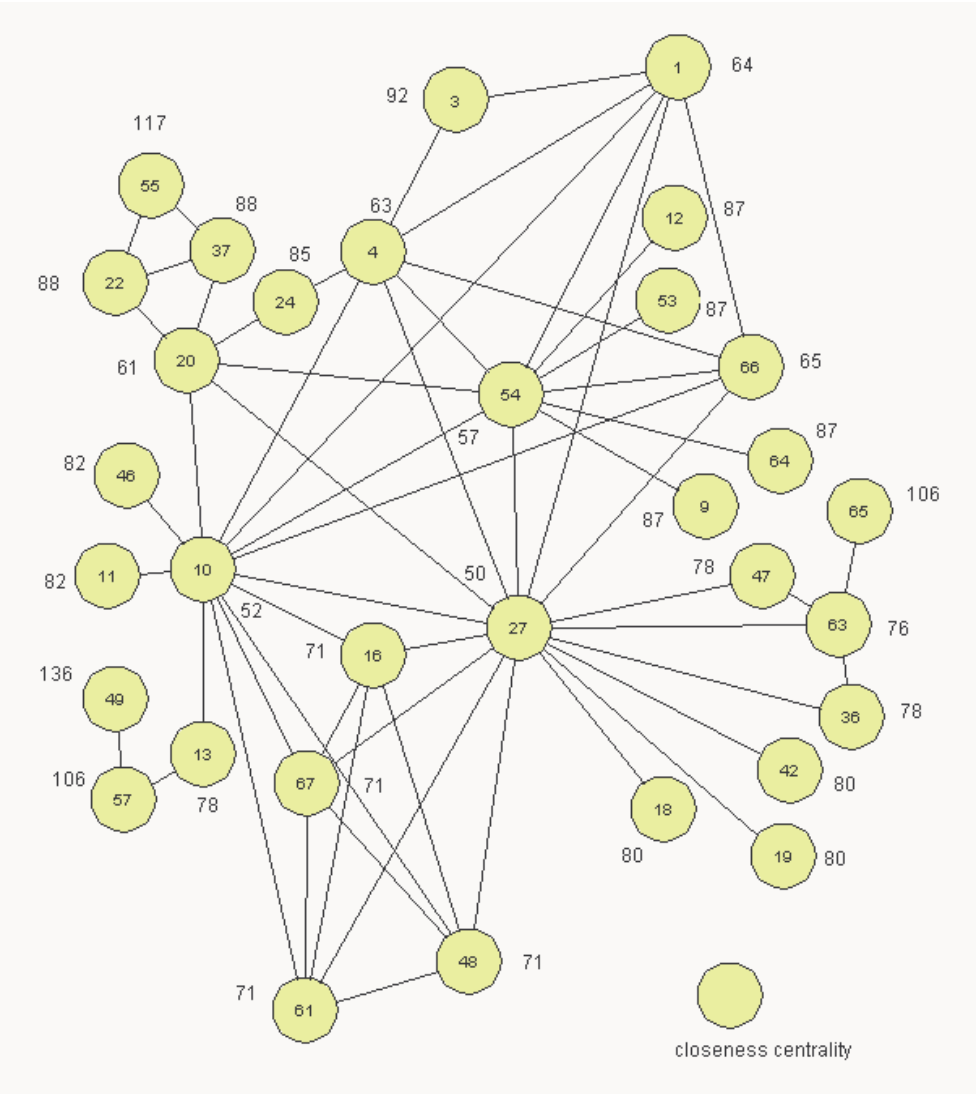

Figure 35. **Disaggregated network of node weights under OSCC for open source JI**

of the network. In this latter subgroup, the liaison node, 27, is adjacent to every other member of the subgroup, whereas in the former subgroup, the liaison node, 20 is not adjacent to node 55, which has the greatest weight. Consider also the clique on six nodes, $\{10, 27, 16, 67, 48, 61\}$. It is worth pointing out that the way the network is drawn make nodes 48 and 61 appear to be less involved in the network than nodes 16 or 67. However, all four nodes are structurally the same, as evidenced by their node measures: all a weight of 71. In fact, a review of the network in Figure 31, listing the individual degree centrality weights, shows these four nodes to have the same weight as each other, which is not surprising if they are structurally the same. Of course it should be recalled that the data used is all open source and potentially biased by the large number of apprehended individuals in the data.

Figure 36 shows the closeness centrality weights for the subgroups and remaining individuals after the aggregation step. Each subgroup receives a weight equal to the minimum weight of each member in the subgroup. Individuals who are not aggregated keep the same weight after the remainder of the network is aggregated.

Node 27, who has an individual measure of 50, gives his weight to each of the aggregated subgroups of which he is a member. These weights represent how quickly some individual in each subgroup can disseminate goods or information to everyone else in the network. For this to be a realistic interpretation, it is necessary to assume that the groups are very cohesive, and an individual can represent the entire group. This may be a reasonable assumption if it is known each group has a leader or a go between who is responsible for goods and information flowing into and out of his subgroup.

5.4.2 Aggregate before assigning closeness centrality node measures

This aggregation procedure, in which the subgroups are aggregated before node measures are assigned to each aggregated subgroup or individual, examines the interactions between subgroups without the noise of within subgroup interactions. If data on relationships in JI are taken regularly

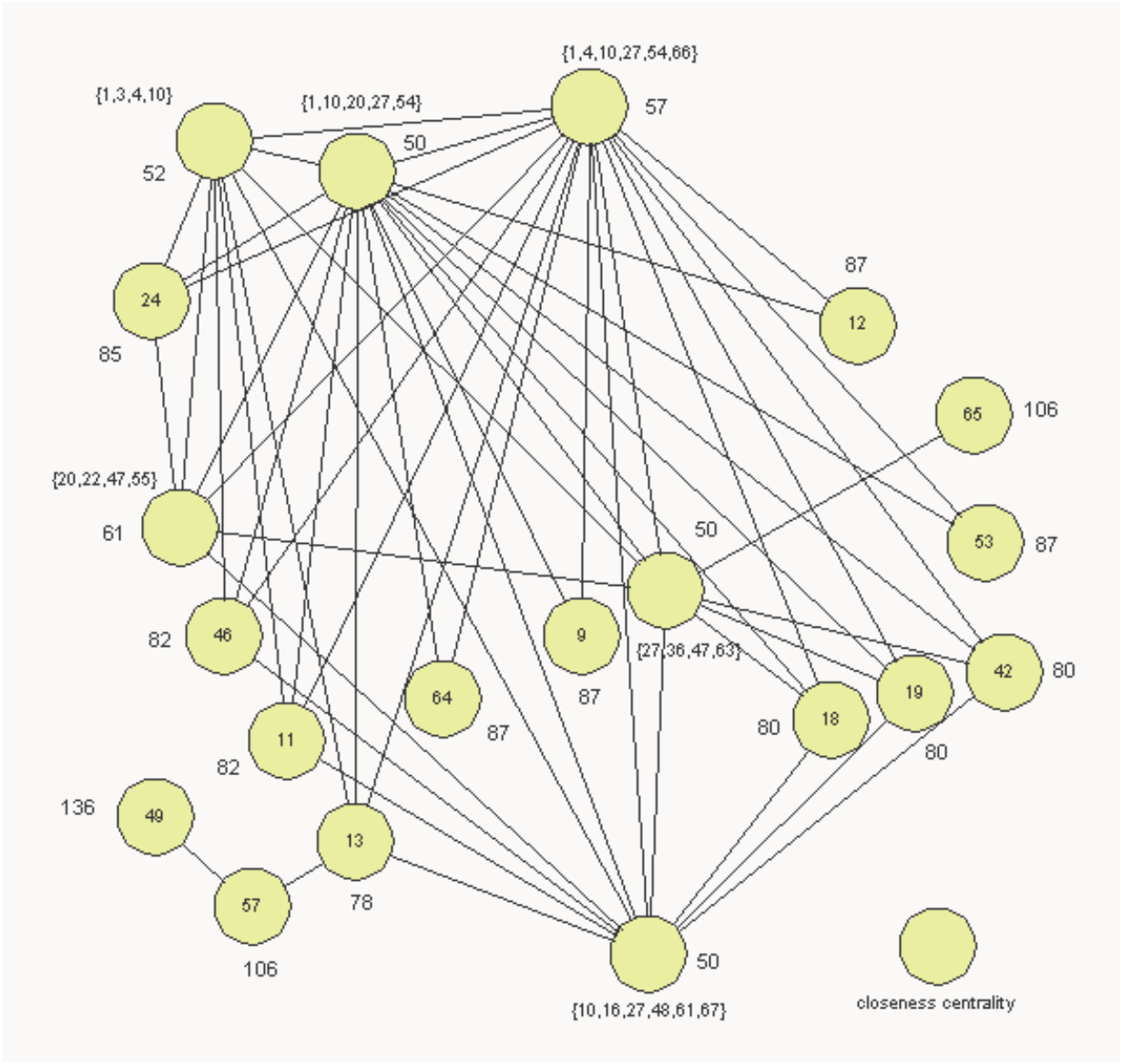

Figure 36. **2-plex aggregated node weights for OSCC for open source JI**

over time, then this procedure can help to determine subtle changes in subgroup interactions, which may be indicative of increased activity and future hostile actions.

This procedure assigns to an arc connecting two aggregated nodes a weight equal to the minimum of arcs with exactly one endpoint in each subgroup. Since every arc has unity weight, every arc in the aggregated network also has unity weight.

Now each arc is equal in its length for traversal. Since the closeness centrality measure is a global property, the subgroup nodes weights show positional superiority in the network for the dissemination of goods and information. These weights can be seen in Figure 37, where a low weight indicates a greater ability to disseminate information throughout the network. Not surprisingly, the individuals at the fringe of the network have the highest weights. It is interesting to consider the disparity of the weights on those individual nodes of degree 1, such as nodes 46 (weight of 38), 53 (weight of 40), and 49 (weight of 63). Merely having degree 1 does not make a node unworthy of further study. It has already been noted that a leaf could be a person new to the network, and easily influenced, or an individual who interacts with an important person in the network in a different context, in which case the leaf can be exploited. Thus when considering node 49, who has the highest weight in the network, it may be reasonable to deduce that he does not wield a great deal of influence. However, when considering node 18, who has a fairly low weight of 37, it is clear that he is an individual who interacts directly with node 27, a person who does exert a lot of influence in the network as a whole.

It was supposed earlier in this section that the subgroup $\{1, 10, 20, 27, 54\}$ would appear most efficient as disseminating information and goods, since its members are the five individuals with the lowest weight. Indeed, this subgroup does hold the lowest weight, but it shares that position with another subgroup, $\{1, 4, 10, 27, 54, 66\}$. These two subgroups share three members, so it is not too surprising that their ability to reach out to the rest of the network is similar.

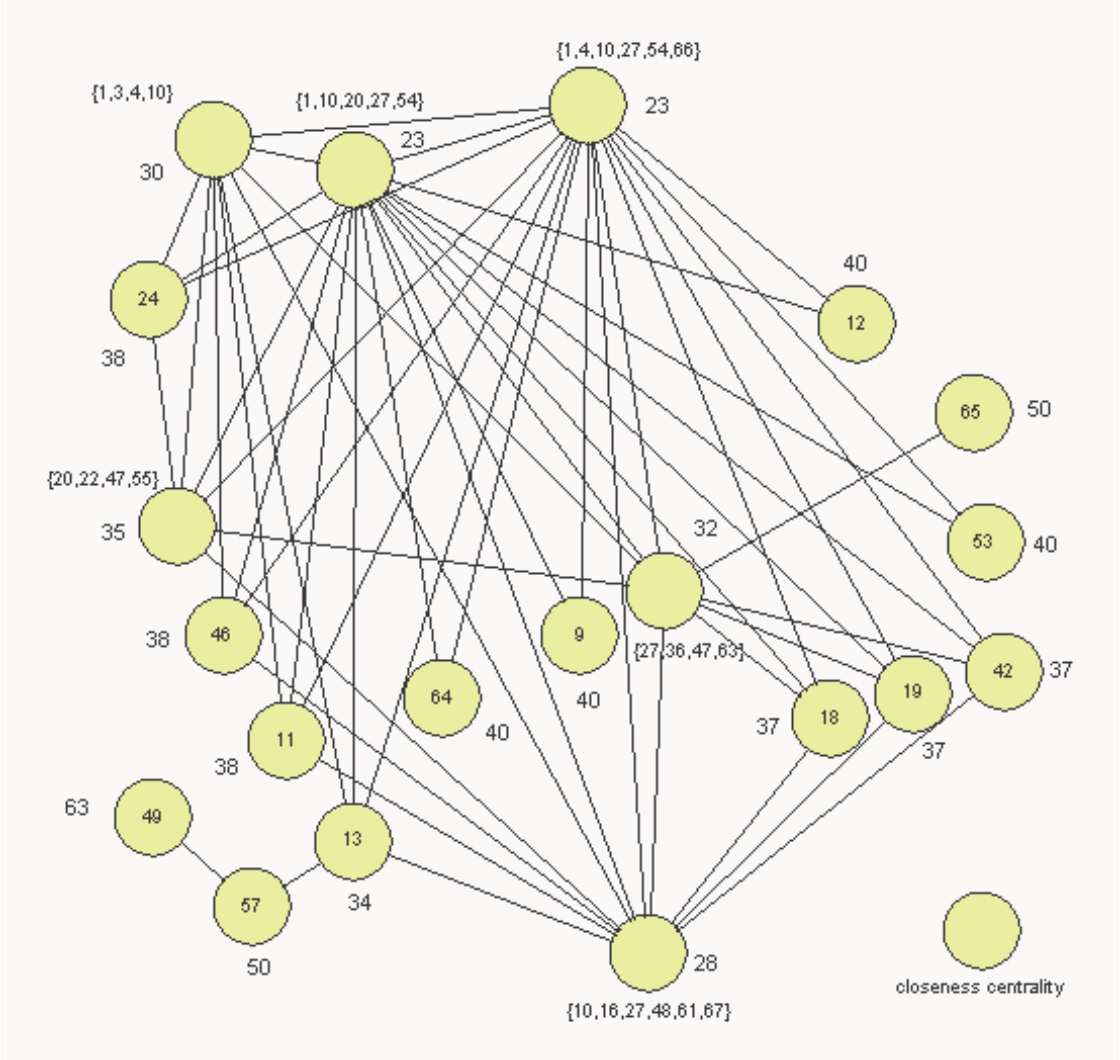

Figure 37. **2-plex aggregated subgroup weights under OSCC for open source JI**

5.5 JI Analysis Summary

This chapter used the methodology introduced in Chapter 3 to apply to a scenario *not* built specifically to demonstrate a technique (as Chapter 4 was). The open source JI data is real-world, and probably biased by the number of apprehended subjects. It is therefore "messier" than the notional networks examined in Chapter 4. The first step in the aggregation analysis of a new network is to determine appropriate subgroups. This network is small enough to allow a visual examination of the network structure, which showed several overlapping groups, along with some leaves. There is no obvious distinction in the subgroups, and detection of the pure cliques showed them to not appropriately define subgroups. Therefore, the non-overlapping subgroups methodology for k-plexes was used, and the 3-plexes were determined to be appropriate for the JI analysis

Since the interactions in the Jema'ah Islamiyah data are basically unknown, this research examined both the local and global properties to understand as much as possible about the structure and relationships of the network. The individuals acting as liaisons between the larger cliques, nodes 27, 10, 54, and 1, became the focus of the analysis very quickly. Every calculation performed on the network showed some subset of these individuals to exhibit the greater amount of influence. Often individuals connected to these four nodes also showed better measures than individuals who seem structurally the same in other parts of the network. It is therefore recommended to obtain more information on how to influence the network behavior, that these four individuals, and their immediate neighbors be the target of further investigation.

It is important when considering the analysis in this chapter, that this is not presented as a meaningful intelligence analysis of JI. This investigation is offered as a demonstration of the methods in Chapter 3 on a network of incomplete and inaccurate information. Thus, the reader's focus in this chapter should be on the application of aggregation techniques to gain information about the network rather than the actual results.

Chapter 6 - Exploration of Imperfect Information

6.1 Introduction

Modeling the effects of missing information in social networks is of great importance. Rarely do analysts have perfect information of the social group they wish to study. It is therefore necessary for the analyst to understand how missing information impacts the analysis of the network. This chapter provides the analyst with an introduction to the effects of missing information, by considering how unknown arcs can make it difficult for the analyst to identify appropriate subgroups in a network. Four distinct methods of arc removal protocol are identified in Section 5.2 and implemented in Section 5.3 to show how missing arcs inhibits the analyst's ability to find appropriate subgroups. This work is demonstrated on a network previously investigated in Chapter 4.

6.2 Methodology

The network used for this assessment is the one introduced in Section 4.2.2, with overlapping cliques, shown again here as Figure 38. Since this network has overlapping subgroups, some individuals will be in multiple subgroups. Recall the appropriate subgroups chosen for aggregation are the following set of 2-plexes:

$\{1, 2, 3, 4, 5\}$
$\{4, 5, 6, 7\}$
$\{8, 9, 10, 11\}$
$\{11, 12, 13, 14, 15\}$
$\{15, 16, 17, 18, 19, 20\}$

The impact of missing arcs is determined by removal of a given percentage of the arcs in the network, and then reassessing the subgroups formed. The user inputs a percentage of the total number of arc in the network to be removed and an adjacency matrix for the network under study. A MATLAB script outputs an adjacency matrix and boolean variable indicating whether the new adjacency matrix can produce the subgroups originally found in the complete network. A confidence interval is placed on $\widehat{p}$, the proportion of networks which still produce the same 2-plexes, for each of the four removal methods at several levels of percent arcs missing.

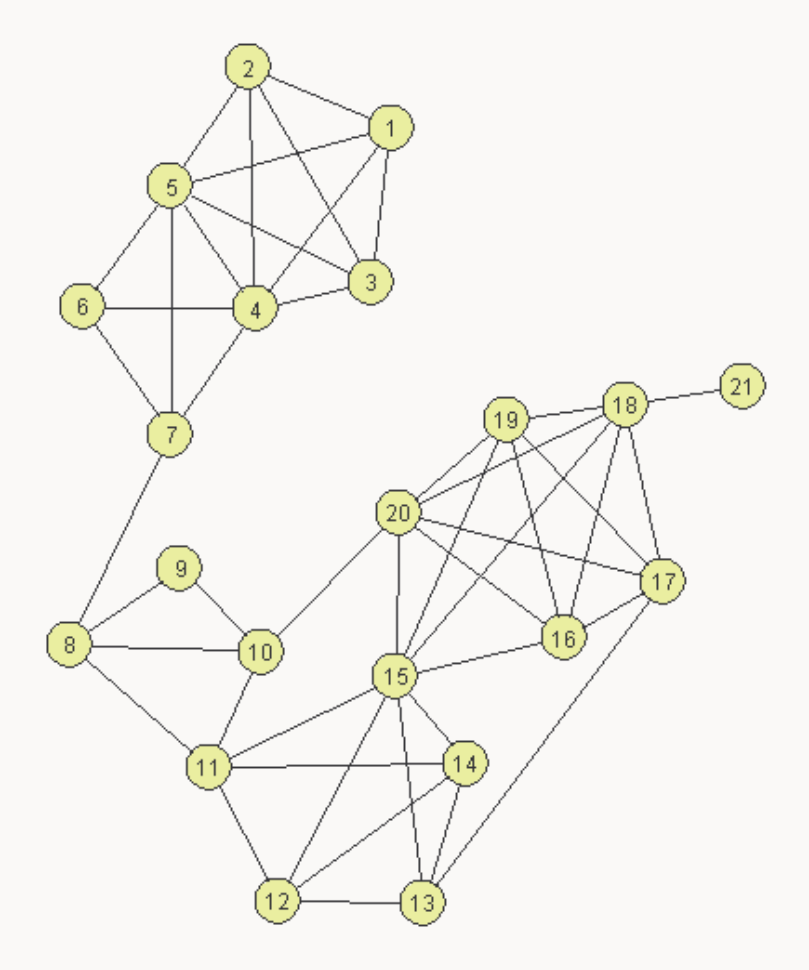

Figure 38. **Disaggregated network of complete arc structure**

A user-given percentage of the arcs are targeted for removal. In the experiment, the percentage starts at 2.5%, and increments by 2.5% until $\widehat{p} = 0$. The arcs are targeted in the following four ways:

1. any random arc
2. any random arc with each endpoint at nodes of lower than average degree
3. any random arc incident to individuals of higher than average degree
4. any random arc incident to individuals in multiple subgroups

The first approach assumes that any arc in the network is equally likely to be unrecorded. It is appropriate to consider this situation when there is a high degree of uncertainty about the individuals in the network. Suppose, for example, that a previously unknown group of people are under investigation, and the arcs in the network represent observed contacts. Any of the contacts are equally likely to be missed.

The second method for arc removal considers only those arcs connecting two individuals of low degree. This is appropriate when individuals believed to be low level associates are not being tracked with as much diligence. Again, if the arcs represent contacts, and only the prominent members are being watched, then any contact between individuals of low importance in the network may continue unnoticed.

The third method for arc removal focuses only on individuals with high degree. This method is appropriate when such individuals know they are highly interconnected and practice good operational security. Because of this heightened OPSEC, their interactions with other individuals are assumed to be more likely to be unknown.

As in the third method, the fourth investigates those individuals who act as liaisons between two groups. This occurs when the liaison individuals know they are chokepoints for the goods and information passing through the network and practice good operational security, to continue being effective in their roles.

Each run of the program analyzes a total of 500 randomly reduced networks, in twenty samples of size twenty-five networks. Twenty samples were drawn to have enough data to ensure the distribution of the $\widehat{p}$ (the percent of reduced networks with the same 2-plexes as the complete net-

work) is approximately normal. This is confirmed through a *normal probability plot*, a graphical method for determining whether the data follows a normal distribution. If the data points approximate a straight line on the normal probability plot, it is reasonable to use the normal distribution (Montgomery, 2001: 110).

It is necessary to choose an appropriate sample size to ensure 95% confidence intervals. The information for finding the sample size is in Wackerly, Mendenhall, and Scheaffer's text, *Mathematical Statistics with Applications,* in Section 8.7, starting on page 395. Since each of the distributions of $\widehat{p}$ are normal (shown in section 5.2), approximately 95% of the data will fall within two standard deviations of the mean of $\widehat{p}$. This research is interested in having a 95% confidence interval with error of estimation of $\widehat{p}$ less than 5%.

$$2\sigma_{\widehat{p}} = 0.05 \tag{6.1}$$

The estimate of the standard deviation for $\widehat{p}$ is

$$\sigma_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{6.2}$$

Substituting the value for $\sigma_{\widehat{p}}$ into Equation 6.1 yields

$$\begin{aligned} 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &= 0.05 \\ n &= \frac{\widehat{p}(1-\widehat{p})}{(0.1)^2} \end{aligned} \tag{6.3}$$

Since an estimate for $\widehat{p}$ is not yet known, the value $\widehat{p} = 0.5$ is chosen to maximize n.

Substituting $\widehat{p} = 0.5$ in Equation 6.3 yields

$$n = \frac{0.5(1-0.5)}{(0.1)^2} = 25$$

Therefore, the sample size chosen is 25, which will allow 95% confidence intervals with an error of estimation less than 0.05 for $\widehat{p}$, regardless of what value $\widehat{p}$ actually takes on.

The upper and lower confidence intervals (UCI and LCI, respectively) are calculated by the following formula from Wackerly, Mendenhall, and Scheaffer:

$$UCI = \widehat{p} + z_{\alpha/2}\sigma_{\widehat{p}} \tag{6.4}$$

$$LCI = \widehat{p} - z_{\alpha/2}\sigma_{\widehat{p}} \tag{6.5}$$

Using $\alpha = 0.05$ to find $z_{\alpha/2}$ in a normal probability table and substituting Equation 6.2 into the UCI and LCI in Equations 6.4 and 6.5, respectively, the confidence limits become

$$\begin{aligned} UCI &= \widehat{p} + 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \\ LCI &= \widehat{p} - 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \end{aligned}$$

6.3 Analysis

This section reports on the 20 runs of 25 samples for each of the four methods described in section 5.1, starting on page 135. For each of the methods, the percent of missing information increases iteratively by 2.5% until $\widehat{p}$, the proportion of networks producing the appropriate subgroups, is 0. Once normality has been assessed for $\widehat{p}$, then 95% and 90% confidence intervals are placed on each of runs of 500 samples.

6.3.1 Random Arc Removal (RAR)

The first arc removal method examined is one in which every arc in the network is equally likely to be unknown. Table 13 shows the data, as well as some summary statistics for random arc removal. The number of networks with appropriate subgroups is first shown for each of the twenty samples. A $\widehat{p}$, and 95% normal and binomial confidence intervals are then given.

Recall that if the data for $\widehat{p}$ falls along a straight line on a normal probability plot, then the distribution of $\widehat{p}$ can be assumed to be approximately normal. This allows confidence intervals of $\widehat{p}$ to be calculated using standard normal values.

The normal probability plots in Figures 39 and 40 show the counts of networks with appropriate subgroups is approximately normal. Therefore, the distribution of $\widehat{p}$ is approximately normal, and

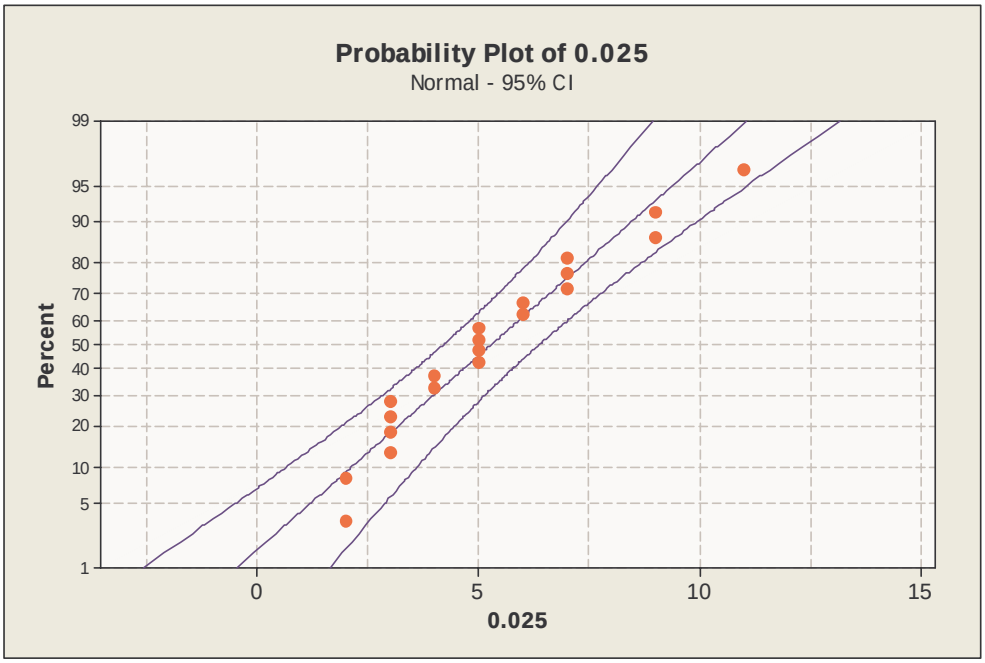

Figure 39. **Normal Probability Plot for 0.025 fraction of arcs missing for RAR**

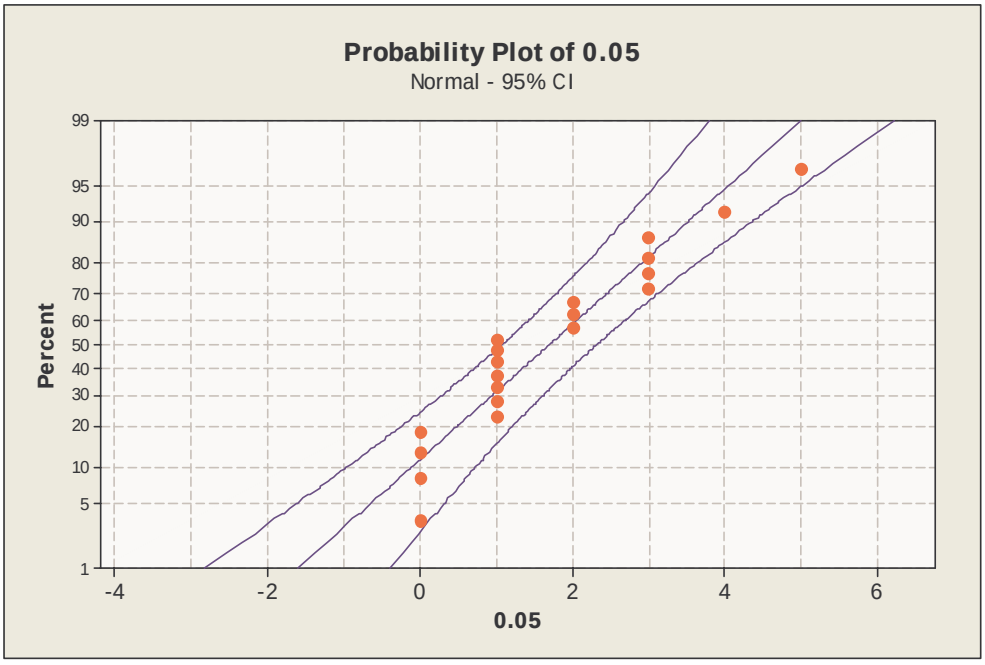

Figure 40. **Normal Probability Plot for 0.050 fraction of arcs missing for RAR**

Table 13. **RAR: Data and statistics for arcs missing at random**

fraction missing	0.025	0.05	0.075	0.10
1	4	3	0	0
2	11	1	0	0
3	7	1	0	0
4	3	0	1	0
5	6	0	1	0
6	2	1	0	0
7	7	4	0	0
8	3	1	0	0
9	5	0	0	0
10	6	1	0	0
11	3	5	1	0
12	4	2	0	0
13	2	3	0	0
14	5	2	0	0
15	5	0	1	0
16	5	3	1	0
17	3	2	0	0
18	9	1	1	0
19	9	1	0	0
20	7	3	0	0
$\widehat{p}$	0.212	0.068	0.012	0
95% Normal LCI	0.174	0.045	0.002	0
95% Normal UCI	0.250	0.091	0.022	0
95% Binomial LCI	0.177	0.048	0.004	0
95% Binomial UCI	0.251	0.094	0.026	0

the normal distribution can be used to derive the confidence intervals. Although the use of normal confidence intervals are common in practice in the application of the central limit theorem, the true confidence intervals in this example are calculated from the binomial distributions. Both confidence intervals are found in Table 13. The normal 95% confidence intervals are symmetric about $\widehat{p}$, while the binomial confidence intervals are not.

Figure 41 shows the point estimate, $\widehat{p}$ with its 95% binomial confidence intervals. Figure 42 shows the shows the same trend 95% normal confidence intervals. The trend line, showing the decrease in $\widehat{p}$ is the same in both plots; the only difference is the confidence interval lines on the points at 0.025, 0.05 and 0.075. The plots are presented separately, because if the two confidence intervals were located on the same plot, they would be nearly indistinguishable. This plot shows how quickly $\widehat{p}$ goes to 0, underscoring the extent to which randomly missing information impedes

subgroup detection. Recall that $\widehat{p} = 0$ when none of the 500 runs produces the correct 2-plexes. Thus when $\widehat{p} = 0$, statistically, the correct subgroups will not be determined.

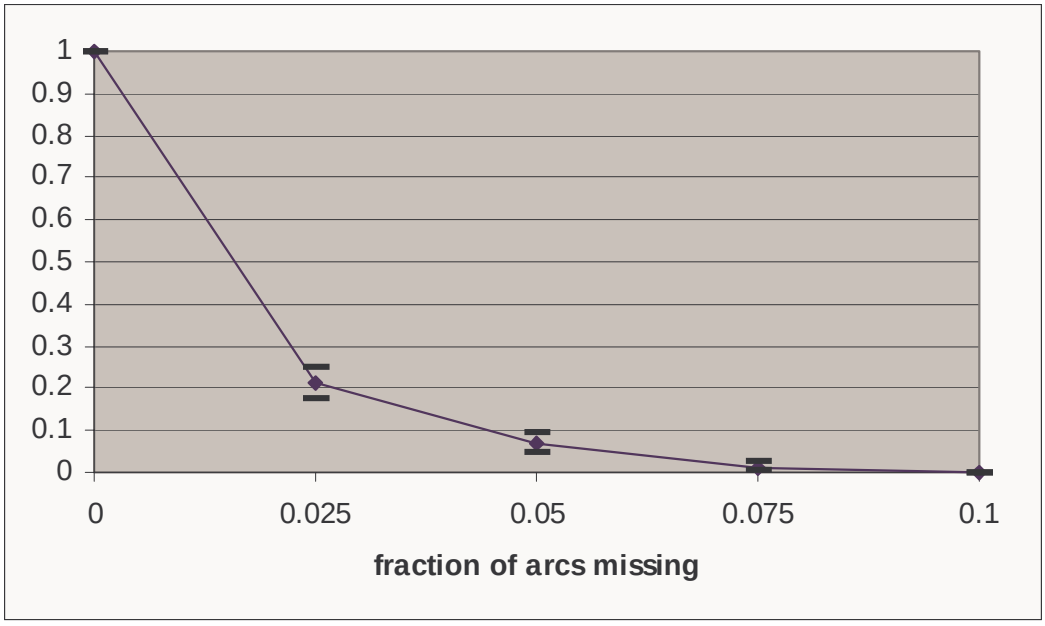

Figure 41. $\widehat{p}$ **and 95% Binomial CI by fraction of missing arcs for RAR**

6.3.2 Random arc removal between nodes of low degree (RARLD)

The second arc removal method examined is one in which only those arcs which connect nodes of less than average degree. Table 14 shows the data as well as some summary statistics for random arc removal. First the number of networks with appropriate subgroups is shown for each of the twenty samples. A $\widehat{p}$, and 95% normal and binomial confidence intervals are given for each level of fraction of arcs missing.

Figure 43 shows the point estimate, $\widehat{p}$ with its 95% binomial confidence intervals. Since the confidence intervals calculated using the normal and binomial distributions are nearly identical, it is not necessary to show a plot of both. Removing arcs between individuals of low degree breaks the subgroups apart very quickly. Removal of just 7.5% of the arcs already yields $\widehat{p} = 0$. Since the arcs are removed only between individuals of low degree, these individuals quickly become either completely disconnected from the network or adjacent to a very small number of individuals. The

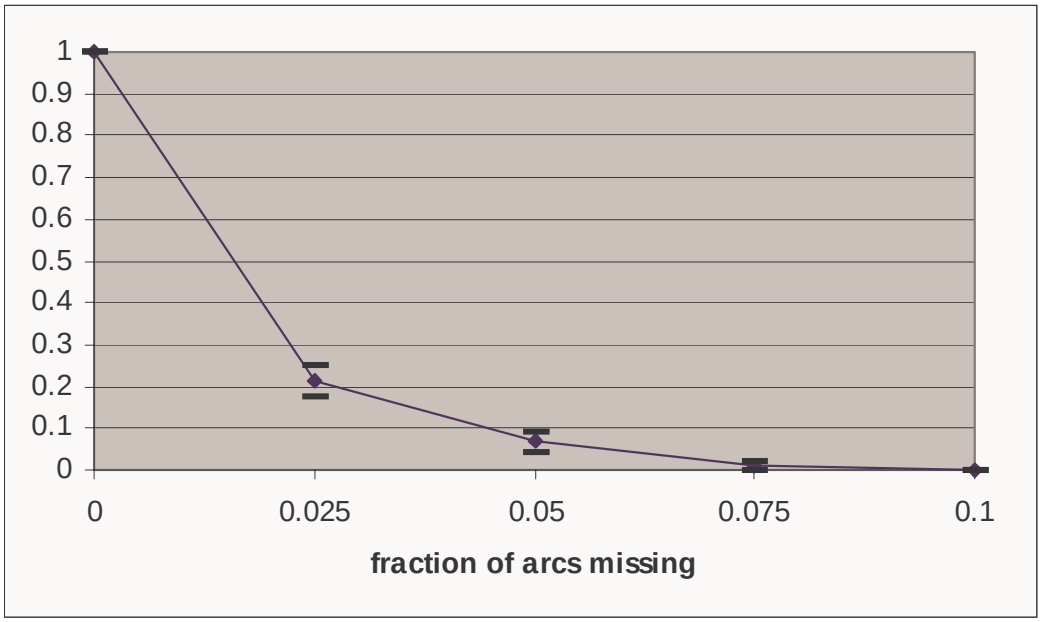

Figure 42. $\widehat{p}$ **and 95% Normal CI by fraction of missing arcs for RAR**

Table 14. **RARLD: Data and statistics for arcs missing at random between nodes of low degree**

fraction missing	0.025	0.05	0.075
1	5	1	0
2	0	1	0
3	5	0	0
4	4	1	0
5	3	1	0
6	2	2	0
7	3	1	0
8	2	0	0
9	4	1	0
10	5	1	0
11	1	0	0
12	5	0	0
13	3	0	0
14	4	3	0
15	4	0	0
16	2	1	0
17	4	1	0
18	2	1	0
19	0	2	0
20	2	2	0
$\widehat{p}$	0.120	0.038	0
95% Normal LCI	0.090	0.020	0
95% Normal UCI	0.150	0.056	0
95% Binomial LCI	0.093	0.023	0
95% Binomial UCI	0.152	0.059	0

2-plex subgroups require each individual to be adjacent to at least $n-2$ other members of the n individuals in the subgroup. When the degree of individuals of low degree starts decreasing even further, these individuals fall out of the subgroups. Although the 2-plexes may still contain many of the same members, the requirement of this experiment is that the subgroups be exactly the same in the reduced adjacency matrix.

6.3.3 Random arc removal incident to nodes of high degree (RARHD)

The third arc removal method examined is one in which only those arcs incident to nodes of higher than average degree are targeted for removal. Even if these individuals are being searched for, they may be able to keep some of their communications hidden through good OPSEC. Table 15 shows the data as well as some summary statistics for random arc removal. First the number

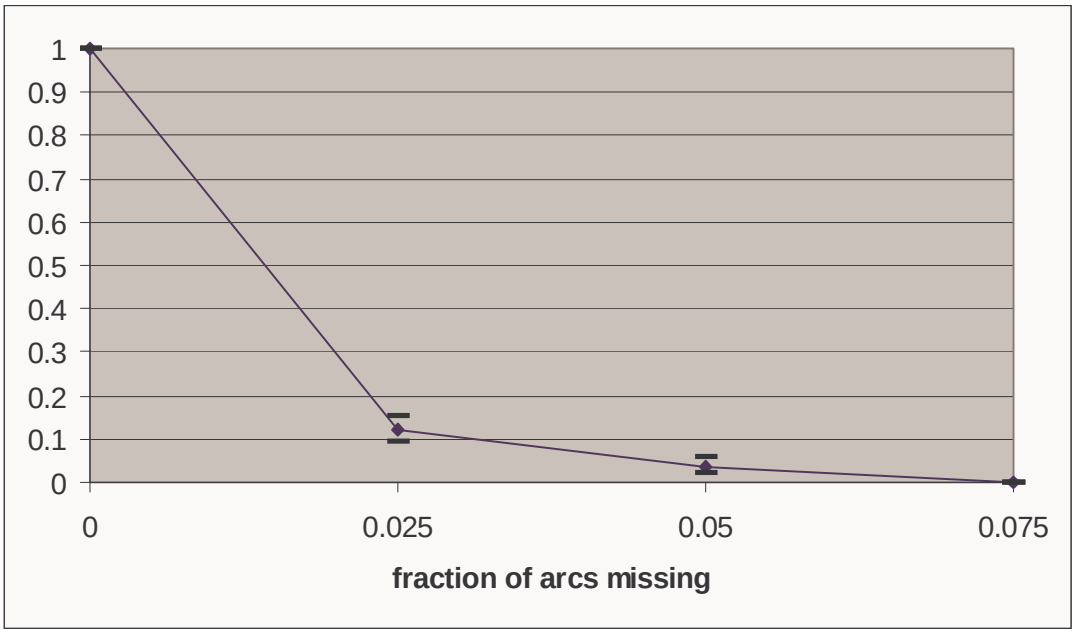

Figure 43. $\widehat{p}$ **and 95% Binomial CI by fraction of missing arcs for RARLD**

of networks with appropriate subgroups is shown for each of the twenty samples. A $\widehat{p}$, and 95% normal and binomial confidence intervals are then given for each level of fraction of arcs missing.

Table 15. **RARHD: Data and statistics for arcs missing at random incident to nodes of high degree**

fraction missing	0.025	0.05	0.075	0.10	0.125
1	6	2	1	0	0
2	6	5	1	0	0
3	12	2	0	0	0
4	13	2	1	0	0
5	9	4	0	0	0
6	6	3	3	0	0
7	9	6	1	0	0
8	7	5	0	0	0
9	6	5	0	0	0
10	6	1	0	0	0
11	6	3	0	0	0
12	9	0	1	0	0
13	6	5	0	1	0
14	5	3	1	0	0
15	6	4	1	1	0
16	11	1	2	1	0
17	4	3	0	2	0
18	9	3	0	0	0
19	7	2	0	0	0
20	7	0	0	0	0
$\widehat{p}$	0.300	0.118	0.024	0.008	0
95% Normal LCI	0.258	0.088	0.010	-0.000	0
95% Normal UCI	0.342	0.148	0.038	0.016	0
95% Binomial LCI	0.260	0.091	0.013	0.002	0
95% Binomial UCI	0.342	0.150	0.042	0.020	0

Removing arcs in this method is less sensitive than previous methods; $\widehat{p}$ does not fall to 0 until 12.5% of the arcs are deleted. Unlike removing arcs between individuals of low degree, removing some of the arcs incident to individuals of high degree does not do as much structural damage to the network. Individuals of higher than average degree often have more relationships with other individuals in the subgroup than necessary in order to form the 2-plexes. It is worth noting the lower 95% normal lower confidence limit for 10% missing information is shown as -0.00. The number is rounded up from -0.0023 and displayed as negative to contrast with the binary lower confidence

limit which never falls below 0. Although the normal confidence interval is an approximation, the fact that it includes 0 means at the $\alpha = 0.05$ level, $\widehat{p}$ is not statistically different from 0.

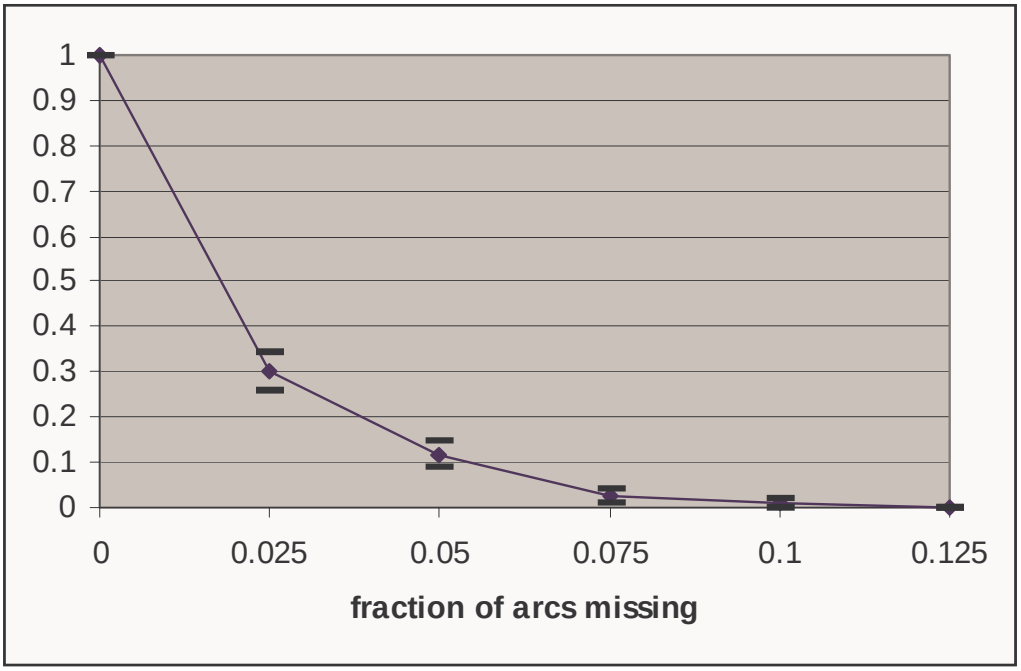

Figure 44. $\widehat{p}$ **and 95% Binomial CI by fraction of missing arcs for RARHD**

Figure 44 shows $\widehat{p}$ and its 95% binomial confidence intervals as a function of the fraction of missing arcs. Since the confidence limits are calculated using the binomial distribution, they are not symmetric, though on a plot of this detail, they are not noticeably different from the normal confidence limits, which are symmetric. $\widehat{p}$ still drops dramatically after just 2.5% of the arcs missing, though not as far as seen in Figures 41 and 43, the similar plots for Methods 1 and 2, respectively.

6.3.4 Random arc removal incident to liaison individuals (RARLI)

The fourth arc removal method examined is one in which only those arcs incident to nodes in multiple subgroups are targeted for removal. Recall in the network under consideration (see Figure 38 on page 134), these liaison individuals are represented by nodes 4, 5, 11, and 15. A lack of information on these arcs may occur when individuals who know they hold two subgroups together

practice good operational security, for example. Even if these individuals are being targeted, they may be able to keep some of their communications hidden.

Table 16 shows the data as well as some summary statistics for random arc removal. First the number of networks with appropriate subgroups is shown for each of the twenty samples. A $\widehat{p}$, and 95% normal and binomial confidence intervals are then given for each level of fraction of arcs missing.

It can be seen that determination of the 2-plexes is more robust under this method of removing arcs than any of the previous three methods. Even removal of 50% of the arcs produces a $\widehat{p} > 0$, though barely. This is due to the fact that these liaison individuals have high degree in the network. Removal of their incident arcs still leaves them well connected with the subgroups of which they are members.

In Table 16, the fraction of missing arcs is shown only in increases of 0.05, not 0.025, as in the previous three methods. This change in fraction missing is due to the number of level of fraction missing required to test using this arc removal method.

The plot with the 95% binomial confidence intervals in Figure 45 shows the 0.025 fraction iterative increase up to 0.50. It is interesting to note first that the value for $\widehat{p}$ is not monotonically decreasing after 0.35. The value of $\widehat{p}_{0.375} > \widehat{p}_{0.350}$ and $\widehat{p}_{0.425} > \widehat{p}_{0.40}$. However, the confidence interval for $\widehat{p}_{0.375}$ includes $\widehat{p}_{0.350}$ (and vice versa); similarly, the confidence interval for $\widehat{p}_{0.425}$ includes $\widehat{p}_{0.40}$(and vice versa). Therefore, this slight increase in the point estimator is not a concern. Also of interest is the sharp decrease in $\widehat{p}$ between 0.125 and 0.150.

6.4 Summary

This chapter has provided an introduction for understanding the impact of missing arcs on the ability to determine appropriate subgroups in the network. Four arc removal methods were shown to target specific random arcs: 1. any arc, 2. any arc connecting nodes of lower than average degree, 3. any arc incident to a node of higher than average degree, and 4. any arc incident to nodes in multiple subgroups (in the complete network). Then each of the four methods were analyzed individually as a percentage of the arcs were removed. The percentage of arcs removed increased

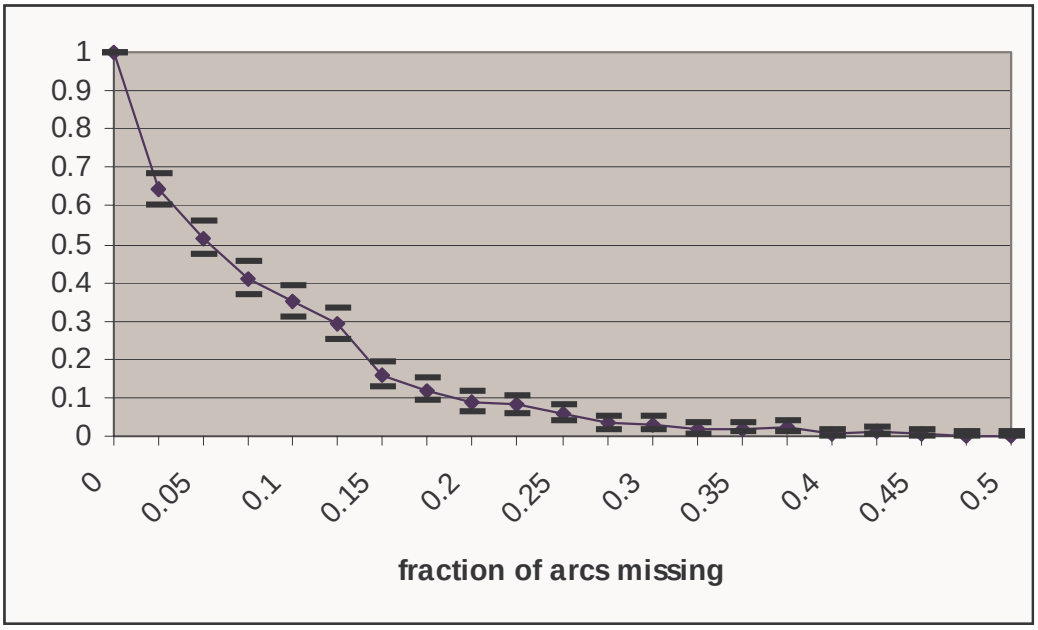

Figure 45. $\widehat{p}$ **and 95% Binomial CI by fraction of missing arcs for RARLI**

Table 16. **RARLI: Data and statistics for arcs missing at random incident to nodes of high degree**

fraction missing	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	12	9	3	2	0	1	0	0	0	0
2	12	10	7	3	2	2	1	1	0	0
3	13	8	1	3	3	0	0	0	0	0
4	15	10	3	2	2	1	0	1	0	1
5	17	7	6	2	2	1	1	0	0	0
6	14	11	4	4	0	1	2	0	0	0
7	11	5	5	0	0	0	0	0	0	0
8	7	10	6	3	1	3	0	0	0	0
9	12	9	5	2	0	1	0	0	0	0
10	16	10	3	3	3	1	0	0	1	0
11	12	11	5	0	1	1	0	1	0	0
12	12	5	5	2	3	1	0	0	1	0
13	14	7	8	1	1	1	1	0	0	0
14	10	8	2	2	2	0	0	0	0	0
15	12	3	2	3	1	0	2	0	0	0
16	15	13	2	5	4	0	1	0	0	0
17	13	6	4	2	2	0	1	0	1	0
18	15	10	3	1	2	1	1	0	0	0
19	13	11	4	2	0	1	0	0	0	0
20	13	12	2	2	0	0	0	0	0	0
$\widehat{p}$	0.516	0.350	0.160	0.088	0.058	0.032	0.020	0.006	0.006	0.002
95% Nml LCI	0.470	0.306	0.126	0.062	0.036	0.016	0.007	-0.001	-0.001	-0.002
95% Nml UCI	0.562	0.394	0.194	0.114	0.080	0.048	0.033	0.013	0.013	0.006
95% Bin LCI	0.471	0.308	0.128	0.065	0.039	0.018	0.010	0.001	0.001	0.001
95% Bin UCI	0.561	0.394	0.195	0.116	0.082	0.052	0.036	0.017	0.017	0.011

incrementally by 2.5% until the point estimate for proportion of networks producing the appropriate 2-plexes was reduced to 0. Each of the runs of 500 networks determines a point estimate, $\widehat{p}$, as well as 95% binomial and normal confidence intervals.

It was found that detection of the appropriate 2-plexes in the network examined in Section 5.2 was most robust under the fourth method of arc removal. $\widehat{p}$ goes to 0 very quickly for the first three methods (0.10, 0.075, and 0.125, respectively), but is still greater than 0 at 0.50 for the fourth method. This is because the liaison individuals in the network examined have fairly high degree, and removal of their incident arcs still leaves them well connected with their subgroups.

This chapter provides only an indication of what an analyst must consider when faced with a network of imperfect information. The methods detailed in Chapter 3 and demonstrated in Chapters 4 and 5 of this thesis rely on being able to find appropriate subgroups for aggregation. While the work

in missing information in this chapter provides insight about how quickly the subgroup detection methods break down, it does not provide the analyst with information on how the outcome of the aggregation techniques are affected. The actual amount of impact missing arcs has on subgroup detection is scenario dependent.

Chapter 7 - Contributions, Limitations, and Future Research

7.1 Contributions

The main contribution of this thesis is in using aggregation to gain insight on structure and activity in social networks. First, application of the subgroup detection methods provides information on size, density, and structure of the subgroups. Whether the subgroups are distinct or overlapping shows how integrated the subgroups are. These subgroups are not required to be pure cliques, which are probably rare in practice. k-plexes were chosen as an appropriate relaxation of cliques, while still allowing cohesiveness within a subgroup. Since these techniques allow the network structure to have ovelapping cliques or distinct cliques, a wide variety of networks can be analyzed.

The node measures chosen to utilize the arc measures are degree and closeness centrality. These are traditional node measures, used often in physical systems, but also applicable in social network analysis as well. These two were chosen first for their ease of use and second for the fact that they consider local and global properties, to provide a wider variety of information about the network. Unless the available network data discounts the use of one of the measures, it is recommended to use both in order to gain more information about the social network properties.

The combination of two subgroup types and node measures provides four areas for calculations. Further, each of those four combinations, two sequences of analysis were detailed, used to examine more subtle interactions within and between subgroups and individuals.

All of the aggregation in this thesis was performed in two sequences: 1. assign to each individual a centrality node measure and then aggregate individuals into subgroups, and 2. aggregate individuals into subgroups and then assign to each subgroup a centrality node measure. The first gives an indication of network activity both within and between subgroups. The second, however, shows only interactions between subgroups, offering more subtle indication of intergroup activity that can be lost in the first aggregation. Therefore, it is recommended that analysis of aggregation be performed in both sequences.

For each of these eight combinations (two subgroup structures, two node measures, and two orders of aggregation analysis), the calculations for the aggregation step are defined, along with an interpretation of the measures. Robustness of aggregation order is also proven. All of these techniques can be used on any additive measure, such as distance, time, speed, cost, length, etc. This allows for analysis of social networks under many different conditions of structure and data definition.

The exploration of incomplete information offers an introduction to how missing arc information inhibits an analysts ability to detect appropriate subgroups. Since the aggregation methodology rests on identification of these subgroups, the impact of detecting them incorrectly is a start of imperfect information analysis.

7.2 Limitations

In general, this thesis does not offer a complete method for analysis of social networks; rather it focuses on using aggregation for gaining insight into the network. Information on network connectivity, cutsets, and susceptibility to attack, among others, are also necessary when performing social network analysis.

One of the three properties of social networks was ignored during this research: high local clustering. Though apparent in many social networks, it is not necessarily appropriate for many groups of interest to the intelligence community, since these groups practice OPSEC by not allowing individuals to become too close. Similarly, to keep one or a small number of subgroups unaware of activities in the rest of the subgroup, the diameter can be larger than that typically found in social networks. While the techniques in this thesis can be used on any social network, they were demonstrated only on networks that are assumed to accurately model these groups of interest.

7.3 Future Research

The chapter on incomplete information offers only a place to begin analysis of such networks. The next step, if following the research in this thesis, is to determine quantitatively how poor identification of subgroups affects the two node measures and interpretation of network activity.

The only piece of missing information examined is relationships. Unknown nodes or poor weighting of arc measures can also be explored for their impact on the node measures.

The subgroup detection methods determine subgroups based solely on network structure. The strength of relationship between individuals can also be used to determine appropriately cohesive subgroups. The White and Harary paper examines cohesion and adhesion in social networks. The two researchers provide a foundation for exploring how to fracture a network, and how the network will split once a fracture occurs.

The research in this thesis considered only undirected networks of deterministic weight. Yan's paper (further detailed in Appendix A) offers methods to address both of these limitations. First, he defines fuzzy cliques which relaxes the pure clique concept to find an appropriate structure in directed networks. Then, he introduces the concept of using fuzzy numbers, as intervals or distributions, as arc weights. This may be appropriate when some strength of relationship is known, but not definitively. It is beneficial to use whatever information is known about a network, and not be reduced to using unity weights if possible.

The two node measures were chosen to display one local and one global property of networks. Many other node measures and network performance measures can be used to gain insight into network activity. These can be further explored using aggregation to show subgroup interactions under the measure chosen.

The closeness centrality node measure considers how quickly information and materiel can be disseminated through the network. Percolation theory offers a rigorous mathematical basis on which this idea can be expanded.

7.4 Conclusion

The research in this thesis can be used by an analyst as an aid to understand levels of activity in a social network. Though it does not require the analyst to have any information other than the hypothesized structure of the network, any outside information the analyst has on individual or subgroup relationships can be used to obtain a more robust model of the network.

APPENDIX A - Clique Detection on Directed Networks

This appendix offers an example for implementation of Yan's fuzzy clique detection algorithm, introduced in the Fuzzy Cliques section of Chapter 2. This method is appropriate for directed networks, since traditional cliques definitions are appropriate only for undirected networks. This example demonstrates not only how to find fuzzy cliques, but also calculates several measures of interest on the arcs and nodes of the network.

Recall the definition (Yan, 1988: 378):

> "A *fuzzy clique* is defined as a maximum strongly connected node subgroup in which each node is connected to all the others directly or indirectly, regardless of the number of intermediate nodes. The core members are those nodes whose distances to and from all clique members are less than or equal to a given fuzzy or non-fuzzy number D."

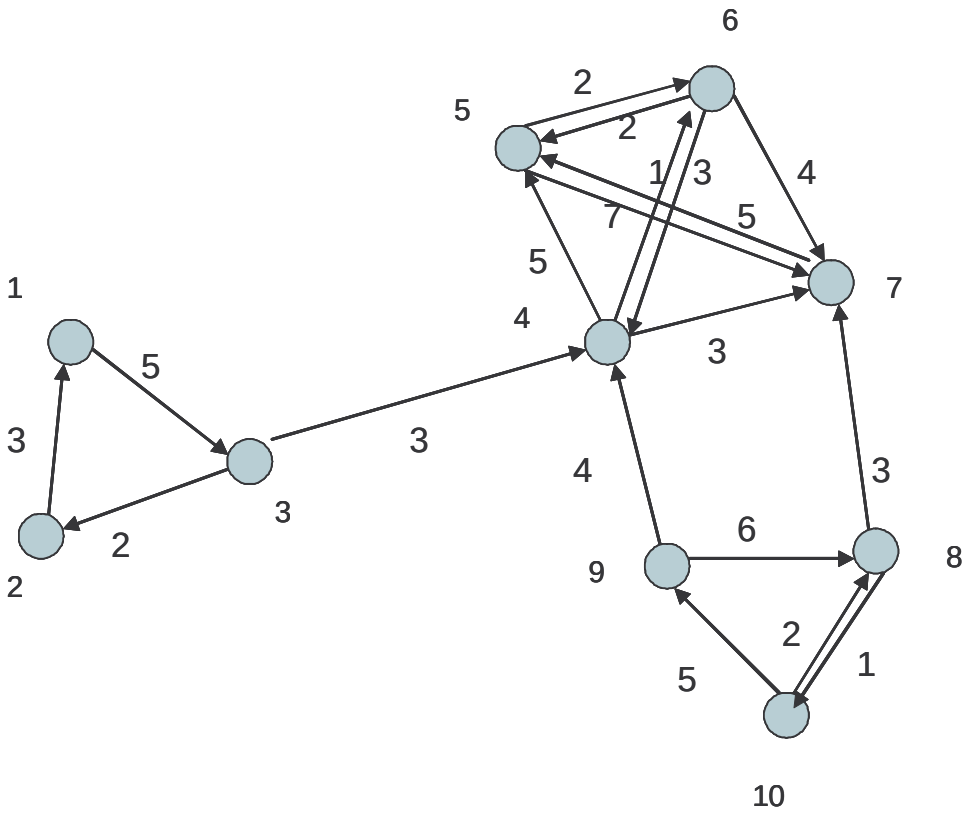

Figure 46. **Disaggregated notional directed network**

The notional network in Figure 46 will be used to demonstrate the process of finding fuzzy cliques as well as assigning several measures Yan has also defined. The adjacency matrix and matrix of arc weights for Figure 46 are as follows:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 1 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 1 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 5 & 0 \end{bmatrix}$$

Following Yan's clique detection algorithm, the network can be partitioned into maximal connected subgraphs who union is the entire network. For this network, the following three sets of nodes are each maximal connected subgraphs, and together they make up the entire network.

$$\begin{aligned} C_1 &= \{1, 2, 3\} \\ C_2 &= \{4, 5, 6, 7\} \\ C_3 &= \{8, 9, 10\} \end{aligned}$$

Several measures can now be defined on the network:

1. **membership value of each node:** a measure of a node's value within its own clique
2. **node-clique coefficient:** a measure of a node's relationship to a clique of which it is *not* a member
3. **clique-clique coefficient:** a measure of relationship between two (possibly aggregated) cliques

A.1 Membership Value of Each Node

Yan defines the membership value to be P'/P where P' is the number of nodes in the clique whose distance from the member is less than D and P is the number of members in the clique (Yan, 1988: 378). To accomplish this, some threshold number D must be chosen. For each clique, this example chooses D to be the average of the edge weights wholly within the clique.

Let $D_{C_i} = \sum_{(i,j) \in A_{C_i}} a_{i,j} / |A_{C_i}|$ be the threshold for clique $C_i \qquad i = 1, 2, 3 \qquad A_{C_i}$ is the set of edges wholly contained within C_i

$$
\begin{aligned}
D_{C_1} &= (3+2+5)/3 = \frac{10}{3} \\
D_{C_2} &= (5+3+7+1+3+5+4+2+2)/9 = \frac{32}{9} \\
D_{C_3} &= (5+6+2+1)/4 = \frac{14}{4}
\end{aligned}
$$

Let M_i be the membership value for node i

i	M_i
1	0
2	$\frac{1}{3}$
3	$\frac{1}{3}$
4	$\frac{3}{4}$
5	$\frac{1}{4}$
6	$\frac{1}{2}$
7	$\frac{1}{2}$
8	$\frac{1}{3}$
9	0
10	$\frac{1}{3}$

A.2 Node-Clique Coefficient

This number provides a measure of how closely related a node is to a clique of which it is not a member.

Yan's measure for this relationship for a clique C and a node N is:

$$K(C,N) = \sum_{\substack{i \in C \\ Q_i \neq 0}} \frac{1}{Q_i} / |C|$$

where Q_i is the directed distance from node N (not in C) to every node $i \in C$ (or vice versa)

Below are a sample of some node-clique connections:

$$
\begin{aligned}
K(C_1,4) &= \frac{\left(\frac{1}{8}+\frac{1}{11}+\frac{1}{3}\right)}{3} = \frac{145}{792} \approx 0.183 \\
K(C_3,4) &= \frac{\left(\frac{1}{4}+\frac{1}{10}+\frac{1}{9}\right)}{3} = \frac{83}{540} \approx 0.154 \\
K(C_3,6) &= \frac{\left(\frac{1}{5}+\frac{1}{7}+\frac{1}{9}\right)}{3} = \frac{143}{945} \approx 0.151 \\
K(C_3,7) &= \frac{\left(\frac{1}{7}+\frac{1}{3}+\frac{1}{5}\right)}{3} = \frac{71}{315} \approx 0.225
\end{aligned}
$$

This implies that the relationship of node 7 with fuzzy clique C_3 is the strongest of the four node-clique pairs chosen. In fact, node 6's relationship with the fuzzy clique C_3 is much weaker than 7s (as evidenced by $K(C_3,6) \approx 0.151$). Note that this measure shows a directed relationship from the fuzzy clique *to* the individual node. Thus, the implication is that C_3 is in some way "closer" to node 7 than node 6. This seems clear when looking at the network.

A.3 Clique-Clique Coefficient

Perhaps most critical when aggregation is important, since this measure offers a relationship between cliques. This can be used in an aggregated network as an edge weight between cliques.

Yan's clique-clique measure is

$$
J(C_1,C_2) = \sum_{i \in C_1} \sum_{\substack{j \in C_2 \\ Q_{i,j} \neq 0}} \frac{1}{Q_{i,j}} \diagup |C_1||C_2|
$$

where $Q_{i,j}$ is the directed distance from node $i \in C_1$ to $j \in C_2$ (or vice versa)

The clique-clique measures for this example network are as follows:

$$
\begin{aligned}
J(C_1,C_2) &= \left(\frac{1}{8}+\frac{1}{11}+\frac{1}{9}+\frac{1}{11}+\frac{1}{11}+\frac{1}{14}+\frac{1}{12}+\frac{1}{14}+\frac{1}{3}+\frac{1}{6}+\frac{1}{4}+\frac{1}{6}\right) \diagup 12 = \frac{9157}{66528} \approx 0.138 \\
J(C_1,C_3) &= 0 \\
J(C_2,C_3) &= \left(\frac{1}{7}+\frac{1}{4}+\frac{1}{5}+\frac{1}{7}+\frac{1}{5}+\frac{1}{9}+\frac{1}{10}+\frac{1}{7}+\frac{1}{3}+\frac{1}{8}+\frac{1}{10}+\frac{1}{10}\right) \diagup 12 = \frac{4909}{30240} \approx 0.162
\end{aligned}
$$

Again, this measure shows the directed membership from one fuzzy clique to another, since the relationship can only go one way between fuzzy cliques. $J(C_1, C_3) = 0$, since it is impossible for anything to flow from either C_1 to C_3 or C_3 to C_1. In this notional example, the relationships flow from C_1 and C_3 to C_2. C_3 has the stronger relationship with C_2 than C_1 does. If, for example, the arc weights represent influence, then the clique-clique measures as defined indicate that C_3 has a greater ability to influence C_2 than C_1 does.

The aggregated network based on these measures is shown in Figure 47.

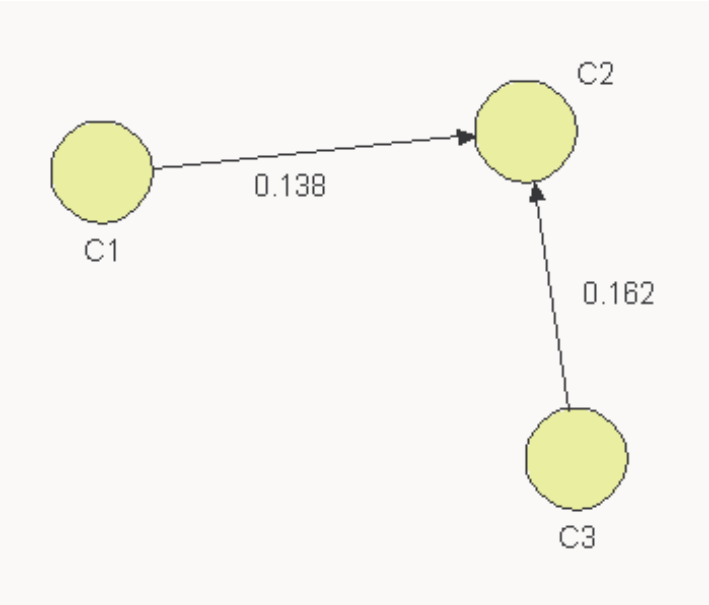

Figure 47. **Aggregated directed network with clique-clique measures**

APPENDIX B - Building Social Networks

None of the techniques described in this thesis can be used without a social network for analysis. Chapter 4 demonstrates the techniques on notional social networks of complete information using a method developed in this section, which provides a method for building social networks appropriate for modelling terrorist social networks. In general, the three properties required for social networks are:

1. small average diameter
2. high local clustering
3. the node degree distribution follows the power law

As previously discussed, the second property is not necessarily appropriate for hostile networks, since operational security measures often keep individuals apart who would otherwise exhibit local clustering behavior.

The method for building these networks begins by determining a node degree distribution for the network. The power law distribution is an inverse log-linear relationship between the degree and number of nodes of that degree. The total number of nodes, $|V|$, the slope of that linear relationship, m, and the degree of the node with the highest degree (the y intercept), b, are user inputs. The nodes are numbered 1 through $|V|$, and each node, i, is assigned a degree equal to $\lceil e^{-m\ln(i)+b} \rceil$. This produces a monotonically decreasing degree distribution, where node 1 has the highest degree, b, and node $|V|$ has the lowest degree $\lceil e^{-m\ln(|V|)+b} \rceil$.

B.1 Method for generating social networks

Let $|V|$ be the number of nodes, m be the slope of the linear relationship (this determines the network density), b be the greatest node degree in the network (degree of node 1)

for $i = 1 : |V|$

$degree(i) = \lceil e^{-m\ln(i)+b} \rceil$

for $j = i + 1 : |V|$

if i and j do not yet have their full degree, then $A(i, j) = A(j, i) = 1$

end

end

B.2 How to vary m and b and what they represent:

It is first necessary to understand the function which determines the degree distribution. As previously described, the power law used here states that the natural log of the number of nodes of a given degree is inversely linearly proportional to the natural log of the degree of the network. This linear relationship has a negative slope, since more people in the network have lower degree, while a small number of people have a high degree. The function is $\ln(\deg(i)) = -m\ln(i) + b$ where $\deg(i)$ is the degree of node i, and (i) is the index of the node in the network.

The multiplier, m, is the slope of the function. Keeping b constant, as m increases the slope of the function becomes steeper. This means that a relatively large m defines a degree distribution in which an increasingly small number of nodes have higher degree, and most nodes have small degree. Even when b is quite large, $m = 3$ yields only two nodes of degree greater than 1. When the degree distribution shows many nodes of degree 1, this tends to produce pairs of nodes connected to each other, but disconnected from the rest of the network. As m gets smaller, the slope of the function becomes shallower. Thus very few nodes have small degree, and there is not much gradation in the degree. Again, even when b is large, $m = 0.8$ has no nodes of degree 1.

The intercept, b, determines the degree of the node of highest degree. This affects the degree of all other nodes, since they must follow the power law function described in Section 3.5. b is chosen to be the natural log of the degree of the node having highest degree. Holding m constant, a large b provides a wider range of degrees for the vertices. Since b controls the highest degree in the network, a small b produces a network in which there is not much variance in the degrees.

B.3 Properties of Social Networks:

Before using these networks as examples of social networks, it is necessary to show that this method does actually produce networks that possess the properties necessary for modeling terrorist

networks. Small average diameter, high local clustering, and power law degree distribution on the nodes are examined in this section.

B.3.1 Small Diameter

Let d_{ij} be the shortest path between vertices i and j. The small diameter property requires $\max_{i,j \in A} d_{ij}$ be small. This will be shown for $m = 1$ and $m > 1$. It is not necessarily true for $m < 1$.

B.3.1.1 Suppose $m = 1$ Recall the degree of a node, i, is equal to $\lceil e^{-m \ln(i)+b} \rceil$. A node has degree 1 if

$$\begin{aligned} e^{-m \ln(i)+b} &< 1 \\ -m \ln(i) + b &< \ln(1) \end{aligned}$$

The first indexed node with degree < 1 is:

$$\begin{aligned} \min_i \{-m \ln(i) + b &< 0\} \\ -m \ln(i) + b &< 0 \\ -m \ln(i) &< -b \\ \ln(i) &> \frac{b}{m} \\ i &> e^{\frac{b}{m}} \end{aligned}$$

The minimum indexed node, i, which satisfies this equation is $i = \lceil e^{\frac{b}{m}} \rceil$. Now consider the node of highest degree. Its degree is

$$\lceil e^{-m \ln(1)+b} \rceil = \lceil e^b \rceil$$

The first node of degree 1 is $\lceil e^b \rceil$, and the degree of node 1 is also $\lceil e^b \rceil$. Since the method makes node j adjacent with the smallest indexed node possible, node 1 is adjacent to nodes 2, 3, $\cdots$, i, $i + 1$, where i is the first node of degree 1. This means every vertex j such that $degree(j) > 1$ is adjacent to node 1, making every pair of nodes with degree greater than 1 only a path of length 2

apart. Nodes of degree 1 are either adjacent to another vertex of degree 1 or a vertex with degree greater than 1.

First consider two nodes of degree 1 adjacent. Then they form a disconnected arc from the remainder of the network. The analysis of the subgroup detection and aggregation techniques developed in this chapter uses only the large component from this network. These disconnected components will be removed from the network. There are, however, situations when this method can be used to build social networks where the disconnected components are appropriate in the network structure. Disconnected components can be an indication of missing information, pointing intelligence analysts to a place to allocate more resources. These smaller components can represent sleeper cells that are unconnected at one point in the analysis. If further analysis shows disconnected components forming relationships with the larger component, or among themselves, then this could omen increased communication and future actions.

Consider now two nodes of degree 1 that are not adjacent, where neither is part of a disconnected component on two nodes. They must be connected to each other through the large component, through nodes of degree greater than 1. Since it has been shown that any pair of nodes of degree greater than 1 are at most a path of length 2 apart, then the nodes of degree 1 must be at most a path of length 4 apart.

B.3.1.2 Suppose $m > 1$ Then the first node of degree 1 is $\lceil e^{\frac{b}{m}} \rceil < \lceil e^{b} \rceil$, since e is a monotonically increasing function. As described in the previous case, every vertex of degree >1 is adjacent to node 1, thus making them at most a path of length 2 apart. Then vertices of degree one adjacent to a vertex of degree >1 is of distance at most 4 away from every other vertex in the large component. When b is small, however, there are many vertices of degree 1 adjacent to each other. This makes for a sparse, disconnected network.

B.3.1.3 Suppose $m < 1$ There is no guarantee that a network built with $m < 1$ as an input parameter will have a small diameter. The choice of b determines the size of the largest component;

a large b ensures the cliques and near cliques produced are connected through a small number of arcs. However, the diameter can become increasingly large, as the clique or near clique containing node 1 can be arbitrarily far from the near clique containing $|V|$. If a small m is desired to obtain a sufficiently dense social network, then it may be necessary to make multiple small networks and connect them together into a united whole.

B.3.2 High local clustering

High local clustering is evident when two nodes, j and k, adjacent to a node i are also adjacent themselves. A network built with small m will exhibit this property, since the variation in degree of the nodes is not large, especially as b decreases. However, in terrorist networks, many members are kept purposefully ignorant of one another for security reasons. Thus, high local clustering is not necessary or even appropriate for terrorist networks.

B.3.3 Power law vertex degree distribution

It has been shown in social networks having perfect information that the degree distribution must follow the power law, as described in the section explaining the impact of varying m and b, Section 3.5.2. This method, by design, gives the networks a vertex degree distribution that follows the power law.

B.4 Demonstration of Method

The two examples shown in this appendix demonstrate different structures the method can generate. The first is denser, connected, and shows two, possibly three interconnected subgroups. The second is not connected and less dense, as many of the nodes have degree 1. This can represent a subgroup where node 1 is a recruiter and nodes 2-7 are people who have recently been brought into the network. Furthermore, the disconnected component, nodes 8 and 9, may indicate missing information or erroneous data.

B.4.1 Example 1

$n = 10$; $b = 7$; $m = 0.35$

This produces a network on 10 nodes, where the node of greatest degree has degree 7, and is fairly dense (indicated by a low m value)

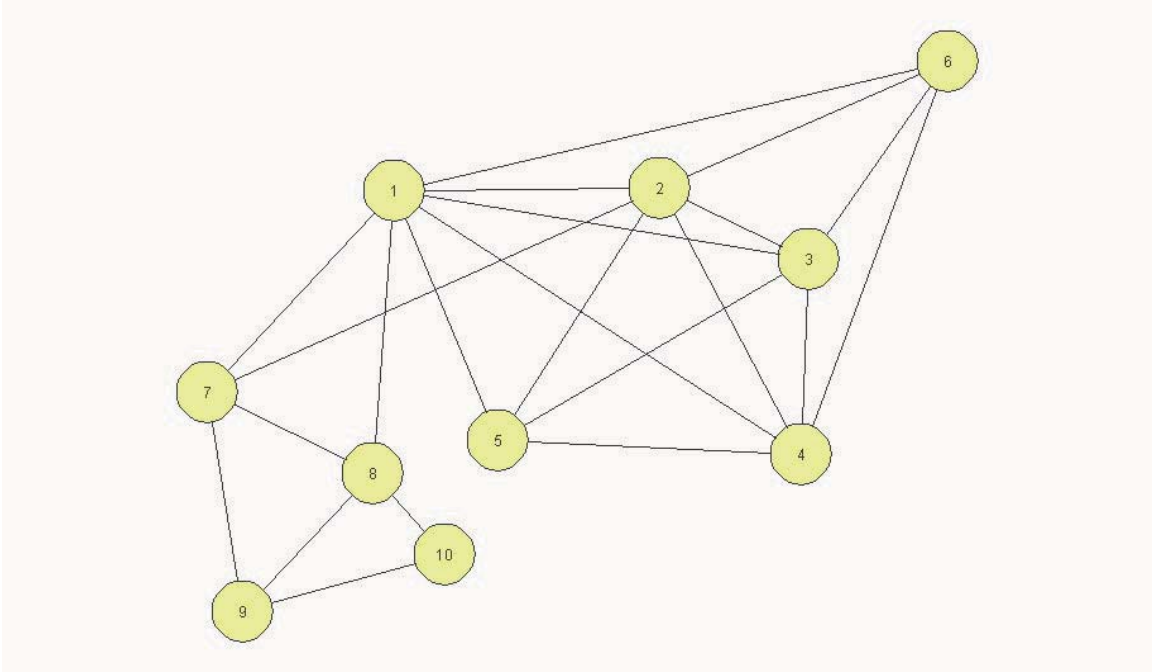

Figure 48. **Example Social Network 1**

B.4.2 Example 2

$n = 10$; $b = 6$; $m = 1$

This produces a network on 10 nodes, in which the node of highest degree has degree 6, and is considerably less dense than the one in Example 1. Note that there is an edge connecting two vertices of degree 1 that are disconnected from the large component. In such an example, it may be appropriate to discard the smaller component, if the analysis is concerned only with connected networks. However, this may be used to represent two subgroups that are unaware of each other, or simulate missing information.

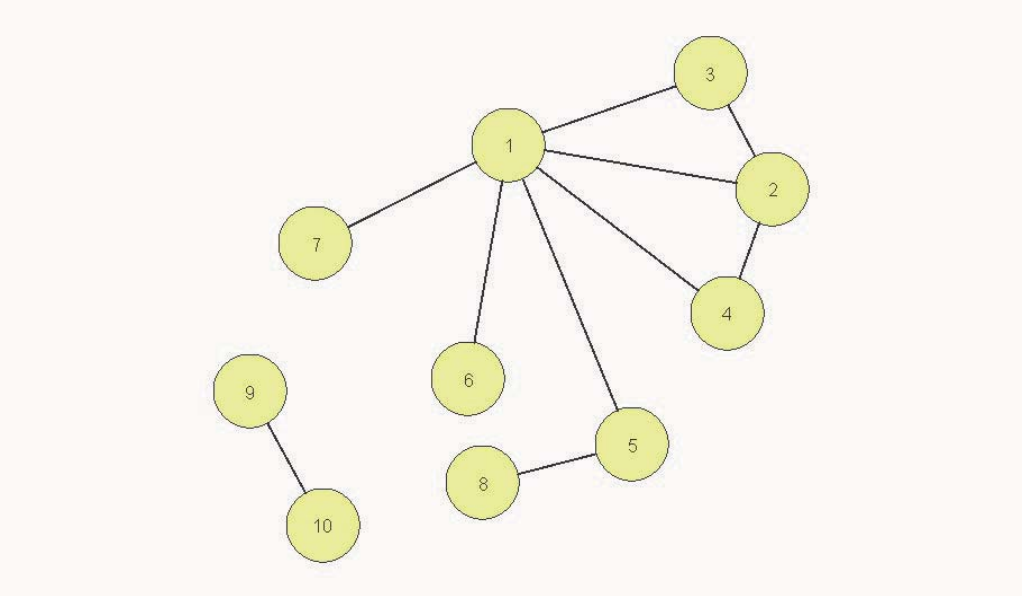

Figure 49. **Example Social Network 2**

APPENDIX C - Jema'ah Islamiyah Names and Sources

C.1 JI names by numbered index

1 Abdullah Sungkar (see Abu Bakar Baasyir)

3 Abdul Wahid Kadungga

4 Abu Bakar Baasyir (or Bashir or Ba'asyir) (see Abdullah Sungkar)

9 Agus

10 Ali Gufron (alias Muchlas) (alias Huda bin Abdul Haq)

11 Ali Imron

12 Amin

13 Amrozi

16 (Dr) Azahari

18 Ending Isomudin

19 Eni Maryani

20 Faiz bin Abu Bakar Bafana

24 Fuad Amsyari

27 Hambali (also called Riduan Isamuddin or Ensep Nurjaman)

36 Khalid Almihdhar

37 Mike

42 Mohammed Iqbal A Rahman (or Abdurrahman) (also called Abu Jibril) (also called Fikiruddin Muqti)

46 Nasir Abbas

47 Nawaf Alhazmi

48 Noordin Mohammed Top (or Thob)

49 Nur Fitrotullah

53 Rauf

54 Samudra (Imam) – Abdul Aziz, Kudama, Abu Omar, Faiz Yunshar

55 Sammy

57 Sumarno

61 Wan Min bin Wan Mat

63 Yazid Sufaat

64 Yudi

65 Zacarias Moussaoui

66 Zulkarnaen al Arif Sunarso al Daud

67 Zulkifli Marzuki

C.2 Sources of Information on JI relationships

- Source: ABC News Online
  http://abc.net.au/news/indepth/featureitems/s737774.htm
- Source: The Age
  http://www.theage.com.au/articles/2003/02/10/1044725712936.html
  http://www.theage.com.au/articles/2003/02/04/1044122333063.html
- Source: Asia Times
  http://www.atimes.com/se-asia/DB06Ae01.html
- Source: Associated Press
  http://abcnews.go.com/wire/World/ap20021013_556.html
- Source: BBC
  http://news.bbc.co.uk/2/hi/asia-pacific/1844871.stm
  http://news.bbc.co.uk/2/hi/asia-pacific/2284645.stm
  http://news.bbc.co.uk/2/hi/asia-pacific/2339693.stm
  http://news.bbc.co.uk/2/hi/asia-pacific/2346225.stm
  http://news.bbc.co.uk/2/hi/asia-pacific/2385323.stm
  http://news.bbc.co.uk/2/hi/asia-pacific/2542863.stm
  http://news.bbc.co.uk/2/hi/asia-pacific/2602747.stm
- Source: Brunei Direct
  http://www.brudirect.com/DailyInfo/News/Archive/Dec02/211202/wn06.htm
- Source: Christian Science Monitor
  http://www.csmonitor.com/2002/0212/p06s02-wosc.html
- Source: Crisis web
  http://www.crisisweb.org/projects/showreport.cfm?reportid=845
- Source: FAS
  http://www.fas.org/irp/world/para/ji.htm
- Source: Feral News
  http://www.feralnews.com/issues/bali/pastika_case_summary_0301.html
- Source: Gulf News
  http://www.gulf-news.com/Articles/news.asp?ArticleID=80591

- Source: Herald Sun Australia
  http://heraldsun.news.com.au/common/story_page/0,5478,5627306%255E401,00.html
- Source: HindustanTimes.com
  http://www.hindustantimes.com/news/181_239371,00050004.htm
- Source: HM Treasury, 24 October 2002
  http://www.britain-info.org/asia/xq/asp/SarticleType.1/Article_ID.2771/qx/articles_show.htm
- Source: ICG
  http://www.crisisweb.org/projects/asia/indonesia/reports/A400845_11122002.pdf
  http://www.asia-pacific-action.org/southeastasia/indonesia/resources/reports/igc%20report%20on%20ji.htmv
  http://www.intl-crisis-group.org/projects/asia/indonesia/reports/A400845_11122002.pdf
- Source: intellnet
  http://www.intellnet.org/news/2003/01/28/15915-1.html
- Source: LA Times
  http://www.latimes.com/news/nationworld/world/la-fg-bali30jan30,0,3649555.story?coll=la%2Dheadlines%2Dworld
- Source: MSNBC
  http://www.msnbc.com/modules/wtc/wtc_globaldragnet/custody_malaysia.htm
- Source: Nanyang Tech University, Singapore
  http://www.ntu.edu.sg/idss/Perspective/research_050221.htm
- Source: NEWS.com.au
  http://news.com.au/common/story_page/0,4057,6287121%255E2,00.html
- Source: Reuters
  http://www.singapore-window.org/sw02/021108re.htm
  http://news.lycosasia.com/SGEN/290,697,4.asp
  http://www.alertnet.org/thenews/newsdesk/JAK314622.htm
- Source: Sydney Morning Herald
  http://www.smh.com.au/articles/2003/01/26/1043533953389.html
  http://www.smh.com.au/articles/2003/02/03/1044122322016.html
- Source: TEMPO Interactive
  http://www.tempo.co.id/news/2003/1/2/1,1,5,uk.html
- Source: U.S. Embassy, Jakarta
  http://www.usembassyjakarta.org/terrorism/2-JIterrorist.html
- http://www.buckyogi.addr.com/footnotes/natgj.htm#indoisl

BIBLIOGRAPHY

[1] Batagelj, V., A. Ferligoj, and P. Doreian. "Generalized Blockmodeling," *Informatica,* 23/4:501-6 (December 1999).

[2] Beauchamp, M.A. "An improved index of centrality," *Behavioral Science,* 10: 161-163 (1965).

[3] Bonacich, P. "Power and Centrality: A family of measures" *American Journal of Sociology,* 92: 1170-1182. (1987).

[4] Bonacich, Phillip, and Maureen J. McConaghy. "The algebra of blockmodeling," *Sociological Methodology*, 489-522:1980.

[5] Breiger, Ronald L. *Explorations in Structural Analysis: Dual and Multiple Networks of Social Interaction.* New York: Garland Publishing, 1991.

[6] Bron, C. and J. Kerbosch. "Finding all cliques in an undirected graph," *Communications of the ACM*, 16:575–577 (1973).

[7] Callaway, Duncan S., M.E.J. Newman, Steven H. Strogatz, and Duncan J. Watts. "Network Robustness and Fragility: Percolation on Random Graphs," *Physical Review Letters*, 85/25: 5468-5471 (18 December 2000).

[8] Chang, Kuo-Chu and R Fung. "Node aggregation for distributed inference in Bayesian networks," *IJCAI-89 Proceedings of the Eleventh International Joint Conference on Artificial Intelligence in Detroit MI, 20-25 Aug 1989* (Vol 1). Palo Alto CA: Morgan Kaufmann: 265-70, 1989.

[9] Comellas, Francesc, Javier Ozon, and Joseph G Peters. "Deterministic small-world communication networks," *Information Processing Letters*, 76/1-2: 83-90 (30 November 2000).

[10] Everett, M.G. "A Graph Theoretic Blocking Procedure for Social Networks," *Social Networks*, 4: 147-167 (1982).

[11] Freeman, L.C. "A set of measures of centrality based on betweenness," *Sociometry*, 40: 35-41 (1977).

[12] Godehardt, Erhard. *Graphs as Structural Models.* (Second Edition) Braunschweig: Viewag and Sohn Verlagsgesellschaft, 1990.

[13] Gomez, Daniel, Enrique Gonzalez-Aranguena, Conrado Manuel, Guillermo Owen, Monica del Pozo and Juan Tejada. "Centrality and power in social neworks: a game theoretic approach," *Mathematical Social Sciences*, 46/1: 27-55 (August 2003).

[14] Hoff, Peter D., Adrian E. Rafferty, and Mark S. Handcock. "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97/460: 1090-1099 (December 2002).

[15] http://www.cnn.com/2002/WORLD/asiapcf/southeast/10/22/seasia.terrorplans/index.html. 10 February 2004.

[16] http://www.cnn.com/2003/WORLD/meast/11/15/turkey.explosion/index.html. 10 February 2004.

[17] http://www.cnn.com/WORLD/africa/9808/07/terror.chronology/index.html. 10 February 2004.

[18] Kumar, Ravi, Prabakar Raghaven, Sridhar Rajagopalan, and Andrew Tomkins. "The Web and Social Networks," *Computer Networks*, 35/11: 32-36 (November 2002).

[19] Lopez, L. and M.A.F. Sanjuan. "Relation between structure and size in social networks," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics),* 65/3: 036107/1-7 (March 2002).

[20] McAndrew, Duncan Ross. *The Structure of Criminal Networks.* Thesis (Ph.D.) University of Liverpool. 1999.

[21] Montgomery, Douglas C. *Introduction to Statistical Quality Control.* New York: John Wiley & Sons, Inc., 2001.

[22] Morris, Peter. *Introduction to Game Theory.* New York: Springer-Verlag, Inc., 1994.

[23] Nagpaul, P.S. "Visualizing cooperation networks of elite institutions in India," *Scientometrics*, 54/2: 213-228 (2002).

[24] Nair, P.S. and S.C. Cheng. "Cliques and fuzzy cliques in fuzzy graphs," *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, 4: 2277-2280 (2001).

[25] Newman, M.E.J. "Mixing patterns in networks," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 67/2: 26126/1-13 (Februaru 2003).

[26] Newman, M.E.J., S. Forrest, and J. Balthrop. "Email networks and the spread of computer viruses," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics),* 66/3: 35101/1-4 (September 2002).

[27] Newman, M.E.J., S.H. Strogatz, and D.J. Watts. "Random graphs with arbitrary degree distributions and their applications," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics),*64/2, 026118/1-17 (August 2001).

[28] Newman, M.E.J., D.J. Watts, and S.H. Strogatz. "Random graph models of social networks," *PNAS*, 99/Suppl1: 2566-2572 (19 February 2002).

[29] Parsons, Simon. "Current Approaches to Handling Imperfect Information in Data and Knowledge Bases," *IEEE Transactions on Knowledge and Data Engineering,* 8/3: 353-372 (June 1996).

[30] Renfro, Robert S. II. *Modeling and Analysis of Social Networks* Thesis (Ph.D.)–Air Force Institute of Technology, 2001.

[31] Robins, Garry, Philippa Pattison, and Peter Elliot. "Network Models for Social Influence Processes," *Psychometrika*, 66/2: 161-190 (June 2001).

[32] Roth, Philip L. "Missing data: a conceptual review for applied psychologists," *Personnel Psychology*, 47/3: 537-562 (Autumn 1994).

[33] Seidman, Stephen B., and Brian L. Foster. "A graph-theoretic generalization of the clique concept," *J. Mathematical Society*, 6: 139-154 (1978).

[34] Shaw, M.E. "Group structure and the behavior of individuals in small groups," *Journal of Psychology,* 38:139-149 (1954).

[35] Sparrow, M.K. "The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects," *Social Networks*, 13: 251-274 (1991).

[36] Van Mieghem, P. "Topology information condensation in hierarchical networks," *Computer Networks*, 31/20: 2115-37 (27 September 1999).

[37] Wackerly, Dennis D., William Mendenhall III, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. (Sixth Edition) Pacific Grove CA: Duxbury, Thompson Learning, 2002.

[38] Wang, Y.J. and G.Y. Wong. "Stochastic Blockmodeling for Directed Graphs," *Journal of the American Statistical Association*, 82: 8-19 (1987).

[39] Watts, Duncan J. *Small worlds: the dynamics of networks between order and randomness*. Princeton NJ: Princeton University Press, 1999.

[40] West, Douglas B. *Introduction to Graph Theory*. (Second Edition) New Jersey: Prentice Hall, 2001.

[41] White, Douglas R., and Frank Harary. "The Cohesiveness of Blocks in Social Networks: Node Connectivity and Conditional Density," *Sociological Methodology*, 31/1: 305-360 (January 2001).

[42] Wilson, Robin J. and John J. Watkins. *Graphs: an introductory approach*. New York: John Wiley & Sons, Inc., 1990.

[43] Yan, Xiaoyan. "On fuzzy cliques in fuzzy networks," *Journal of Mathematical Sociology*, 13/4: 359-389 (1988).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
01-09-2004	**Master's Thesis**	Jun 2004 – Mar 2004

4. TITLE AND SUBTITLE

AGGREGATION TECHNIQUES FOR SOCIAL NETWORK ANALYSIS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Sterling, Sara, E., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 642
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/04-12

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

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Social network analysis focuses on modeling and understanding individuals of interest and their relationships. Aggregation of social networks can be used both to make analysis computationally easier on large networks, and to gain insight in subgroup interactions. Aggregation requires determining appropriate closely knit subgroups as well as choosing a measure or measures to represent the network data.

This thesis provides the analyst with several techniques for using aggregation to analyze the characteristics of social networks. The contribution of this research lies in its ability to analyze a wide variety of social network structures and available data through two methods for subgroup detection and application of two network measures. These techniques are demonstrated on notional social networks. Since analysts rarely have perfect information of the network structure, an exploration of the effects of missing arcs on subgroup detection is presented.

15. SUBJECT TERMS

social network analysis, aggregation, cliques, *k*-plexes, node centrality measures

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**Dr. Richard F. Deckro**
U	U	U	UU	187	**19b. TELEPHONE NUMBER** *(Include area code)* *(937) 255-6565, ext 4325*

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