Detecting Changes in a Dynamic Social Network

Ian McCulloh

March 31, 2009
CMU-ISR-09-104

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Kathleen M. Carley, Chair
Carolyn Rose
Cosma Shalizi (Statistics)
Kevin Huggins (U.S. Military Academy)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Program in Computation, Organization, and Society

Keywords: social network analysis, statistical process control, longitudinal network analysis, change detection, network statistics, network dynamics

Abstract

Social network analysis (SNA) has become an important analytic tool for analyzing terrorist networks, friendly command and control structures, arms trade, biological warfare, the spread of diseases, among other applications. Detecting dynamic changes over time from an SNA perspective, may signal an underlying change within an organization, and may even predict significant events or behaviors. The challenges in detecting network change includes the lack of underlying statistical distributions to quantify significant change, as well as high relational dependence affecting assumptions of independence and normality. Additional challenges involve determining an algorithm that maximizes the probability of detecting change, given a risk level for false alarm.

A suite of computational and statistical approaches for detecting change are identified and compared. The Neyman-Pearson most powerful test of simple hypotheses is extended as a cumulative sum statistical process control chart to detect network change over time. Anomaly detection approaches using exponentially weighted moving average or scan statistics investigate performance under conditions of potential time-series dependence. Fourier analysis and wavelets are applied to a spectral analysis of social networks over time. Parameter values are varied for all approaches. The results are put in a computational decision support framework.

This new approach is demonstrated in multi-agent simulation as well as on eight different real-world data sets. The results indicate that this approach is able to detect change even with high levels of uncertainty inherent in the data. The ability to systematically, statistically, effectively and efficiently detect these changes has the potential to enable the anticipation of change, provide early warning of change, and enable faster appropriate response to change.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

Terrorists from al-Qaeda attacked America on 11 September 2001. Some suggest that these terrorists began to plan and resource this attack as early as 1997. If social network analysts could monitor the social, email, or phone networks of these terrorists and detect organizational changes quickly, they may enable military leaders to respond prior to the successful completion of their attack. Social network change detection (SNCD) is a novel approach to this problem. It combines the area of statistical process control and social network analysis. The combination of these two disciplines is likely to produce significant insight into organizational behavior and social dynamics.

Statistical process control is a statistical approach for detecting anomalies in the behavior of a stochastic process over time. This approach is widely used in manufacturing as a means for quality control. Manufacturing systems experience similar issues of high correlation, dependence, and non-ergodicity that is common in relational network data. I posit that applying statistical process control to graph-level network measures is effective at rapidly detecting changes in longitudinal network data.

It is important to note that I am not predicting change, but rather detecting that a change occurred quickly and making some inference about the actual time of change. For example, before a terrorist commits an attack, there will be a change in the social network as the organization plans and resources the attack. SNCD may allow an analyst to detect the change in the social network, prior to the successful completion of the attack. In a similar fashion, corporate managers may wish to detect changes in the organizational behavior of their companies to capitalize on innovation or prevent problems. For example, the CEO of Dupont became aware of the U.S. recession in late 2008 in time to enact a crisis management plan averting financial disaster for the company. In this example, the economic change had already occurred. Dupont's success was not in predicting a recession, but rather detecting that it had occurred quickly, in time to respond.

SNCD may offer executives and military analysts a tool to operate inside the normal decision cycle. Figure 1 represents some measure of interest over time. It could be the revenue of a company, the combat power of an enemy, or for our purposes a measure of interest from a social network. When do we conclude from this measure that a change may have occurred? Let us assume that by conventional methods we can detect a change in organizational behavior as of "today", the vertical line in Figure 1. This time point might be too late to take preventative or mitigating action. In other words, this could be the point of inevitable bankruptcy for the company, or the successful culmination of a terrorist attack. Identifying that a change occurred by time period E might allow the analyst to respond to the change before it is too late; get inside the decision cycle.

Change detection is more challenging than it may seem at first. We can see a sudden change in the measure between time D and time E, however, this may look very similar to the peak at time A. Furthermore, if we assert that a change in fact occurs at time A,

there may exist a large amount of time periods to investigate for the cause of any change. If we can identify more likely points in time when change may have occurred, we can reduce the costs in terms of time and resources to search for the potential causes of change. Identifying the likely time that a change may have occurred is called *change point identification.*

Another problem that we face is detecting the change as quickly as possible after the change occurred. Can we improve the ability to get inside the decision cycle by detecting the change at time D, or even better at time B? This is called *change detection.* This thesis is a first attempt to investigate this challenging problem in longitudinal network analysis.

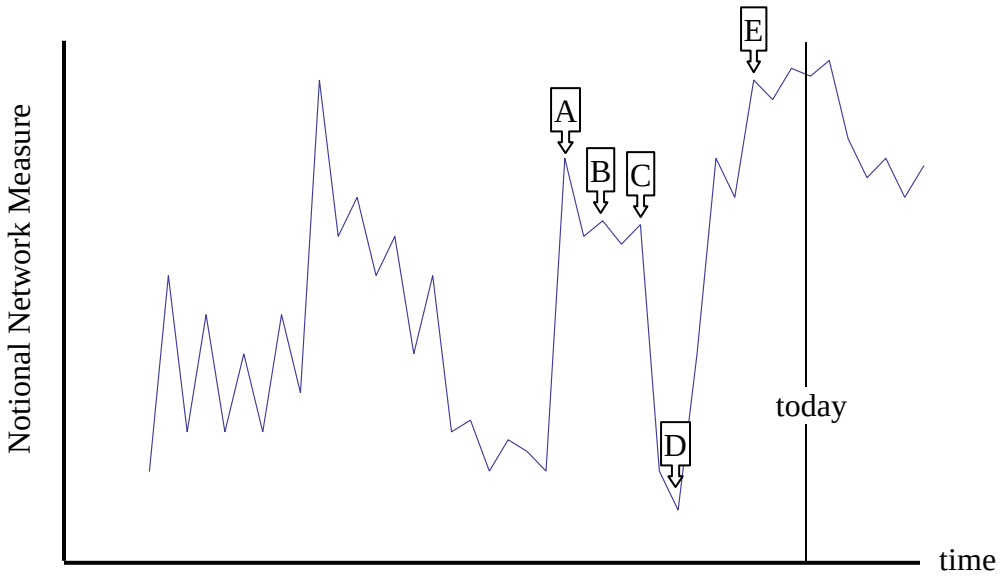


Figure 1. Example of Change Detection

1.1 Importance of Change in Longitudinal Social Networks

This thesis addresses a new area of research that is a national need. Research agencies throughout the Department of Defense (DoD) and the U.S. Government have demonstrated recent interest in pursuing research in the area of social network analysis. Particular interest is in stochastic and predictive modeling of these networks. The National Research Council (NRC) (2005) in a recent report on Network Science identified a lack of understanding in the stochastic behavior of networks. They further stated that there existed a great need for this understanding in order to develop effective predictive models. Twenty percent of the research tasks in the Office of Naval Research's (ONR) recent broad agency announcement 07-036 were in the area of social networks. One of the research tasks were for "real time methods for the analysis of networks." Another task was to develop "metrics extracted in real time to diagnose effective or ineffective collaboration or negotiation," and for creating "unobtrusive data collection methodologies" for social networks. The U.S. Army Research Institute for the

Behavioral and Social Sciences (ARI) has requested research in social networks to "investigate individual unit and organizational behavior within the context of complex networked environments" in their fiscal year 2008 BAA. The U.S. Army Research Office has already budgeted over $1 Million per year for faculty and cadets at the U.S. Military Academy to study the stochastic behavior of networks. The National Academies identified the need for research in this area as early as 2003 in the Dynamic Social Network Modeling and Analysis workshop in Washington, DC.

While this research will not predict network behavior, it will provide an approach for more accurately detecting *that* a change occurred and *when* that change likely occurred. This is an important first step for any predictive analysis. If a social scientist can accurately detect change and the time change occurred, only then can he investigate the cause of change with any real success. Therefore, I posit that this approach will contribute to longitudinal network analysis in general, enabling future researchers to address the problem of prediction.

Much research has been focused in the area of longitudinal social networks (Sampson, 1969; Newcomb, 1961; Romney, 1989; Sanil, Banks, and Carley, 1995; Snijders, 1990, 2007; Frank, 1991; Huisman and Snijders, 2003; Johnson et al, 2003; McCulloh et al, 2007a, 2007b). Wasserman et al. (2007) state that, "The analysis of social networks over time has long been recognized as something of a Holy Grail for network researchers." Doreian and Stokman (1997) produced a seminal text on the evolution of social networks. In their book they identified as a minimum, 47 articles published in *Social Networks* that included some use of time, as of 1994. They also noted several articles that used over time data, but discarded the temporal component, presumably because the authors lacked the methods to properly analyze such data. An excellent example of this is the Newcomb (1961) fraternity data, which has been widely used throughout the social network literature. More recently, this data has been analyzed with its' temporal component (Doreian et al., 1997; Krackhardt, 1998; Baller, et al. 2008).

Methods for the analysis of over time network data has actually been present in the social sciences literature for quite some time (Katz and Proctor, 1959; Holland and Leinhardt, 1977; Wasserman, 1977; Wasserman and Iacobuccci 1988; Frank, 1991). The dominant methods of longitudinal social network analysis include Markov chain models, multi-agent simulation models, and statistical models. Continuous time Markov chains for modeling longitudinal networks were proposed as early as 1977 by Holland and Leinhardt and by Wasserman. Their early work has been significantly improved upon (Wasserman, 1979; 1980; Leenders, 1995; Snijders and van Duijn, 1997; Snijders, 2001; Robins and Pattison, 2001) and Markovian methods of longitudinal analysis have even been automated in a popular social network analysis software package SIENA. A related body of research focuses on the evolution of social networks (Dorien, 1983; Carley, 1991; Carley, 1995; Carley 1997; Dorien and Stokman, 1997) to include three special issues in the Journal of Mathematical Sociology (JMS Vol 21, 1-2; JMS Vol 25, 1; JMS Vol 27, 1). Evolutionary models often use multi-agent simulation. Others have focused on statistical models of network change (Feld, 1997; Sanil, Banks, and Carley, 1995;

Snijders, 1990, 1996; Van de Bunt et al, 1999; Snijders and Van Duijn, 1997). Robins and Pattison (2001, 2007) have used dependence graphs to account for dependence in over-time network evolution. We can clearly see that the development of longitudinal network analysis methods is a well established problem in the field of social networks. Table 1 provides a comparison of the dominant methods for longitudinal network analysis.

The literature shows that there exist four network dynamic states in longitudinal social networks. A network can exhibit *stability*. This occurs when the underlying relationships in a group remain the same over time. Observations of the network can vary between time periods due to observation error, survey error, or normal fluctuations in communication. A network can *evolve.* This occurs when interactions between agents in the network cause the relationships to change over time. A network can experience *shock.* This type of change is exogenous to the social group. Finally, a network can experience a *mutation.* This occurs when an exogenous change initiates evolutionary behavior.

Much of the research in longitudinal social networks has focused on evolutionary change. Markov methods and multi-agent simulation are effective at helping social scientists understand evolutionary change. However, a careful review of the literature did not reveal any research in detecting shock or mutations in the network.

SNCD provides a statistical approach for detecting changes in a network over time. In addition to change detection, change point identification is also possible. Identifying changes and change points in empirical data, will allow social scientists to better isolate factors affecting network evolution as well as the relatively new concept of shock. Moreover, knowing when a network change occurs provides an analyst insight in how to bifurcate longitudinal network data for analysis.

A complete review of methods for longitudinal social network analysis is beyond the scope of this thesis. The reader is referred to Wasserman and Faust (1994); Dorien and Stokman (1997); and Carrington, Scott and Wasserman (2007). Essentially, methods for longitudinal social network analysis have been focused on modeling and testing for the significance of social theories in empirical data. These methods have not been designed to detect change over time. This thesis is focused on detecting change in a social network over time.

Table 1. Dominant Methods for Longitudinal Network Analysis

	Markov Chain	Multi-Agent	Statistical	SNCD
Problem Addressed	1. Network evolution based on Markovian assumptions. 2. Determine how underlying social theories affect group dynamics.	1. Network evolution based on node-level behavior. 2. Evaluate the impact of social intervention on group dynamics.	1. Compare the properties of networks at different points in time.	1. Detect change (shock, evolution, or mutation) over time in empirical networks.
Key Assumptions	1. Future behavior of network is independent of the past. 2. There is no exogenous change in the network.	1. Node level behavior can drive group behavior. 2. Underlying social theories affecting group dynamics are known.	Assumptions vary, but include such things as dyadic independence/ dependence, over-time independence, one node class.	Group behavior can be inferred from longitudinal social networks
Limitations for change detection	1. Does not account for exogenous change. 2. Markov assumption.	1. Used to model both exogenous and evolutionary change, but not to detect change. 2. Underlying social theories must be known.	1. Does not handle over-time dependence. 2. Not a longitudinal approach.	1. Ergodicity and dependence is not fully addressed.
Strengths	Determining significant social theories affecting group dynamics.	Simulating group dynamics in a social network.	Comparing social networks.	Detecting changes in empirical social networks over time.

1.2 Application

This thesis will provide insight into the stochastic behavior of social networks. In addition, algorithms will be proposed that detect subtle changes in a social network. Imagine Joe Analyst working in an intelligence center trying to understand the dynamics of global terrorism. He currently has a wide array of tools to assist him. He can piece together social networks from news papers and broadcasts, intercepted voice

communication, and intelligence gathered from field agents. He can model this information with social networks and use various measures to identify individuals who are well connected, influential, or connect otherwise disconnect terrorist cells. In other words, he can tell you who was likely responsible for an attack in the past, and who was influential in the organization. But, what about today? Have influential members become less important? Are other members of the organization assuming more influential positions in the social network? Can we detect a change in the social network of a terrorist organization as they increase their communication before they are able to execute their planned terrorist attack? These are the questions that this research will help answer.

Applications are not limited to the military. Consider a civilian company, whose managers can identify major leadership challenges before they affect the productivity of the company. The introduction of e-mail and cell phones into the workplace has significantly changes the dynamics of communication. In the past, workers had limited peers available that they could ask about problems, before they had to seek guidance from senior management. Today, the available peers to consult are limited only by a person's social network. With growing on-line communities of practice, this network is becoming larger and larger. While it may be good that workers are able to resolve problems at a lower level, senior managers are unable to influence decisions with their senior judgment and experience. This research will provide those managers with a tool to detect potential problems in their organization, by detecting subtle changes in the social network of employees.

1.3 Contribution

Several contributions to science are derived from this thesis. I formally present a framework to understand longitudinal networks building off of Doreian's (1997) work in network dynamics. I introduce a methodology to detect changes in networks over time. Issues of over-time dependence and periodicity are addressed. I introduce a model of a network in dynamic equilibrium and compare it to competing models in the literature. I then tie this model to a multi-agent simulation, thereby contributing to the validation of multi-agent models for over time network data. I empirically derived an analytic expression to determine the decision interval used in a cumulative sum statistical process control chart that can not only be used in SNCD, but in any automated quality control application in industry. I proposed a statistical model for the degree distribution of nodes in a network that unifies the random networks of Erdős and Rénia with the scale-free networks of Barabasi. Finally, I have written software that has implemented these methods in a powerful social network analysis tool. Various analysts and scientists throughout DoD and academia are currently using these methods.

The demand for these tools is already here. Major General Lovelace of the Army's Central Command has requested an officer to begin measuring and monitoring the social network of senior leaders in their organization to improve the efficiency of their organizational behavior. The National Security Agency and U.S. Special

Operations Command have also requested tools to aid in detecting real time changes in social networks that they monitor. These important tools have become a reality with this research and are automated in ORA, a powerful SNA software package.

1.4 Organization

This thesis is essentially a collection of original published papers in the area of SNCD. As such each chapter is written such that it can be read independently. There will therefore be some intentional redundancy between chapters.

Chapter 2 introduces several statistical concepts related to networks. Random networks are contrasted with scale-free networks. Important issues such as defining the context of relational data are highlighted. This chapter provides rationale to make several assumptions used in my derivation of change detection approaches that are presented in Chapter 4.

Chapter 3 presents a model for stable, non-evolving networks over time. More importantly several network simulation approaches are compared. Multi-agent simulation is found to be a good approach for modeling networks over time. This is an important issue in understanding network change. With many real-world networks, there are many competing factors that may cause a change in the network. Isolating the "real" cause of change can be very difficult. Simulation offers a controlled environment, where change can be introduced a defined, known point in time, allowing change detection approaches to be evaluated fairly. As such, an important aspect of this thesis is the validation and verification of multi-agent simulation models used to explore change detection in networks. The performance of competing change detection methods are evaluated using simulation in subsequent chapters.

Chapter 4 provides an overview of statistical process control and introduces its application to longitudinal networks. Several different statistical process control approaches are derived and compared using simulated network data.

Chapter 5 introduces a method to handle over-time dependence and periodicity in network data. An example of periodicity occurs in email networks. People are more likely to communicate at certain times of the day, days of the week, etc. These typical fluctuations in communication can affect SNCD and increase the probability of false-positives (incorrectly determining that the network changed). Fourier analysis is used to identify periodicity in the data. Two approaches to handling the periodicity are proposed.

Chapter 6 demonstrates the proposed SNCD approach on real-world longitudinal network data. Real-world data is used to further demonstrate SNCD in practical application, highlighting the insight that change detection provides a social scientist.

Chapter 7 proposes a method for conducting SNCD in situations where network measures exhibit high variance or when there are only a few nodes in the network.

Chapter 8 presents some investigation into the robustness of SNCD to missing links in the network data. In addition the correlation between various network measures are also explored and implications are drawn for SNCD.

Chapter 9 summarizes the contributions of the thesis, identifies several limitations of the proposed approach, and provides directions for future research. The appendices provide more details and tutorials on certain experimental aspects of the thesis.

Appendix A is a primer on social network analysis. Since this thesis draws on methods from diverse disciplines, some readers may not be familiar with social network analysis. This appendix will provide an introduction for these readers and should be read before proceeding to Chapter 2 if necessary.

I have developed multiple methods to collect longitudinal networks from email. These methods are explained and compared in Appendix B. While longitudinal networks can be obtained from many different sources, this appendix will provide an approach to quickly and unobtrusively gather longitudinal social network data.

In Appendix C, I have empirically derived an analytic equation for determining the decision interval in a cumulative sum statistical process control chart, so that it can be implemented in software. This equation has application both within SNCD, but also with any automated statistical process control chart designed to detect small changes in a process. This appendix is helpful for analysts attempting to determine appropriate parameter values for the cumulative sum control chart.

Finally, Appendix D provides further details on the simulation used to demonstrate SNCD in Chapter 4. The multi-agent model specification is presented. It is docked with other simulation models used within the DoD.

2 Statistical Distribution of Networks

A major challenge in describing change in a network is to first understand the statistical distributions that describe the typical behavior of a network in the first place. This challenge has been a rich area for research for more than 60 years. For more information on network statistics the reader is referred to Wasserman and Faust (1994), Carrington, Scott, and Wasserman (2007), Barabasi, Watts, and Newman (2005). This chapter will focus on a single property of a network, the degree. The degree, *k* of a node *i* is the number of other nodes linked to node *i*.

A popular topological property of networks is the distribution of the degree of nodes in the network. Many have fit power law distributions to this property (Barabasi and Albert, 1999; Goh, Kahng, and Kim, 2001; Barabasi, 2002; Barabasi et al, 2002; Bollobas and Riordan, 2003; Pastor-Satorras and Vespigani, 2004; Newman, 2005; Caldarelli, 2007). This has given rise to the Scale-Free network inspired by statistical mechanics, which is characterized by having a power law degree distribution which tends to create hub nodes that are highly connected, while most nodes have relatively few connections in the network. Some have criticized this approach for social networks on the grounds that it does not consider the context of the network data and assumes universal network behavior (Doyle et al, 2005; Wasserman, Scott and Carrington, 2007; Alderson, 2008).

Another popular model for degree distribution is the random graph (Erdos and Renyi, 1959, 1960, 1961; Wasserman and Faust, 1994). Under this model, nodes in the network are connected with equal probability. This has been used in many cases as a null hypothesis to test the "randomness" of empirical network data (Wasserman and Faust, 1994; McCulloh et al, 2007; Alderson, 2008). Under this model of the distribution of node degree, the degree tends to appear normally distributed as the number of nodes increases and probability of connection remains the same. Some argue that this model is therefore not representative of empirical data (Watts and Strogatz, 1998; Barabasi, 2002; Pastor-Satorras and Vesppignani, 2004; Newman, 2005; Caldarelli, 2007).

Between these extremes, more realistic network models have been proposed by introducing more detailed models of the probability of node connection. Some of these models have been based on established social theory and multi-agent simulation (Carley, 1990, 1995, 1999; Doreian and Stokman, 1997), while others have been based on structural properties within the network (Frank, 1991; Snijders, 2007; McCulloh, Lospinoso and Carley, 2007; Lospinoso, 2008). One recent model involved a new node randomly attaching to the network and then "burning" through the network making new connections with other nodes like a wildfire spreads (Leskovec, Kleinberg, and Faloutsos, 2005). Unfortunately, these models do not leave us with a tractable analytic distribution of degree.

This chapter will highlight three novel points relevant to network science in general and this thesis in particular. First, different statistical distributions are fit to the same simulated data, showing that it is extremely difficult to assert with any certainty the

actual distribution of network properties. In other words, there are several candidate distributions which may all fit the data equally well. Second, the context that defines a relationship may significantly affect the structure of the network. For example, a social network where a relationship is defined as person *i* has seen person *j* before will be much more dense than a network where a relationship is defined as person *i* spends at least 1 hour in conversation with person *j* each day. Third, a power series distribution is proposed for the degree measure in a network. The binomial distribution common in random networks is shown to be a special case of the power series distribution. The power series distribution is also a power law distribution, which can fit the distribution of degree in other network structures. The parameters of the power series distribution can be viewed as a constraint function on the utility and costs associated with establishing a link. In the previous example, a node incurs greater cost in terms of time to spend one hour per day with another node than the cost associated with simply seeing another node one time. Together these three points provide direction for future research and identify an area of caution for detecting change which will be articulated later in this thesis.

2.1 Random Networks

The simplest model of a network is the *random network* (Erdos and Renyi, 1959, 1960, 1961). This model contains a fixed number of nodes, *n*. Between each ordered pair of nodes, a link is drawn with some probability, *p*. Under this model, all nodes have an equal probability of being connected to every other node and therefore they behave similarly to one another. Individual nodes can occupy very different positions in the network with the same random probability. Some nodes will be on the periphery of the network, while others will occupy a more central position. The degree to which this occurs depends only upon the random realization of an instance of the network.

If we fit statistical distributions to the nodal measures in a random network, we can create a statistical model to evaluate the position nodes occupy in the network. If a node occupies a position that is significantly anomalous to positions expected under random network assumptions, we might conclude that the network has an unusual node or that the topology of the network is not random. This insight is the first step towards a better statistical understanding and description of network behavior.

The degree, k_i, of node *i* is defined as the number of other nodes directly linked to *i*. The degree of a node will contribute to their influence in the overall group (Freeman, 1977). The degree will not determine influence alone, however. In Figure 4, it is clear that nodes 5 and 9 have the highest degree, but node 10 may have greater influence in the network due to its' more central position (Freeman, 1977; Wasserman and Faust, 1994).

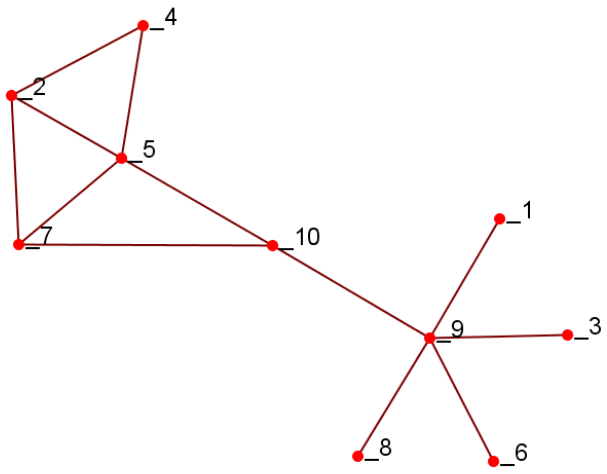


Figure 2. Example Network.

The distribution of degree in a random network has been well established in the literature (Erdos and Renyi, 1960; Donninger, 1986; Bollobas, 1998; Caldarelli, 2007). Suppose a network with n nodes follows the Erdos and Renyi model with probability p. Then, a node of degree k occurs if and only if there are exactly k link successes of the $n-1$ possible links. The degree k follows a binomial distribution, satisfying the four properties of the binomial distribution: 1) There are a fixed number of nodes, n, in the network; 2) There is a fixed probability, p, of any two ordered pairs of nodes being linked; 3) The trials are identical, which occurs when nodes behave similarly; 4) The trials are independent, which means that a node's choice of connections is not influenced by other nodes in the network. The binomial degree distribution is given by,

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

The average degree is therefore given by $p(n-1)$ and the variance is given by $(n-1)p(1-p)$. The maximum likelihood estimate (MLE) of the average degree is given by

$$\bar{k} = \frac{1}{(n-1)} \sum_{i=1}^{n-1} k_i = p(n-1).$$

Therefore, the MLE of $p = \bar{k}/n$. If we assume that the node degrees are independent[1], then by the Central Limit Theorem, for large n, the average degree is an approximately normally distributed random variable. Since the probability of link occurrence, p, is a linear combination of a normal random variable, it too is a normal random variable. We can now proceed to derive a confidence interval about the estimate of p. This is done

[1] This assumption is not valid for all networks. The impact of ergodicity in networks is still an open research area.

more easily by relating the distribution of the number of links in the network to the parameter *p*.

We can derive a similar distribution for the number of links, *L*, in a random network. There are $n(n-1)$ possible links in a social network if we exclude reflexive links. If the probability of a link occurring is a constant probability, *p*, then the distribution of links in a random network also follows a binomial distribution given by,

$$P(L) = \binom{n(n-1)}{L} p^{L} (1-p)^{n-1-L}.$$

If the number of links follows a binomial distribution, then the average number of links is given by $np(n-1)$ and the variance of the links is given by $np(n-1)(1-p)$. The MLE of *p* is therefore $\bar{p} = L / n(n-1)$. Recall that we concluded that *p* is a normally distributed random variable when *n* is large and node degree is independent. If we take the difference of a normally distributed random variable and its' MLE and divide the result by the standard deviation of that random variable, we have a standard normal random variable. Therefore,

$$z_L = \frac{P - \hat{P}}{\sqrt{\hat{P}(1-\hat{P})/n(n-1)}} \sim \text{Normal } (0,1)$$

This algebraically reduces to $P = \hat{P} \pm z_{L\alpha/2}\sqrt{\hat{P}(1-\hat{P})/n(n-1)}$, which defines a confidence interval on $\bar{p}$ at the $1-\alpha$ confidence level, where $\hat{P} = L / n(n-1)$.

Any value of $L / n(n-1)$ or of $\bar{k}/n$ that exceeds the constructed confidence interval is therefore statistically anomalous at the $1-\alpha$ confidence level for a random network. If we conclude that a network does appear random, we can conduct a similar procedure for the degree to determine statistically anomalous nodes in the network.

$$z_k = \frac{k - \bar{k}}{\sqrt{np(1-p)}} = \frac{k - p(n-1)}{\sqrt{np(1-p)}} \sim \text{Normal } (0,1).$$

This reduces to $k = p(n-1) \pm z_{k_{\alpha/2}}\sqrt{np(1-p)}$. Any value of *k* that exceeds this value is statistically anomalous at the $1-\alpha$ confidence level for a random network.

2.2 Scale-Free Networks

Scale-free networks were introduced by Barabasi and Albert (1999). The term scale-free refers to their observation that "the distribution of their local connections is

free of scale, following a power law." The density function of the power law distribution used by Barabasi and Albert is given by,

$$f(k) = ck^{-\gamma},$$

where c is a positive finite constant and $\gamma > 0$. The parameter γ can be referred to as the shape parameter and c can be referred to as the scale parameter. The particular range of values for γ that makes a network "scale-free" has not been defined (Bollobas and Riordan, 2003; Albert, 2008; Alderson, 2008; Barabasi, 2008). This is an important distinction, because a value of γ that approaches 0 approximates a uniform distribution, which has very different properties than the networks we typically think of as scale-free. At the U.S. Military Academy's 3rd Network Science Workshop, Barabasi stated, "the interesting networks that we look at have a [shape parameter] between 2 and 3." For these networks the mean and higher order moments of the degree distribution are undefined. When the moments of a distribution are undefined, it becomes difficult to characterize the modeled process in a compact form. This is perhaps one of the intriguing challenges presented by modeling network structure.

Scale-free networks present a much more fundamental problem in terms of degree distribution. So far the networks most commonly modeled with a scale-free network are dichotomous, large networks with hundreds to millions of nodes. The degree of a node in a dichotomous network is always a discrete, countable, integer. Therefore, the power law density function applied in the network science community is merely an approximation of what is really a discrete probability mass function. While this approximation might be totally acceptable for networks with many nodes, keeping the distribution discrete allows us to explore the relationship between the random and scale-free networks.

A discrete version of the power law distribution is a power series distribution and is given by,

$$p(k) = \frac{a_k \gamma^k}{\sum a_k \gamma^k},$$

where the coefficient function $a_k > 0$, and γ is the shape parameter of the distribution. It is also possible to use a Zipf distribution (Newman, 2005), which is more analogous to the power-law distribution used by Barabasi and Albert. Both distributions follow a power law and have the ability to model data with a similar shape as I will show later in the chapter. I choose to use the power series distribution to show the similarity between random and scale-free networks. Both the Zipf and power series distribution are more appropriate distributions for modeling the degree of a node in a dichotomous network with a fixed number of nodes. Using the power series distribution, I show an interesting variate relationship that relates scale-free and random networks.

2.3 Variate Relationship Between Binomial and Power Series

The following derivation shows an interesting relationship between random and scale-free networks. As shown earlier, the degree distribution in a random network follows a binomial distribution. The degree distribution in a scale-free network follows a power series distribution. Binomially distributed data may appear as a discrete power series distribution when p is small (Evans Hastings and Peacock, 2000). As I stated earlier, the power series distribution is a more appropriate distribution for node degree in a dichotomous network, since the degree of a node is a countable integer bounded between 0 and $n - 1$.

The variate relationship between the binomial and the power series occurs where $k \sim \text{Binomial}(n, p)$, then $k \sim \text{PowerSeries}(\gamma)$, where the Power Series distribution is defined as,

$$\Pr(k) = \frac{a_k \gamma^k}{\sum a_k \gamma^k}$$

when $\gamma = p / (1 - p)$ and $\sum a_k \gamma^k = (1+\gamma)^n = (1-p)^{-n}$. Substituting these values into the expression for the power series probability mass function above, the mass function can be expressed as,

$$\Pr(k) = \frac{a_k (p/(1-p))^k}{(1-p)^{-n}} = a_k p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}.$$

The resulting mass function is the binomial probability mass function. It is important to note that the power series distribution can also be used to represent non-binomial distributed data by modifying the coefficient function, a_k. This modification allows the power series distribution to model the degree distributions commonly associated with scale-free networks. However, this variate relationship presents a single distribution that can be used to model both the degree distribution found in random and scale-free networks.

A limitation of the binomial is that the expected degree is always $n*p$, which can become unrealistic in very large networks. It can also be shown that for a critical value of p the network will undergo a phase transition between a network with many unconnected components to a single, giant component (Bollobas, 1998). These two behaviors typically distinguish random networks from scale-free networks.

2.4 Constrained Node Network

The random network is not always an unreasonable model for the distribution of degree in social networks (McCulloh et al, 2008; Ring, McCulloh and Henderson 2008). This is especially true of small networks under 100 nodes. I have conducted an extensive

search of the literature and modeled the degree distribution of many social networks (Newcomb, 1961; Sampson, 1969; Freeman and Freeman, 1980; Krackhardt, 1987; Johnson, Boster and Palinkas, 2003; McCulloh et al, 2007). When the relationship between nodes requires a meaningful investment of time, the social networks are not scale free. A person's website on Facebook is unconstrained in the sense that it can maintain an unlimited number of links pointing to it (in-degree). It is slightly constrained in the number of links that it points to (out-degree) in that it takes some amount of time and effort to find friends and establish a link to their web page. When we stipulate that two individuals must spend an hour together and engage in meaningful conversation each week for a relationship (link) to exist, we have severely constrained the degree of the node. The scarce resource of time prevents any node from maintaining a large number of connections as they might in a scale free network.

The size of the network is also an important consideration in degree distribution. A node's choice of which other nodes to connect with can be limited by the size of the network. For a group of 20 individuals, for example, it is quite possible that all people have a connection. When the size of the group contains thousands of individuals, it becomes virtually impossible for all nodes to be connected. Simply consider the amount of time required to hold a meaningful conversation between a given node and all others in the network. This further suggests that there is some kind of limitation on the number of connections a node can have. In addition to the cognitive limitations of a person, nodes can be constrained by proximity, the utility of the connection, and in some cases the cost of the connection.

The constraints on link formation between nodes in a network will be context specific. The number of social ties that a person can maintain might be limited by their time, cognitive capacity, proximity, and other factors. A web-page, on the other hand, is not limited in the number of other web pages that can connect to it. There are limitations, or at least a cost in terms of building the web site, in the number of links to other pages, and still other factors. Relationships in a social network can also be very general, such as "been to the same country", which would connect many individuals that do not know each other at all.

The probability of link formation could therefore be modeled as a function of the context specific constraints. The constraints could prevent the occurrence of hub nodes, characteristic of scale-free networks in one context, but not in another. While the functional forms could be similar between different contexts, and therefore their network structure will be similar; the underlying functions governing constraints in the network may be quite different. Therefore, I propose modeling the constraints in the network as some form of the coefficient function a_k. An understanding of how network constraints can affect topological properties within the network is dependent upon how relationships are defined in network data. This is critical to understanding the behavior of the network.

2.5 Virtual Experiment

A virtual experiment is conducted to show how power series distributions with differing constraint functions can be used to model scale-free networks generated under the Barabasi-Albert model. I remind the reader that real-world data does not follow this model for many applications, particularly when the relationships are constrained by time or other scarce resources. However, scale-free networks have been successfully used to model the internet, electrical networks, protein networks, and other real-world data (Goh, Kahng and Kim, 2001; Pastor-Satorras and Vespignani, 2004; Leskovec, Kleinberg and Faloutsos, 2005; Caldarelli, 2007). Since random networks are already an established model in the social network community (Wasserman and Faust, 1994), I compare the power series distribution to the scale-free network in an attempt to unify the competing models.

In order to demonstrate the flexibility of the power series distribution and demonstrate differing constraint functions, 3000 scale-free networks were simulated consisting of 1000 nodes each. Networks were generated with degree distributions having γ = 2, 2.5, and 3. These values of γ were chosen because they were said to be the "interesting" networks by Barabasi (2008). Since he coined the term "scale-free" network, these seemed to be the best point of comparison found in the literature. I generated 1000 networks for each value of γ, therefore there were 3000 total scale-free networks.

Four statistical distributions are fit to the degree of the networks using the method of least squares. This method is chosen, because it minimizes the error in the fit of the distribution, which can be used to compare the quality of the fit among the four different distributions. The four distributions include the Barabasi and Albert power law distribution and three power series distributions, each with a different constraint function. The first constraint function is the *binomial constraint* function given by, $a_k = \binom{n-1}{k} = \binom{999}{k}$. The second power series distribution uses a *constant constraint* function given by, $a_k = 1$. The third power series distribution uses a more flexible *inverse constraint* function given by, $a_k = \left(\frac{1}{k}\right)^{\lambda}$.

The coefficient of determination is calculated for each distribution's fit to each of the 3000 networks. The coefficient of determination is calculated using the formula,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{p}_i - p_i)^2}{\sum_{i=1}^{n}(p_i - \bar{p}_i)^2},$$

where p_i is proportion of the empirical data that has a degree less than the degree of node i, $\hat{p}_i$ is the probability of observing a random value less than the degree of node i under proposed distribution, and $\bar{p}_i$ is the average of the proportion of the empirical data which is equal to 0.5. As the error in the fit of the distribution increases with respect to the variance, the value of R^2 decreases. Using the method least squares will minimize the error in the fit of the distribution. Since we are comparing which distributions are appropriate for the data, this method maximizes the value of R^2 and gives each distribution a best case fit for comparison. This is particularly important with heavy tail data such as the power-law, since observed values in the tail of the distribution can

significantly bias the estimation of distribution parameters. I acknowledge that the parameters of those distributions are not minimally variant unbiased estimators for the distributions. In practice, it is desirable to use a Maximum Likelihood Estimate of the distribution parameters. For this application, the method of least squares is clearly more appropriate.

The coefficients of determination were compared between a Barabasi - Albert power law distribution fit to the data, and three power series distributions with different constraint functions fit to the same data. This was repeated for all 3000 networks. A two-sample t-test is used to evaluate the null hypothesis that the coefficients of determination of the best fit power series distribution are the same as the Barabasi - Albert power law distribution. An important distinction is made. In this virtual experiment, the underlying stochastic process that generates the data is the Barabasi - Albert power law distribution (Albert and Barabasi, 1999). The coefficient of determination for this distribution is therefore a measure of the error in the simulated data. When there is insufficient evidence to reject the null hypothesis, this only suggests that the power series distribution is a possible alternative distribution. In real-world data, we cannot know the underlying distribution. Therefore, a lack of evidence to reject the null hypothesis in this virtual experiment has important implications for empirical studies. It means that there is no evidence to suggest that the Barabasi - Albert power law distribution is a better fit to the empirical data than the best fit power series distribution.

2.6 Results

The 1000 scale-free networks for each parameter of γ were successfully created for a total of 3000 independently seeded networks. The node degree for all 1000 nodes in each network was recorded. Figure 3 displays the empirical distribution function for a representative instance of the degree distribution for the networks where γ = 2, 2.5, and 3 respectively. This figure illustrates the importance of the parameter γ in defining a scale-free network. As the value of γ approaches 1, a uniform distribution can be fit to the data. As the value of γ increases, the curvature in the data increases.

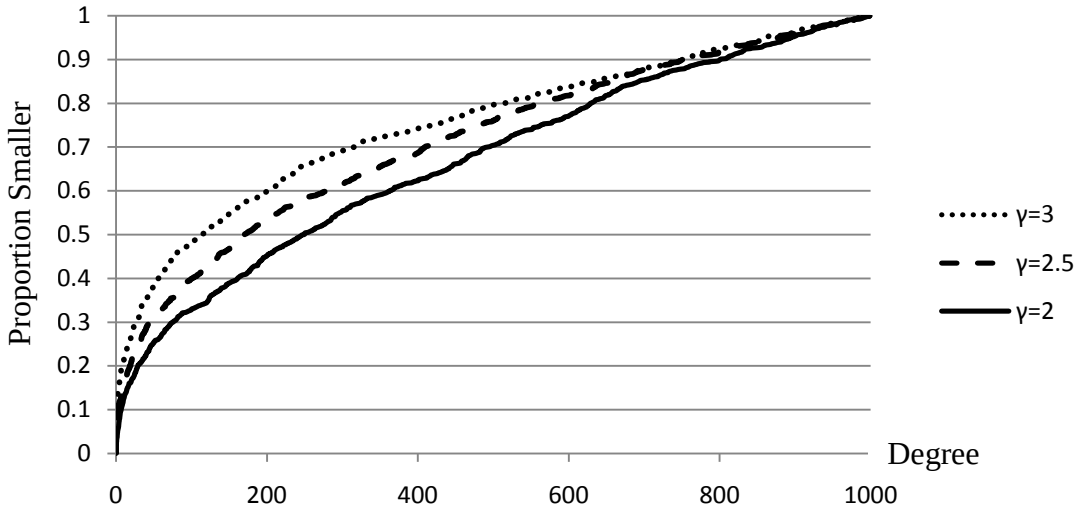


Figure 3. Empirical distribution functions for the degree of a scale-free network

The three power series distributions and the Barabasi-Albert power law distribution were fit to each of the 3000 generated networks. Figure 4 shows a representative instance of the empirical distribution, each of the three power series distributions using the different constraint functions, and the Barabasi-Albert power law distribution fit to the most extreme degree distribution where $\gamma = 3$. There is a noticeable difference between the fit of the power series distribution using the binomial constraint function and the empirical distribution function. This has been extensively noted in much of the network science literature. There is not a noticeable difference between the power series distribution using the inverse constraint function and the empirical distribution function. This suggests that a power series distribution may provide an equally good fit to network data as the Barabasi-Albert power law distribution provides.

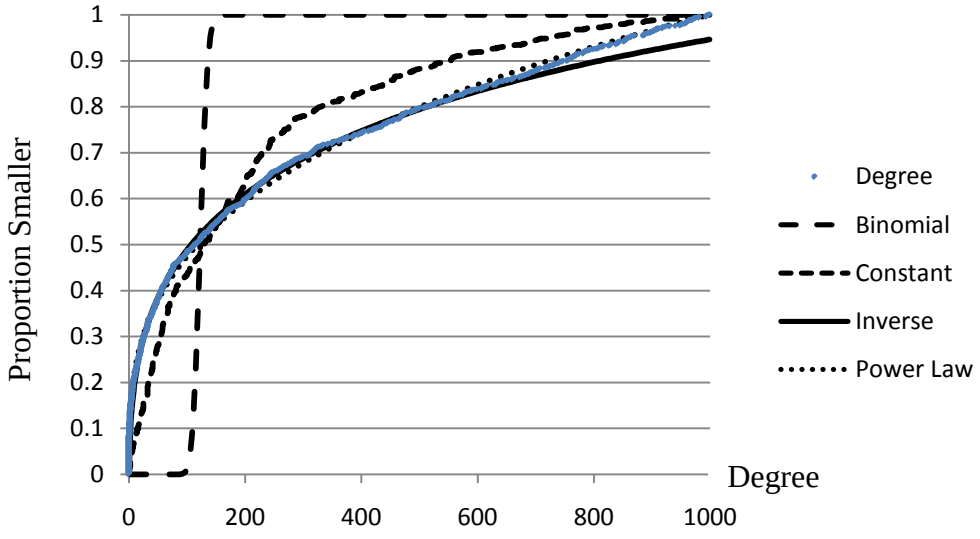


Figure 4. Various distributions fit to scale-free empirical degree distribution.

The coefficients of determination provide a quantitative measure of the quality of fit of the distribution. Table 2 shows the average coefficients of determination across the 1000 simulated networks of each parameter value γ, for the four distributions fit to each empirical distribution function. A value of 1.0 indicates that the distribution fits the data with no error. A value of 0.0 indicates that the distribution has no explanatory power in describing the data.

Table 2. Coefficients of Determination for Four Distributions.

	Barabasi-Albert	Binomial	Constant	Inverse
$\gamma = 2.0$	0.9984	0.0929	0.9444	0.9970
$\gamma = 2.5$	0.9989	0.0927	0.9111	0.9971
$\gamma = 3.0$	0.9966	0.0926	0.8531	0.9937

The power series distribution with the binomial constraint function does not fit the empirical data well. It is possible to improve the fit of this distribution if I allowed the size parameter, n, to increase. Since the size of the network was set at 1000 nodes, I decided it would be a more appropriate comparison to fix this parameter at $n = 1000$ to be consistent with the other distributions. The power series distributions with both the constant and inverse constraint functions provide a reasonably good fit to the data. I again remind the reader that the values in Table 2 compare the power series distributions to an empirical data set that definitely has a Barabasi-Albert power law underlying stochastic process. It has not been established that the Barabasi-Albert power law distribution models real-world empirical data as well as the power series with constant constraint function. In fact there have been several papers that have questioned the appropriateness of the Barabasi-Albert power law distribution applied to empirical findings (Doyle et al, 2005; Wasserman, Scott and Carrington, 2007; Alderson, 2008).

The power series distribution with the inverse constraint function is compared to the Barabasi-Albert power law distribution using a two sample t-test. The t-test is repeated for each set of 1000 networks corresponding to $\gamma = 2$, 2.5, and 3. Table 3 shows the results of the t-test.

Table 3. Statistical Comparison of Power-Law and Power-Series Distributions.

	Barabasi-Albert	Inverse Constraint	Test Statistic, T	p-Value
$\gamma = 2.0$	0.9984	0.9970	0.933	0.3510
$\gamma = 2.5$	0.9989	0.9971	1.294	0.1960
$\gamma = 3.0$	0.9966	0.9937	1.611	0.1075

There is no p-value < 0.10 that would indicate a statistically significant difference at the 90% confidence level. Therefore, we do not have enough evidence to reject the null hypothesis that the distributions are different. The reality of this virtual experiment is that they are different. This can be seen in the marginal trend, where each of the coefficients of determination for the power series with inverse constraint function are less

than the true underlying distribution. However, the fact that there was not a statistically significant difference with 1000 simulation replications, suggests that the two distributions are suitable to model the same data. It can also be seen in Table 3 that the p-value has a decreasing trend as γ increases. The power series distribution may not be appropriate for modeling data generated with a Barabasi-Albert power law distribution with $\gamma > 3$.

2.7 Discussion

I believe that the power series family of distributions may be a more appropriate statistical distribution for modeling the degree distribution of nodes in network data. Selecting a statistical distribution based purely on empirical fit can be misleading. Certain statistical distributions have properties that can have implications for the underlying mechanics of the system being modeled. For example, the exponential distribution has a memoryless property. This implies that future observations are independent of past observations. While this is appropriate for modeling the inter-arrival time of customers at a bank, it can be very misleading when modeling the arrival time of a bus, which comes at regular intervals. While the application of waiting for arrival is the same, the underlying mechanics are very different. In one situation the time between arrivals is independent of the past. In the other situation, the time past since the last bus arrived holds a great deal of information about when the next bus will arrive. A similar analogy may be true for the degree distribution of a network.

It is therefore necessary to consider the fit of various candidate power-series distributions when investigating the distribution of degree in network data. If a power-series distribution can be fit reasonably well to the data, the nature of the coefficient function may provide some insight into constraints on the nodes within the network. Further investigation into the coefficient functions of various network data may provide greater insight into the behavior of networks and possibly lead to predictive network models. At the very least, a single family of distributions for degree may lead to understanding appropriate statistical distributions for other network measures that will allow greater hypothesis testing in network data.

The power series distribution's coefficient function has been suggested to represent constraints on node degree. Further investigation is warranted. I have only suggested that this is a possible explanation for network degree distributions and shown some interesting variate relationships. Perhaps multi-agent simulations could constrain individual node degree and provide insight into the constraint function. Constraint functions could be modeled for various real-world network data and inference could be drawn about constraints based on the known or assumed behavior of the application.
We must be careful about the conclusions we draw from the degree distribution of network data. Just because a particular candidate distribution fits empirical data well, does not mean that the underlying distribution generating the data is the same. Scale-free networks that have a power law distribution were not necessarily evolved through preferential attachment (Albert and Barabasi, 1999) as some may assume. Preferential

attachment will produce scale-free networks, but the observation that the network is scale-free does not allow us to infer how it was created.

The power series distribution offers an alternative model to the Barabasi-Albert power law distribution. I believe that this is an appropriate statistical model due to the explanatory nature of the constraint function in the distribution. There is currently no intuitive explanation for the occurrence of power-law distributions in networks. In some sense the statistical support of the binomial distribution is more accurately defined for degree distribution than a continuous power-law distribution, due to the countable, integer values that degree can assume. Hopefully, future researchers will explore the topology of network data, looking at a wider range of candidate distributions and the implications for science that they may uncover.

3 Stochastic Nature of Networks

Social networks often exhibit stochastic behavior. For example, an agent in a network might communicate with a friend several times during a given day and not at all during another day. In this example, the underlying relationship remains the same; however, the observed network ties fluctuate. This is an intuitive example, however the accuracy of observed network data has been well documented in the literature (Killworth, et al, 1976, 1979; Bernard, et al, 1977, 1980, 1982; Krackhardt, 1990; Ellis, et al, 1991; Kashy and Kenny, 1990; Wasserman and Faust, 1994). Furthermore, it is possible that the underlying relationships in a social network may change (Carley, 1991; Doreian and Stokman, 1997; Snijders, 2007). This relatively common behavior will also cause fluctuations in observed network data. Therefore statistical models of social networks are necessary for any kind of meaningful inference on network data.

A necessary prerequisite for statistical inference of social networks is an underlying probability structure for the presence of links in the network. Detecting changes over time, comparing multiple networks, or evaluating a wide range of potential hypothesis all depend upon a method to estimate the probability of links occurring in an observed network. Several statistical models have been proposed. The *p** model was introduced by Frank and Strauss (1986). This model describes the distribution of a Markov random graph. Many others have contributed to developing this family of models (Strauss and Ikeda, 1990; Wasserman and Pattison, 1996; Anderson, et al, 1999; Wasserman and Faust, 1994), especially in the area of parameter estimation. A common approach to describe the link probability is the Exponential Random Graph Model (ERGM) (Krackhardt, 1998; Handcock, 2002, 2003; Hunter, 2006; Goodreau, 2007; Robins, et al, 2007; Hunter, et al, 2008). The ERGM is based on a regression of structural variables in the network that may explain the probability of links occurring in the network. Several have used the ERGM to simulate many instances of a given network and then estimate statistical properties of various network measures (Handcock, 2007; Butts, 2007; Goodreau, 2007). I introduce an alternative approach with the Link Probability Model (LPM) that uses the historical presence of links to estimate the link probability. I demonstrate both simulation approaches on a range of empirical data and show that for a limited number of longitudinal data sets, the LPM provides a better fit to the data than the ERGM.

The ERGM is a family of statistical models that describe the probability of a link being present between two nodes and is a common statistical model for social network analysis. The models are based on logistic regression, where model terms are usually structural variables in the network. The model is used to explore statistically significant properties of networks. The ERGM notation is also flexible, allowing it to represent a wide range of network variables. Unfortunately, many ERGM models are degenerate, meaning that observed data might be highly improbable given the model (Handcock, 2003, 2002). The ERGM is not typically used for over-time network analysis, however Mark Handcock presented an application of the ERGM for simulating networks at the 28th Sunbelt Conference (2008). The Link Probability Model (LPM) is not a statistical model, but rather a matrix of probabilities of a link being present between ordered pairs of nodes. The LPM is estimated from longitudinal networks based on the frequency of

links being present over time. The LPM avoids issues of model degeneracy because the model is not dependent upon highly correlated terms and there are more data points than parameter estimates. The LPM is particularly useful for my application, because I am only interested in modeling over-time data.

First, I briefly review the ERGM. Then the LPM is described and presented as an alternative model to the ERGM. Then the LPM and ERGM are both used to model four data sets: the Sampson (1969) Monastery data, the Newcomb (1961) Fraternity data, and then two sets of data from Fort Leavenworth (Graham, 2005; Baller, et al, 2008). These are four interesting data sets because they all have a temporal component and have been well documented in the literature. The fit of each of these models is compared to the data. I find that the ERGM is degenerate for the Fort Leavenworth data and that the LPM provides a better fit in the other two data sets under certain conditions. I conclude by discussing the strengths and limitations of LPM and its general usefulness to network analysts.

3.1 Exponential Random Graph Model

The ERGM is used in social network analysis as a statistical model that enables an analyst to conduct inference on dependent relational data (Goodreau, 2007; Robins, et. al., 2007). The ERGM is therefore less restrictive than the Holland and Leinhardt (1981) p_1 model that assumed dyadic independence. In many social network applications the relationship between two individuals depends on relationships between the individual and others in the network; cognitive limits on the number of relationships that can be maintained; similarity between individuals; and more. The ERGM framework for relaxing the dyadic independence assumption is thus essential for accurate inference in many data sets.

Exponential random graph models (ERGM) have been studied a great deal in the literature as a model for the probability of links occurring in a social network. The ERGM was first proposed in 1986 (Frank and Strauss) as a very general model. The ERGM can thus be used to model a wide range of explanatory variables. The basic ERGM is given by,

$$P(Y) \propto \theta_1 g_1(y) + \theta_2 g_2 + \ldots + \theta_k g_k(y) \qquad (1)$$

where Y is a graph, θ's are model coefficients, and $g(y)$ is a covariate or term in the model. Covariate terms are general and can represent many features of a graph. These terms are often structural properties of the graph such as the number of links, dyadic relations, and transitive properties, among others.

Estimating ERGM terms and parameters can be computationally challenging in large networks (Snijders, 2002; Pattison and Robins, 2002). Markov chain Monte Carlo estimation of ERGM has been used to fit these models to data (Goodreau, 2007; Robins, et. al., 2007; Handcock, 2003, 2002; Snijders, 2002; Pattison and Robins, 2002). The

Markov dependence in these models leads to problems of degeneracy, which is discussed in detail by Handcock (Handcock, 2003, 2002). Essentially, model degeneracy occurs when the observed data is almost impossible under the specified model. This often occurs when explanatory terms are highly correlated and there is insufficient data to construct an appropriate model. Many of the terms used in ERGM are correlated and it is difficult to define enough terms to preclude networks that do not represent the data, when they spuriously satisfy the ERGM terms. Several advances in ERGM have been proposed to include curved exponential family models (Hunter and Handcock, 2006) and neighborhood models (Robins, et. al., 2005). However, these advances have not completely removed issues of model degeneracy.

3.2 Link Probability Model Formulation

The LPM framework for viewing the probability space of a social network avoids issues of model degeneracy, while preserving flexibility for modeling dyadic relationships. It provides researchers with an improved means to understand the probability space of the network, under certain conditions. The LPM is a square matrix where the rows and columns correspond to the nodes in a social network. The entries are the link probabilities of the directed link from the row node to the column node. This is not to be confused with an adjacency matrix, where the entries are either zero or some number representing the strength of a relationship between nodes. The link probability is a number between 0 and 1, and determines the likelihood of a link being present in an observed adjacency matrix.

The link probabilities can be derived from empirical data in several ways. Given network data collected over multiple time periods on a group of subjects, the link probabilities can be estimated by the proportion of link occurrences, l_{ij}, for each cell in the adjacency matrix, $a_{ij.}$ In the case of communication networks, statistical distributions can be fit to the time between messages for each potential link in the network. For a specified period of time, t, the link probability p for each set of entities i and j can be found. Let x_{ij} be the time between messages in a communication network. The probability density function for any x can then be defined as $f_{ij}(x \mid \theta_{ij})$, where θ_{ij} is the set of parameters for the density function. Then, the probability, p, of a link occurring within some time period t is the probability that $x < t$, which can be expressed as,

$$p = \int_0^t f_{ij}(x \mid \theta_{ij})dx \qquad (2)$$

In practice, the function $f_{ij}(x \mid \theta_{ij})$ must be estimated using techniques such as maximum likelihood estimation from empirical data collected on the group being studied. It may be desirable to construct a network based on a restriction such as, “two emails within a time period demonstrate a relationship, but one does not.” In this case, it is necessary to compose a function of random variables. If $h_{ij}(2 \mid t, \theta_{ij})$ represents the probability density function of time between two sets of two emails and $f_{ij}(x \mid \theta_{ij})$

represents the probability density function of time between one set of two emails, then the following is true under certain assumptions:

$$h_{ij}(2 \mid \theta_{ij}) = \left(\int_0^t f_{ij}(x \mid \theta_{ij}) dx \right)^2 \tag{3}$$

It is possible to generalize this idea; if h_{ij} (x | t, θ_{ij}) is the probability that x or more communications occur within time t, then the following is true:

$$h_{ij}(x \mid \theta_{ij}, t) = \left(\int_0^t f_{ij}(y \mid \theta_{ij}) dy \right)^x \tag{4}$$

The LPM is an important improvement over some traditional models. Individuals in a social network are not connected to other individuals with uniform random probability. The probability structure is much more complex. Intuitively, there are some people whom a person will communicate with or be connected more closely than others. In a study of email communication conducted at the U.S. Military Academy (McCulloh et al, 2007) one subject emailed his wife more than ten times per day on average, while other people that he worked with received an email from him once or twice per month. For this reason, real-world networks tend to have clusters or cliques of nodes that are more closely related than others (Newman, 2003; Topper and Carley, 1999; Carley, 1996). This can be simulated by varying the probabilities that certain nodes will communicate. In this way, stochastic behavior in dynamic social networks can realistically be simulated.

The LPM is a desirable model due to its ability to accurately model empirical data and its ability to avoid degeneracy. The accuracy of the LPM will be discussed in the Results section. The LPM can avoid issues of model degeneracy because the only parameters for the model are the link probabilities. As long as there are at least two time periods for estimating parameters, there are more data points than there are parameters. Each link is treated independently of other links in the model; therefore, none of the terms are correlated. The naïve assumption of independence between links is corrected by the historic presence of links over time. Intuitively, links have some dependence. For example, if an individual chooses to communicate with another, the likelihood of that person reciprocating the communication increases. If we assume a dynamic equilibrium in the underlying relationships of individuals in the network, these patterns of dependent communication will be apparent over time. If node *i* has a high link probability with node *j*, it may be likely that node *j* has a reciprocal high link probability with node *i*. It is not necessary to directly account for this in the model. If the relationship is true, there will be a high expected occurrence of *i* to *j* and *j* to *i* links in the networks over time. The LPM will model these links with high link probability due to their over time frequency, and not directly from their structural dependency. In this way, the LPM can never be over specified, have high variance inflation, or be degenerate. Thus, the LPM may provide an attractive alternative to the ERGM for modeling longitudinal degenerate networks.

3.3 Data for Comparison

Four data sets are used to demonstrate the efficacy of the LPM. The first and second are longitudinal data sets that are well established in the SNA literature, namely the Sampson (1969) Monastery data and the Newcomb (1961) Fraternity data. The third and fourth data sets are larger in size. For the reader's convenience, Table 4 summarizes the similarity and difference among the data sets. All four are explained in more detail.

Table 4. Data Summary.

Name of data set	**Monastery**	**Fraternity**	**Leavenworth '05**	**Leavenworth '07**
Author	Sampson	Newcomb	Graham	Schrieber
Number of nodes	18	17	156	68
Number of time periods	3	15	8	9
Method of collection	Observation	Survey	Survey& Observation	Survey
Link weight	Dichotomous	Weighted	Dichotomous	Dichotomous
Link Relationship	Interpersonal relationship	Preference ranking	Self Reported Communication	Self Reported Communication
Change in density	0.17974-0.18301	0.50000-0.50000	0.01431-0.02906	0.04473-0.04628
Change in average betweenness	0.05556-0.05556	0.33574-0.41176	0.00880-0.00994	0.02009-0.01909
Change in average closeness	0.40158-0.02485	0.66510-0.39859	0.03759-0.05172	0.05739-0.08186
Change in average eigenvector cent	0.23428-0.23247	0.79907-0.74891	0.23591-0.22963	0.2125-0.22243

The first data set was collected in a monastery by Samuel F. Sampson (1969). The participants included 18 monks, and data was recorded on their interpersonal relationships. This is a directed network, where relationships are not necessarily reciprocal. Data was collected over three time periods, representing the time in which a new cohort joined the monastery.

The second data set was collected by Theodore Newcomb (1961) at the University of Michigan. The participants included 17 incoming transfer students, with no prior acquaintance, who were housed together in fraternity housing. The participants were asked to rank their preference of individuals in the house from 1 to 16, where 1 is their first choice. Data was collected each week for 15 weeks, except for week number nine. The relational data recorded between agents were ranks. Both the ERGM and LPM require dichotomous networks to construct a model. I chose to adopt the binarization scheme proposed by David Krackhardt (1998). He dichotomized the network data by assigning a link to preference ratings of 1-8 and having no link for ratings of 9-16. Krackhardt also fit an ERGM to the Newcomb Fraternity data which will be used for comparison with the LPM.

The third data set was collected from an Army war fighting simulation at Fort Leavenworth, Kansas in 2005, by Craig Schreiber and Lieutenant Colonel John Graham. The participants were mid-career U.S. Army officers taking part in a brigade level staff training exercise. This data set contains 156 individual agents that were monitored over the course of four and a half days. Data consists of communication ties between individuals as measured from self reported communications surveys. Surveys were completed at the end of each morning and at the end of the day before the officers went home. Therefore there are nine longitudinal time periods.

The fourth data set was also collected from an Army war fighting simulation at Fort Leavenworth, Kansas by Craig Schreiber; this time in April, 2007. There were 68 participants in this data set, who served as staff members in the headquarters of the brigade conducting a simulated training exercise. The data contains the communication between agents in the network which were collected through self reported communications surveys. Data was collected over a period of four days, twice per day. Thus, there were eight time periods.

3.4 Method of Comparison

The ERGM and LPM are investigated for their strengths and weakness in modeling longitudinal data. For the Sampson (1969) Monastery data, I use the ERGM that was fit to the data by Hunter et al (2008). The Akaike Information Criterion (AIC) is 302.61 and the Bayesian Information Criterion (BIC) is 436.65. The Hunter (2008) ERGM of the Sampson (1969) data was chosen for this study based on its more favorable AIC and BIC compared to other models found in the literature. I feel that this model is therefore an appropriate benchmark for comparison with the LPM. An ERGM is also fit to the Newcomb (1961) fraternity data. Again, I have chosen an ERGM accepted in the literature; this time the model proposed by Krackhardt (1998). An LPM is fit to both the Sampson and Newcomb data sets. Monte Carlo simulation is used to generate instances of the Sampson Monastery social network and the Newcomb Fraternity social network under the ERGM and LPM. In addition, an LPM is also fit to the two Fort Leavenworth data sets (Graham, 2005; Baller, et. al., 2008). For the two Fort Leavenworth data sets, the ERGM was degenerate. The ERGM were not degenerate for the Sampson or Newcomb data sets. The LPM is successfully used to model all data sets.

A distance measure is required to compare the similarity between the dichotomous networks generated using the ERGM, the LPM, and the empirical data. Hamming distance (1950) evaluates a distance between dichotomous networks. If the data were weighted networks and the models generated weighted networks as well, then a Euclidean distance would be appropriate. The quadratic assignment procedure (QAP) (Krackhardt, 1987b) could be used to compare the correlation between networks; however, I focus on network distance, because I intend to demonstrate that the LPM can generate simulated models that are very similar to the original networks in terms of actual distance and not simply a structural isomorphism.

The ERGM and LPM are evaluated on how well they model empirical data using a t-test. I illustrate the method with the Sampson Monastery data. Let the three networks in the Monastery data be labeled N1, N2, and N3. An ERGM is used to simulate networks and they are labeled E1, E2, E3, … E100,000. The LPM is also used to simulate networks and they are Labeled L1, L2, L3, … L100,000. The Hamming distances are calculated between each empirical data set to every simulated ERGM network and I use the following notation,

$$\text{Dist}_{\text{ERGM},1,1} = \text{Hamming}(\text{N1},\text{E1})$$
$$\text{Dist}_{\text{ERGM},1,2} = \text{Hamming}(\text{N1},\text{E2})$$
$$\dots$$
$$\text{Dist}_{\text{ERGM},i,j} = \text{Hamming}(\text{N}i,\text{E}j)$$
$$\dots$$
$$\text{Dist}_{\text{ERGM},3,100000} = \text{Hamming}(\text{N3},\text{E100000}).$$

The Hamming distances are also calculated between each empirical data set and every simulated LPM network and its notation is given by,

$$\text{Dist}_{\text{LPM},i,j} = \text{Hamming}(\text{N}i,\text{L}j).$$

The Hamming distances are calculated between each empirical data set and every other empirical data set and its notation is given by,

$$\text{Dist}_{\text{empirical},i,j} = \text{Hamming}(\text{N}i,\text{N}j), \text{ where } i \neq j.$$

This last set of Hamming distances are a measure of noise or observation error inherent in the data.

The ERGM and LPM are compared using a two-sample T-test between the Hamming distances from the empirical network, N*i*, and all of the simulated networks from the ERGM and the LPM. The test statistic is given by,

$$T_i = \frac{\mu_{ERGM,i} - \mu_{LPM,i}}{S_{P,i}/\sqrt{100{,}000}}$$

where,

$$\mu_{ERGM,i} = \frac{1}{100{,}000}\sum_j Dist_{ERGM,i,j}$$

$$\mu_{LPM,i} = \frac{1}{100{,}000} \sum_{j} Dist_{LPM,i,j}$$

and $S_{P,i}$ is the pooled standard deviation between the ERGM and LPM Hamming distances (Montgomery, 1991). This is repeated for each time period, *i*.

3.5 Results

An ERGM was fit to the Sampson (1969) Monastery data according to the model specification laid out by Hunter, et. al. (2008). Four model terms were used: links, sender, receiver, and mutual. A summary of the model fit is shown in Table 5.

Table 5. Fit Summary for Sampson ERGM.

Model Parameter	Coefficient	Standard Error	MCMC S.E.	p-value
Links	-2.5131	0.3361	0.005	0.0000
sender2	-0.7356	0.6854	0.015	0.2842
sender3	-0.2146	0.7274	0.017	0.7682
… output edited for length …				
receiver17	-1.2015	0.8191	0.018	0.1436
receiver18	-1.0562	0.7193	0.015	0.1432
Mutual	3.6816	0.6731	0.011	0.0000

The Hamming distance from each of the three empirical data sets to each of the ERGM simulated networks was calculated. The Hamming distance from each of the empirical data sets to each of the LPM simulated networks was calculated. The mean and standard deviation of these Hamming distances are displayed in Table 6. A two-sample t-test for each time period illustrates that the networks simulated using the LPM have a smaller average hamming distance to the empirical data sets than the networks simulated using the ERGM. This indicates that the LPM models the Sampson data more accurately than the ERGM model.

Table 6. Sampson Data Hamming Distances and T-test for ERGM and LPM.

Time Period	$\mu_{ERGM,i}$	ERGM Hamming Distance Standard Deviation	$\mu_{LPM,i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	98.70	5.6970	27.67	3.5922	39.43	0.0006
2	99.10	6.2263	24.99	3.5935	37.64	0.0007
3	103.70	6.2902	24.66	3.5945	39.74	0.0006

The Newcomb (1961) Fraternity data was also fit with an ERGM. Three model terms were used: mutual, Simmelian ties, and balance. A summary of the model fit is shown in Table 7. The AIC is 308.93 and the BIC is 319.75, which are more favorable than similar variations of the ERGM.

Table 7. Fit Summary for Newcomb ERGM.

Model Parameter	Coefficient	Standard Error	MCMC S.E.	p-value
Mutual	-1.5745	0.2304	0.0070	0.0000
Simmelian Ties	0.6581	0.0006	0.0001	0.0000
Balance	0.2333	0.0364	0.0010	0.0000

The Hamming distances from each of the fourteen empirical data sets to each of the ERGM simulated networks and each of the LPM simulated networks were calculated. The mean and standard deviation of these Hamming distances are displayed in Table 8. A two-sample t-test for each empirical data set illustrates that the networks simulated using the LPM have a smaller average hamming distance to the empirical data sets than the networks simulated using the ERGM. This indicates that the LPM models the Newcomb fraternity data more accurately than the ERGM model.

Table 8. Newcomb Data Hamming Distances and T-test for ERGM and LPM.

Time Period	$\mu_{ERGM,i}$	ERGM Hamming Distance Standard Deviation	$\mu_{LPM,i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	139.7	8.3938	91.9	5.1913	18.0147	0.0353
2	138.9	8.1847	75.1	5.2128	24.6573	0.0258
3	137.3	8.2872	48.3	5.2226	33.9732	0.0187
4	135.5	9.3363	49.7	5.2340	29.0460	0.0219
5	134.1	8.9870	50.1	5.2319	29.5558	0.0215
6	136.3	8.5251	45.5	5.2440	33.6983	0.0189
7	133.9	9.0609	47.3	5.2397	30.2202	0.0211
8	134.1	7.2946	51.9	5.2591	35.6377	0.0179
10	133.7	5.1865	64.2	5.2223	42.3990	0.0000
11	132.7	6.0562	53.4	5.2074	41.4119	0.0006
12	136.3	8.4466	51.1	5.2147	31.8930	0.0200
13	134.9	9.0117	46.6	5.2311	30.9989	0.0205
14	133.9	5.4457	46.1	5.2230	50.9574	0.0000
15	133.1	5.7242	47.2	5.2378	47.4518	0.0004

The LPM is further investigated using the Fort Leavenworth data. ERGM's with only a single term were found to be degenerate for several common parameter choices; therefore, they are not included in the analysis of this section. For both of the Fort Leavenworth data sets, the Hamming distance between the simulated LPM networks and each empirical network, $\text{Dist}_{\text{LPM},i,j} = \text{Hamming}(\text{N}i,\text{L}j)$, was compared to the Hamming distance between each empirical network to the other empirical networks within the data set, $\text{Dist}_{\text{empirical},i,j} = \text{Hamming}(\text{N}i,\text{N}j)$, where $i \neq j$. Two-sample t-tests were used to determine if there was a significant difference in mean Hamming distance between the empirical networks and the LPM. The t-tests were properly adjusted for heteroscedasticity and unequal sample sizes. Table 9 displays the Hamming distances and the results of the two-sample t-tests for the 2005 Fort Leavenworth data, and Table 10 displays this information for the 2007 Fort Leavenworth data. In all cases the Hamming distance is less for the LPM. The low p-values show a statistically significant difference in mean Hamming distance of the empirical to empirical comparison versus the LPM to empirical comparison. Additionally, since $\mu_{emperical,i} - \mu_{LPM,i} > 0$ it is shown that the simulated LPM networks have, on average, less Hamming distance from each of the empirical data sets than the empirical data sets have from each other. This means that networks generated using the LPM are closer to the original data than the observed empirical networks are to each other. While the t-tests for 2005 Fort Leavenworth time periods 6, 8, and 9 are only marginally significant, they have the same positive trend as the other 14 empirical networks in the 2005 and 2007 data sets.

Table 9. 2005 Fort Leavenworth Data Hamming Distances and T-test for LPM.

Time Period	$\mu_{\text{empirical},i}$	Empirical Hamming Distance Standard Deviation	$\mu_{\text{LPM},i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	1445.000	84.774	1284.338	23.747	3.467	0.001
2	1394.750	67.487	1239.647	23.703	3.765	0.000
3	1296.125	85.436	1151.946	23.671	3.287	0.001
4	1315.875	153.533	1169.665	23.718	2.421	0.015
5	1191.250	112.324	1058.990	23.667	2.732	0.006
6	1204.875	207.944	1071.116	23.623	1.912	0.056
7	1167.375	190.431	1037.713	23.695	1.980	0.048
8	1159.625	204.465	1030.815	23.732	1.888	0.059
9	1170.125	195.266	1040.142	23.618	1.953	0.051

Table 10. 2007 Fort Leavenworth Data Hamming Distances and T-test for LPM.

Time Period	$\mu_{empirical,i}$	Empirical Hamming Distance Standard Deviation	$\mu_{LPM,i}$	LPM Hamming Distance Standard Deviation	T_i t-test	p-value
1	409.286	38.560	358.094	12.775	3.755	0.00
2	365.857	18.298	320.097	12.739	7.073	0.00
3	365.857	29.043	320.164	12.793	4.450	0.00
4	377.857	38.247	330.674	12.773	3.489	0.00
5	375.286	36.100	328.377	12.796	3.675	0.00
6	349.857	38.159	306.078	12.785	3.245	0.00
7	373.857	48.451	327.073	12.826	2.731	0.01
8	362.429	55.635	317.151	12.775	2.302	0.02

3.6 Discussion

The LPM has been used to model longitudinal social network data for four different data sets. In those data sets, the LPM generates simulated networks that are more like the original data than networks generated using the ERGM. In addition, it is generally the case that the networks generated using the LPM are more similar to the original data than any prior time period. The LPM avoids issues of model degeneracy due to its formulation. The probability of link occurrence is based on the historic presence of links and does not use a Markov assumption or over specify a statistical model. For these reasons, the LPM provides an alternative method for modeling and conducting longitudinal social network analysis.

Monte Carlo simulations can be generated using the LPM. Each cell, $a_{ij.}$, in the LPM can be compared to a uniform (0,1) random variable to determine the presence of a link in a simulated adjacency matrix. As demonstrated earlier, these simulated adjacency matrices are very similar to the empirical data as demonstrated by the low Hamming distance between simulated networks and empirical networks. Statistical distributions can then be fit to any social network measures calculated on the simulated networks. These statistical distributions can then be used for inference using traditional statistical methods.

The LPM cannot be used in place of the ERGM in all situations, however. Multiple networks are required to estimate the LPM for a given empirical data set. The ERGM on the other hand, can be estimated from a single observed network. The approach to adding and removing nodes is different for the ERGM and LPM. For the LPM, a missing node would be included in the model with a 0 recorded for all column and row entries of the missing node. Finally, the LPM is formulated based on the assumption that there are fixed probability structures under-laying social networks that do

not change significantly over time. The observed social networks based on the LPM will fluctuate between time periods, but the general patterns of connections remain the same. Table 11 illustrates some differences and similarities between LPM and ERGM data requirements.

Table 11. Comparison of LPM and ERGM.

Data characteristics	LPM	ERGM
Link weighting	Dichotomous	Dichotomous
Number of links	No limit	Probability of degeneracy increases with number of links
Min. no. time period	2	1
Practical no. time period	5+	1
Assumed cause of stochasticity	Dynamic equilibrium	Evolves due to structural properties of the network.

The LPM has several advantages over the ERGM for longitudinal social network analysis; however the ERGM has advantages over the LPM for other types of analysis. Table 12 displays advantages and disadvantages of the LPM and ERGM models. The LPM requires multiple observed networks to estimate model parameters, where the ERGM can be estimated using a single observed network. At a minimum, two observed networks are required to estimate an LPM, however, in practice; the variance of the estimate is proportionate to $1/\sqrt{n}$, where n is the number of observed networks. I nominate five observed networks as a rule of thumb for fitting the LPM as most of the estimate variance is eliminated with this number. The LPM is more computationally efficient than the ERGM. The number of link probabilities for a network is quadratic with the number of nodes. The LPM estimates are then linear with the number of observed networks. The ERGM parameter estimates can be n^n with number of nodes for each term. Heuristics are often used to estimate ERGM model parameters. In addition, the ERGM has problems with model degeneracy as previously discussed. The LPM has been shown to provide a model that can be used to simulate data that is more similar to empirical data than data generated with ERGM simulations. An additional benefit for the LPM is the ability to use link probabilities as dependent variables in regression models for homophily. Homophily is an expression to describe the similarity between individuals in terms of certain attributes that the individuals have. In more complex models, the parameters of link probability densities can serve as dependent variables in homophily regression. Unfortunately, the LPM does not provide any explanation of likely structural causes for the stochastic behavior of networks. Significant terms in an ERGM can be interpreted as the underlying mechanism for network evolution over time. It may be possible to develop similar explanations of behavior through future research in homophily regression using the LPM. Further research is needed on both the ERGM and the LPM to illuminate strengths and limitations. In the interim, there is strong evidence to suggest the use of the LPM whenever degeneracy is a problem among ERGM's, or when the goal is to estimate the normal behavior of a social group that is in dynamic equilibrium.

Table 12. Advantages and Disadvantages of LPM and ERGM.

Considerations	LPM	ERGM
Required no. of observed networks	***Disadvantage***: The LPM requires multiple observed networks to estimate the link probability of a network based on historic frequency of occurrence.	***Advantage***: The ERGM requires only a single network
Computational efficiency	***Advantage***: The computational speed is quadratic with the number of nodes in the network.	***Disadvantage***: The computational speed is n^n which requires heuristic approximations of model parameters.
Model quality	***Advantage***: Stable and consistent model estimates.	***Disadvantage***: Prone to degenerate models.
Accuracy to real data	***Advantage***: Shown to more closely resemble empirical data as measured by Hamming distance.	***Disadvantage***: Has not been shown to consistently model empirical data accurately as measured by Hamming distance.
Explanation of social dynamics	***Disadvantage***: Does not attempt to explain underlying social dynamics of the group or organization.	***Advantage***: Model terms can be interpreted as underlying mechanisms for social dynamics within the modeled group or organization.

Another important area for future research is network periodicity. Intuitively, social networks are subject to periodic trends. An average person's communication patterns may be different during the week, while they are at work, than during the weekend, when they are at home with their family. Future research will hopefully expand both the LPM and ERGM to handle periodic trends in longitudinal data. It will be interesting to compare the performance of the LPM and ERGM for modeling time dependent longitudinal social network data sets.

This Chapter has introduced the Link Probability Model (LPM) for longitudinal social network analysis. The primary strength of the LPM is its ability to accurately model longitudinal network behavior with better goodness of fit than competing models. The LPM also avoids issues of model degeneracy due to the method of its construction. Finally, the LPM is more computationally efficient than the ERGM for both estimation and simulation. Using the LPM, accurate simulation of longitudinal social network data can be performed. This opens the door for researchers to explore an entirely new approach for inference on social networks.

Within this thesis, I will explore the performance of social network change detection by using the LPM to instantiate a multi-agent simulation model, *CONSTRUCT*. The multi-agent simulation model improves on the realism of the LPM by introducing variables such as homophily, socio-demographics, and proximity to modify the LPM at each time step. The multi-agent simulation thereby increases the relational dependence of link occurrence based on established social theories. For a detailed explanation of the *CONSTRUCT* model, refer to Appendix B. Using multi-agent simulation to generate virtual longitudinal social networks, advances can be made in detecting anomalies in network behavior as well as temporal change.

4 Detecting Changes in Social Networks

Social network change detection (SNCD) represents an exciting new area of research. It combines the area of statistical process control and social network analysis. The combination of these two disciplines is likely to produce significant insight into organizational behavior and social dynamics. Immediate applications to counter terrorism and organizational behavior are obvious.

Much research has been focused in the area of longitudinal social networks (Sampson, 1969; Newcomb, 1961; Romney et al, 1989; Sanil, Banks, and Carley, 1995; Snijders, 1990, 2007; Frank, 1991; Huisman and Snijders, 2003; Johnson et al, 2003; McCulloh et al, 2007a, 2007b). Wasserman et al. (2007) state that, "The analysis of social networks over time has long been recognized as something of a Holy Grail for network researchers." Doreian and Stokman (1997) produced a seminal text on the evolution of social networks. In their book they identified as a minimum, 47 articles published in *Social Networks* that included some use of time, as of 1994. They also noted several articles that used over time data, but discarded the temporal component, presumably because the authors lacked the methods to properly analyze such data. An excellent example of this is the Newcomb (1961) fraternity data, which has been widely used throughout the social network literature. More recently, this data has been analyzed with its' temporal component (Doreian et al., 1997; Krackhardt, 1998; Baller, et al. 2008). Methods for the analysis of over time network data has actually been present in the social sciences literature for quite some time (Katz and Proctor, 1959; Holland and Leinhardt, 1977; Wasserman, 1977; Wasserman and Iacobuccci 1988; Frank, 1991). Continuous time Markov chains for modeling longitudinal networks were proposed as early as 1977 by Holland and Leinhardt and by Wasserman. Their early work has been significantly improved upon (Wasserman, 1979; 1980; Leenders, 1995; Snijders and van Duijn, 1997; Snijders, 2001; Robins and Pattison, 2001) and Markovian methods of longitudinal analysis have even been automated in a popular social network analysis software package SIENA. A related body of research focuses on the evolution of social networks (Dorien, 1983; Carley, 1991; Carley, 1995; Carley 1997; Dorien and Stokman, 1997) to include three special issues in the Journal of Mathematical Sociology (JMS Vol 21, 1-2; JMS Vol 25, 1; JMS Vol 27, 1). Others have focused on statistical models of network change (Feld, 1997; Sanil, Banks, and Carley, 1995; Snijders, 1990, 1996; Van de Bunt et al, 1999; Snijders and Van Duijn, 1997). Robins and Pattison (2001, 2007) have used dependence graphs to account for dependence in over-time network evolution. We can clearly see that the development of longitudinal network analysis methods is a well established problem in the field of social networks.

Longitudinal network analysis is not synonymous with network evolution. Doreian and Stokman (1997) are careful to draw the distinction between "network dynamics" and "evolution of networks". They describe "network dynamics" as a more general statement of the network over time. They see the "evolution of networks" as having a stricter meaning that assumes we can understand network change "via some *understood* process." To further illustrate their point, two forms of change are clearly not evolutionary (Nelson, 1995). *Shock* occurs when future events are independent from previous events. This implies that no inference can be drawn from the present model

about the future. Nelson (1995) refers to this as random change, however, I have chosen to use the term shock to distinguish this exogenous change from random noise that may exist in the over-time signal. *Determined change* is also independent of the underlying process. An example of determined changed is more common in equilibrium theory, where the stable state of a system is independent of its' initial state as well as the process to reach the equilibrium (Martin and Sunley, 2006). In a social network context, the individual goals and motives of an individual, among other factors may drive the network to evolve. It is also possible for a shock to impact the network. For example, a military platoon consisting of 20-30 soldiers can experience evolutionary change as individuals interact, share beliefs and experiences. A shock might occur in the form of an enemy attack. During the attack there is something fundamentally different about the relationships among the soldiers. There is nothing about the individual interactions that could predict this change caused by an exogenous source. Shock can occur for many reasons. A shortage of economic resources could lead to job lay-offs which will significantly affect the social network, regardless of evolutionary effects. These are of course drastic changes, presented here to illustrate network shock. It is also possible to have a smaller shock, such as when a new person joins a social group, a company finds new access to less expensive resources, or a group member finds a better way of accomplishing required tasks.

Two other network dynamic behaviors are also possible. A shock may initiate evolutionary behavior. In our military example, it is possible that the heroic or cowardly actions of individuals in the platoon may affect the way other platoon members see them, thereby affecting the interaction among agents in the network and initiating network evolution. I refer to this type of change as a *Mutation.* The final network dynamic behavior that I propose is *Stability.* Stability occurs when the underlying relationship between agents in a network remains the same. It is possible that observed networks may contain error (Killworth and Bernard, 1976; Bernard and Killworth, 1977). If the network is stable, then changes in the network over time are due to observation error alone.

It is important to delineate the difference between stability, shock, determined change, evolutionary change, and mutation if we are to understand network dynamics and any underlying processes governing network behavior. A first step toward this problem is to statistically determine that an organization has changed over time. For example, Johnson et al. (2003) studied people wintering over at the South Pole. There were three similar groups corresponding to three different years. A whole-network survey design was used to collect social network data once per month for eight months for each of the three groups. Johnson studied evolutionary change on the social networks of the three groups. Theoretically, these similar groups should exhibit similar evolutionary behavior. In one of the groups, there was a shock that involved the "disappearance" of an expressive leader "due in part to harassment by a marginalized crewmember." This shock significantly affected the evolutionary behavior of the network. This behavior was only apparent as a result of the similarity between the three groups and the large magnitude of the difference in network behavior, which enabled Johnson to determine the significant cause of this difference. In practice, this type of similarity among groups may

be rare. SNCD offers a method to identify statistically significant change in network behavior in real-time, and to identify a likely change point of when the change occurred. This change point will allow a social scientist to identify potential causes of change, such as the disappearance of the crew member, and isolate that random change from evolutionary change.

My approach for detecting changes in longitudinal networks proposes a technique to rapidly detect the presence of a network change in real-time. I am not predicting a future change, but rather rapidly identifying that a change has occurred; and then providing a statistically sound indication of when that change was likely to have occurred. Rapid detection and identification of change is important for two key reasons. First, it allows a social scientist investigating organizational change to respond quickly to organizational change, facilitating the change if it is positive, and mitigating the effects of negative change on the organization. For example, ideas and policies are discussed and communicated within a network of people, long before organizational implementation. Sometimes, individual politics (network evolution) can prevent the implementation of good ideas (Rogers, 2003). Rapid detection of organizational change may cause a manager to investigate the presence of good initiatives and see them through to implementation. On the other hand, terrorist organizations will begin planning their attacks, long before they are actually carried out. Rapid change detection could alert military intelligence analysts to the shift in planning activities prior to the attack occurring.

The second key reason that rapid change detection is important is that it limits the scope of explanation for network change. A sound statistical estimate of when a network change occurred can help a social scientist identify potential shocks and thereby isolate evolutionary change for investigation. Determining the likely time of change in a network helps us understand where to look for fundamental conditions that cause groups to transform themselves. If we as social scientists could monitor networks in a daily or weekly basis, we could open a new line of research within longitudinal network analysis.

SNCD is essentially a statistical approach for detecting small persistent changes in organizational behavior over time. Organizations are not static, and over time their structure, composition, and patterns of communication may change. These changes may occur quickly, such as when a corporation restructures, but they often happen gradually, as the organization responds to environmental pressures, or individual roles expand or contract. Often, these gradual changes reflect a fundamental qualitative shift in an organization, and may precede other indicators of change. It is important to note, however, that a certain degree of change is expected in the normal course of an unchanging organization, reflecting normal day-to-day variability. The challenge of Social Network Change Detection is whether metrics can be developed to detect signals of meaningful change in social networks in a background of normal variability.

This chapter will introduce an application of statistical process control to detect change in longitudinal network data. A brief background is provided on statistical

process control which is used extensively in manufacturing. Statistical process control is extended to social networks and demonstrated using multi-agent simulation.

4.1 Background

Longitudinal social network data is becoming increasingly more common. Over time network data can be readily obtained in a semi-autonomous fashion from the internet, blogs, and e-mail. Longitudinal network analysis is becoming increasingly relevant for the analysis of online citation networks, internet movie data, massive multi-player on-line games (MMPOG), patent data bases, and more.

Current methods of change detection in social networks, however, are limited. Hamming distance (Hamming, 1950) is often used in binary networks to measure the distance between two networks. Euclidean distance is similarly used for weighted networks (Wasserman and Faust, 1994). While these methods may be effective at quantifying a difference in static networks, they lack an underlying statistical distribution. This prevents an analyst from identifying a statistically significant change, as opposed to normal and spurious fluctuations in the network. The quadratic assignment procedure (QAP) and its regression counterpart MRQAP (Krackhardt, 1987, 1992) has been used to detect structural significance and compare networks in terms of their correlation. This is not the same as detecting a statistically significant change in the network over-time, since error would be propagated through multiple network comparisons. Markovian approaches to longitudinal network analysis such as SIENA are excellent methods for modeling evolutionary change and determining structural factors that affect network change, however, these models are not concerned with rapid detection of significant change. These models also assume an underlying statistical process within the network that drives change, and does not account for random changes or shocks to the network. These methods have been focused on the direction of change. None of these methods were created for the purpose of detecting a change. SNCD has the potential to improve a social scientist's ability to detect organizational change in the same way that Wald (1947) improved on Neyman and Pearson's (1933) most powerful test of simple hypotheses with the sequential probability ratio test.

SNCD is a process of monitoring networks to determine when significant changes to their organizational structure occur and what caused them. I propose that techniques from social network analysis, combined with those from statistical process control can be used to detect when significant changes occur in longitudinal network data. In application, it requires the use of statistical process control charts to detect changes in observable network measures. By taking measures of a network over time, a control chart can be used to signal when significant changes occur in the network. For those unfamiliar with statistical process control, it should be noted that the word "control" can be very misleading. In fact, nothing is controlled at all. Statistical process control is a collection of algorithms that monitor a stochastic process over time and rapidly detect statistically significant departures from typical behavior. Control charts refer to the individual algorithms used to monitor a process. The word "control" is derived from

their application in quality control. Quality engineers attempt to control production lines by monitoring them and investigating any statistical anomalies. Through investigation, they attempt to mitigate negative process behavior and continue any newly discovered process improvements. In our application of SNCD, I use statistical process control to monitor longitudinal social networks and detect any statistically significant departures from typical behavior that may correspond to a shock in the network. While the quality engineer uses this technique to "control" a manufacturing process, I envision that the social scientist will use it to gain insight in network dynamics.

There are many network measures that can be calculated from a given graph. Network measures can be calculated from the entire graph or for each individual node. The SNCD technique is applicable to any measure of the network. In this paper for exposition purposes I focus on graph level measures rather than node level measures. For example I use the average of the betweenness (Freeman, 1977) over all nodes in the graph each time period rather than the betweenness of a single node. I also illustrate SNCD using density (Coleman and Moré, 1983) and average closeness (Freeman, 1979). These are chosen because they are commonly used in the literature and represent a range of the types of measures available for change detection. Additional measures such as the maximum, minimum, and the standard deviation of the above node level measures are considered in a virtual experiment to explore limitations of the proposed method. A complete exploration of all social network measures and all possible types of changes to a network is certainly beyond the scope of this initial paper on the subject, however, I hope to have sufficiently illustrated the promise of this approach. Another concern with these measures is their normalization. In order to compare measures across different time periods, they must be normalized. For a steady sized group this should not be an issue, but in the case of an expanding or contracting group, issues arise as to whether results can be used across the different scales of group size. In other words, the network measures may change in different ways with respect to the current group size and thus provide inconsistent information about the group even absent of any shock within the group. For more detailed information on the standardization of network measures, see Bonacich, Oliver, and Snijders (1998). For this research, the Organizational Risk Analyzer (ORA) developed by Kathleen Carley at the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University is used to compute the average network measures from all group information (Carley, 2007).

4.2 Statistical Process Control

SPC is a technique used by quality engineers to monitor industrial processes. They use control charts to detect changes in an industrial process by taking periodic samples from the process, calculating a statistic based on some process metric, and comparing the statistic against a decision interval. If the statistic exceeds the decision interval, the "control chart" is said to "signal" that a change may have occurred in the process. Once a potential change has been "signaled", quality engineers investigate the process to determine if an actual change occurred, what the most likely time the change occurred was, and whether the process needs to be reset or improved to avoid financial

loss for the company. Control charts are usually optimized for their processes to increase their sensitivity for detecting changes, while minimizing the number of "false positives" – signals when no change has actually occurred in the process.

Three control chart schemes are investigated in this paper; the cumulative sum (CUSUM) (Page, 1961); the Exponentially Weighted Moving Average (Roberts, 1959); and the Scan Statistic (Fisher and MacKenzie, 1922; Naus, 1965; Priebe et al, 2005). The CUSUM will be the primary method considered and recommended for longitudinal network analysis. The other methods are presented here and applied to simulated networks in a virtual experiment to explore limitations of SNCD.

Cumulative Sum Control Chart

Page (1961) proposed the cumulative sum (CUSUM) control chart as an alternative to the $\overline{X}$ chart (Shewhart, 1927). The CUSUM control chart is derived from the sequential probability ratio test (SPRT) which was introduced as an improvement over the Neyman and Pearson most powerful test for a simple hypothesis.

Neyman and Pearson (1933) introduced the most powerful (minimum Type II error) test for a simple hypothesis-testing problem. Neyman and Pearson's test statistic is

$$\Lambda_t = \frac{\prod_{i=1}^{t} f(x_i;\mu_1)}{\prod_{i=1}^{t} f(x_i;\mu_0)}.$$

Neyman and Pearson showed that the most powerful test of H_o against H_1 is obtained by rejecting H_o if $\Lambda_t \geq K$ and concluding in favor of H_o if $\Lambda_t < K$, where K is determined by the level of significance, α. The level of significance is also known as the Type I error and is the probability that H_o is rejected when it is true.

Wald (1945, 1947) demonstrated that the Neyman and Pearson hypothesis testing method could be applied sequentially and could significantly reduce the number of samples required to reach a conclusion. Wald's sequential probability ratio test (SPRT) compares Λ_t to two constants A and B where $0 < B < A < \infty$. Observations are collected and examined one-at-a-time. After the t^{th} observation there are three possible outcomes. If $\Lambda_t < B$, then the test concludes in favor of H_o. If $\Lambda_t > A$, then H_o is rejected in favor of H_1. If $B \leq \Lambda_t \leq A$ then sampling will continue with observation $t + 1$.

The SPRT can be used to test $H_o : \mu = \mu_o$ against $H_1 : \mu = \mu_1$ for normal means. Without loss of generality we will assume that $\mu_1 > \mu_o$. Having observed t observations, the SPR is

$$\Lambda_t = \frac{\prod_{i=1}^{t}\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma}\right)^2\right)\right)}{\prod_{i=1}^{t}\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_i-\mu_0}{\sigma}\right)^2\right)\right)}.$$

This can be reduced algebraically to

$$\Lambda_t = \exp\left(\left(\frac{\mu_1-\mu_0}{\sigma^2}\right)\sum_{i=1}^{t}x_i + t\left(\frac{\mu_0^2-\mu_1^2}{2\sigma^2}\right)\right).$$

The sequential probability ratio, Λ_t, is compared to appropriate constants *A* and *B* as each new observation *t* is formed. Following observation *t*, the test concludes in favor of H_o if $\Lambda_t < B$. If $\Lambda_t > A$, then the test concludes in favor of H_1. If $B \leq \Lambda_t \leq A$, then sample $t + 1$ is obtained and a revised Λ_{t+1} is computed. This procedure continues until either $\Lambda_t < B$ or $\Lambda_t > A$. It is important to note that while I demonstrate this derivation using a normal distribution, any distribution can be used with SPRT.

In an SNCD application of the SPRT, μ_o is some property such as the average density, average transitivity, average balance, etc. of a typical network process and μ_1 is the value of the property when the network process has experienced a change. Since one would never conclude in favor of H_o that the network process is unchanged and stop all sampling, the procedure continues until it signals that the network process may have changed. This implementation of the SPRT procedure leads to the CUSUM control chart.

In an SNCD application of the SPRT, one would continue to monitor the process until $\Lambda_t > A$ when the procedure signals that the network process may have changed. Using our illustration with a normally distributed property, the SPRT leads to the following expression for detecting an increase network property. The procedure would signal when

$$\Lambda_t = \exp\left(\left(\frac{\mu_1-\mu_0}{\sigma^2}\right)\sum_{i=1}^{t}x_i + t\left(\frac{\mu_0^2-\mu_1^2}{2\sigma^2}\right)\right) > A.$$

This expression can be simplified by taking the natural logarithm of both sides of the inequality,

$$\left(\frac{\mu_1-\mu_0}{\sigma^2}\right)\sum_{i=1}^{t}x_i + t\left(\frac{\mu_0^2-\mu_1^2}{2\sigma^2}\right) > \log A.$$

This decision rule can be algebraically reduced to $\sum_{i=1}^{t} x_i - t\left(\frac{\mu_0 + \mu_1}{2}\right) > A'$, where

$$A' = \left(\frac{\sigma^2}{\mu_1 - \mu_0}\right)\log A.$$

By allowing $\mu_1 = \mu_0 + \delta\sigma_{\bar{x}}$, the procedure signals when

$$\sum_{i=1}^{t} x_i - t\left(\frac{\mu_0 + (\mu_0 + \delta)}{2}\right) = \sum_{i=1}^{t}\left(x_i - \mu_0 - \frac{\delta}{2}\right) > A'$$

where δ is the standardized difference in the network property under H_o and H_1. This decision rule can then be further simplified by using the cumulative statistic $C_t = \sum_{i=1}^{t}(Z_i - k)$, where $Z_i = (\bar{x}_i - \mu_0)/\sigma_{\bar{x}}$, and $k = \delta / 2$. The common choice of k is 0.5, which corresponds to a standardized magnitude of change in the network property of $\delta = 1$. Thus, observations are examined sequentially until $C_t > A'$.

The CUSUM control chart sequentially compares the statistic C_t against a decision interval A' until $C_t > A'$. Since one is not interested in concluding that the network process is unchanged, the cumulative statistic is

$$C_t^+ = \max\{0,\ Z_t - k + C_{t-1}^+\}.$$

If this rule was not implemented the control chart would require more observations of the network to signal if $C_t < 0$ at the time of network change. The statistic C_t^+ is compared to a constant, h^+. If $C_t^+ > h^+$, then the control chart signals that a change in the network might have occurred.

For $\delta < 0$, the SPRT similarly leads to the CUSUM procedure for detecting a decrease in the network property. In this case, $C_t^- = \max\{0, -Z_t - k + C_{t-1}^+\}$ is compared to a constant, h^-. If $C_t^- > h^-$, then the control chart also signals that network might have changed.

To monitor for both directions of network change, two one-sided control charts are employed. One chart is used for monitoring for increases in the monitored network property and the other is used for detecting decreases in the property. If the process remains in-control, $C_t^{\pm}$ will fluctuate around zero. If there is an increase in the network property, C_t^+ will tend to increase. Conversely, if there is a decrease in the network

property, then C_t^- will tend to increase. When $C_t^+ > h^+$ or $C_t^- > h^-$, the two one-sided CUSUM control chart scheme signals that the network may have changed.

The CUSUM control chart was selected for two reasons. First, this chart is well suited to detecting small changes in a process over time. In terms of a social network, this is a desired quality because one would not expect a social network to change dramatically between short time periods. By casual observation, one could conclude that a person's friends generally stay the same from week to week and not expect drastic changes in that social network. In addition, drastic changes in the network are normally quite obvious, but since the CUSUM is good at detecting slight changes it may be able to provide rapid detection of drastic changes, or reveal when more subtle changes have occurred. A second benefit of the CUSUM control chart is its built-in change point detection. In all cases, a change in the network precedes the detection of that change. The number of networks that must be observed before the change is detected varies based on the magnitude of change among other factors. After the control chart signals, the most likely change point is found by tracing the *C* statistic back to the last time it was zero. This allows the time of the change in the network to be calculated quickly and easily. This allows the social scientist to limit the scope of investigation for causes of network change. This feature is particularly useful for longitudinal network analysis and for studying the evolution of networks.

Exponentially Weighted Moving Average Control Chart

The exponentially weighted moving average (EWMA) control chart was introduced by Roberts (1959) for monitoring changes in the mean of a process. The EWMA associated with subgroup *t* is $w_t = \lambda \bar{x}_t + (1-\lambda) w_{t-1}$, where $0 < \lambda \leq 1$ is the weight assigned to the current subgroup average and $w_0 = \mu_0$. Common values of λ are $0.1 \leq \lambda \leq 0.3$. Having observed a total of *T* subgroups, the statistic w_T is plotted against the decision interval $\mu_0 \pm L\sigma_{\bar{x}}\left(\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2T}\right]\right)^{1/2}$, where *L* is a constant that scales the width of the decision interval.

Lucas and Saccucci (1987, 1990) investigated the impact of different combinations of *L* and λ on the average number of observations before the EWMA signals a change. The combinations that were investigated were chosen such that the false positive rate for each chart was the same. They found that EWMA charts with small values of λ perform well at detecting small changes in a process mean. Conversely, EWMA charts with large values of λ perform well at detecting large changes in a process mean. Hunter (1986) and Montgomery (1996) investigated the performance of the EWMA chart and concluded that it is similar to the performance of the CUSUM chart. In addition, the EWMA is a time series approach for SPC. Therefore, the EWMA seems a good candidate for comparison to the CUSUM.

Scan Statistic

Scan statistics (Fisher and Mackenzie, 1922; Naus, 1965; Priebe, et. al., 2005), also known as moving window analysis, investigates a random field for the presence of a local signal. A small window of observations is used to calculate a local statistic. In this paper a window size of 7 observations proceeding the current time period is used, and the window mean is used for the local statistic. If the statistic exceeds a decision interval, then inference can be made that a change in the network may have occurred.

4.3 Data

Simulated data is used in order to inject an organizational change at a defined point in time. The CUSUM can then be evaluated on its' ability to identify that change. In real-world data, there are often many changes facing an organization and identifying one specific cause of change can be subjective or questionable. With simulated data, SNCD can be explored in a more controlled series of virtual experiments. For this initial investigation, I use a multi-agent simulation of a 100 node network, using the Construct simulation model (Schreiber and Carley, 2004) set in the context of a U.S. Infantry military organization. Military units have a formal hierarchical chain of command as well as informal leaders and social relationships that extend beyond formally designated units. Both the formal and informal networks are used to share situational awareness, experience, skill development, and resources. Isolation of certain individuals or subordinate elements within a unit, due to radio failure, enemy attack, or poor coordination, can cause serious impacts to the unit's performance. The basic military structure that was simulated was an infantry training model. This is the most basic US military unit and is used for training soldiers and officers across the US Army Training and Doctrine Command (HQ, Dept of the Army, 1992). Within this model, soldiers are organized into four man teams. Two teams and a squad leader form a 9 man squad. Three squads and a three person headquarters form a 30 man platoon. Three platoons and a 10 person command post form a company. Each soldier is trained in various skills that are distributed throughout the organization. Each team for example will have an automatic gunner, a grenadier and two riflemen. One member on a team will also be trained as a medic, another in demolitions, and two will be able to search enemy prisoners of war. Each soldier possesses individual skill in stealth, situational awareness, physical fitness, intelligence, military rank, and motivation. Homophily in these individual skills create stronger bonds between members of a unit which will increase their probability of communication. Organizational proximity will also affect communication, with individuals in the same sub-unit being more likely to communicate. The objective of the simulation will be to model communication within the military unit.

The simulation was run with all agents present for the first 30 time periods. At time 30, some type of change was imposed on the network, isolating some of the agents, thereby simulating radio failure or enemy attack. Figures 5 and 6 show example snapshots of the simulated network before and after the change.

Figure 5. Simulation Before Change.

Figure 6. Simulation After Change.

The simulation was replicated 1000 times to obtain estimates of the average time to detect change as well as the variance.

4.4 Method

Social network change detection algorithms are implemented in much the same way a control chart is implemented in a manufacturing process. Three different graph measures are used for change detection for the sake of illustrating the proposed method. SNCD can be applied to any node or graph measure over time. The graph measures for density, average closeness, and average betweenness centrality are calculated for several consecutive time-periods of the social network. The mean and variance for the measures of the network are calculated by taking a sample average and sample variance from networks that are assumed to be "typical". At least two networks are required to estimate these values, however, more networks will allow a more accurate estimate of the mean and variance of the "typical" network measure. The subsequent, successive social network measures are then used to calculate the CUSUM's C^+ and C^- statistics. These are then compared to a decision interval to determine when or if the control chart signals a change in the mean of the monitored network measure. Upon receiving a signal, the change point is calculated by tracing the signaling C^+ or C^- statistic back to the last time period it was zero. In order to continue running the control chart after a signal, the mean and variance are recalculated after the network measures have stabilized following the change.

The suspected time periods when the network appears to be significantly changing can be estimated using the CUSUM statistic. The network can then be studied in depth across these time periods using a wide variety of network measures to determine the extent of changes to the network structure. Further study can also be directed towards determining changes in the environment in which the network operates during those time-periods.

This methodology is demonstrated on three real-world data sets and explored in more detail through simulation. The real-world data sets are used to illustrate practical application of the approach. In addition, the real-world data sets show relevance of this new analysis technique. The virtual experiments are performed on simulated data to explore some of the limitations of SNCD.

Virtual Experiment

A virtual experiment is conducted using the *Construct* Infantry Model to provide a realistic data set for evaluating SNCD methods. Three different size infantry units (squad, platoon, and company) are simulated for 500 time periods. In these units, four changes are introduced. This creates 9 independent data sets that can be used to evaluate SNCD performance. Three of the changes are not feasible for the squad size element. The four network changes correspond to common military communication problems that might affect an infantry unit.

The first type of network change is the isolation of the Headquarters section. For a squad, this is simply the squad leader. For a platoon, this consists of the Platoon Leader, Platoon Sergeant, and the Radio Telephone Operator (RTO). For a Company, this includes the 10 person command post, also known as the headquarters element. A military headquarters is most often isolated from the rest of the unit as a result of radio failure or a deliberate attack from enemy forces. This is perhaps one of the most significant changes that commonly happen in a military situation, as it requires a rapid and efficient transfer of command and control, as the formal hierarchy is significantly adjusted. In the simulation, this is modeled by isolating the Headquarters section beginning at time period 30. These individuals remain isolated for the remainder of the simulation. Network measures are calculated on the organization for all time periods.

Another significant change in a military organization is the loss of a subordinate element. A subordinate element might be lost as a result of a task organization change, radio failure, or enemy attack. This change is not modeled for the infantry squad, since this would mean losing half of the organization. For the Platoon, this change is modeled by isolating a squad at time period 30 for the remainder of the simulation. For the Company, this is also modeled by isolating a squad at time period 30 for the remainder of the simulation. While it is conceivable to isolate any number of individuals in the simulation, these changes are used to demonstrate the performance of the SNCD methods. Perhaps SNCD methods that have similar performance could be evaluated under greater conditions of change in a future paper. For now, it is beyond the scope of this paper to exhaustively address all conceivable types of network change.

A similar change is the addition of a new subordinate element. This is usually a result of a task organization change. This is modeled by adding a squad in both the Company and Platoon level models. It is not modeled for a squad, because squad organizations are not usually capable of managing an additional subordinate element. Again, this simple change is used to evaluate SNCD and not meant to be an exhaustive comparison of different types of organizational change.

The final type of change simulated, is sporadic communication. Sporadic communication can be either deliberate, or unplanned. An example of deliberate sporadic communication is a reconnaissance operation, where radio power must be conserved and noise discipline is important. An example of unplanned sporadic communication is radio failure. This is modeled in the simulation by introducing a squad from time period 30 to time period 40. Network measures will be recorded throughout the simulation. This change is only modeled for the Platoon and Company level simulations.

The social network measures listed in Table 13 are measured for every simulated network. Table 14 illustrates the combinations of the virtual experiment. The outputs of the simulation are the graph level measures recorded for each simulated time step. Different SNCD methods are then used to identify possible changes in the network over time.

Table 13. Social Network Measures.

Average Betweenness	Standard Deviation of Closeness
Maximum Betweenness	Average Eigenvector Centrality
Standard Deviation of Betweenness	Maximum Eigenvector Centrality
Average Closeness	Minimum Eigenvector Centrality
Maximum Closeness	Standard Deviation of Eigenvector

Table 14. Virtual Experiment.

Variable	Number	Values
Network Size	3	9, 30, 100
Type of Change in Network		
Isolation of leadership	2	Isolated headquarters after 30 time periods
Sporadic communication (Reconnaissance)	2	Initially absent, present for 10 time periods, then absent for remainder of simulation (omitted for squad)
Loss of subordinate unit	2	Removal of the immediate subordinate unit after 30 time periods (omitted for squad)
Gain an attached unit	2	Addition of a squad after 30 time periods. (omitted for squad)
Cells	18	3 Network sizes x 4 Changes x 2 Levels – Squad omissions
Replications	25	
Independent Runs	450	

4.5 Results

Using the social simulation program, *Construct* (Carley, 1990; Carley 1995; Schrieber and Carley, 2004), the performance of SNCD was explored through simulation. A variety of changes are introduced to the network at a known point. The Cumulative Sum (CUSUM), Exponentially Weighted Moving Average (EWMA), and Scan Statistic, statistical process control charts are applied to several social network graph level measures taken on the network at each time step. The number of time steps between the actual change and the time that an SNCD method "signals" a change will be recorded as the Detection Length. The Average Detection Length (ADL) over multiple independently seeded runs is then a measure of the SNCD method's performance. The ADL will be compared for different changes and different SNCD parameters.

Isolation of Headquarters

Investigating the isolation of the headquarters element in three different organizations will provide insight into how the network size affects the performance of change detection measures. In each organization, 30 man platoon, 100 man company, and 9 man squad; 10% of the network was removed. In a sense, the magnitude of change is the same, however, the network size is different.

The isolation of the platoon headquarters is modeled by removing the three headquarters members at time period 30 for the duration of the simulation. Social network measures are recorded for all time periods. Table 15 displays the ADL performance of the SNCD methods. It can be seen that the average of the betweenness is a better measure to use for SNCD than either the maximum or the standard deviation of betweenness. This is generally true for all magnitudes of change and sizes of organization investigated. For the closeness measure, both the maximum closeness and average closeness generally outperform the standard deviation of closeness. However, for an EWMA with $r = 0.3$, the maximum closeness measure has relatively poor performance. This might suggest that the average closeness measure is a more robust measure of change detection. In a single variant, non-network application of the EWMA, the parameter, r, makes the control chart more or less sensitive to a particular magnitude of change (Lucas and Saccucci, 1990; McCulloh, 2004). It is reasonable to consider that for the isolation of a platoon headquarters, the maximum closeness EWMA with $r \leq 0.2$ is sensitive to detecting the change, yet the maximum closeness EWMA with $r \geq 0.3$ is less sensitive. This will be explored with other magnitudes and types of changes throughout the paper. For eigenvector centrality, the maximum eigenvector centrality and the standard deviation of eigenvector centrality appear to be more sensitive measures of change detection than the average or minimum of the eigenvector centrality. It also appears that the eigenvector centrality measures dominate all other measures for performance in this case.

Table 15. ADL Performance of SNCD on Isolation of Platoon Headquarters.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	9.32	8.24	10.16	11.52	6.76
Maximum Betweenness	14.36	14.72	15.72	17.08	13.24
Std Dev. Betweenness	16.44	16.24	16.92	18.52	15.24
Average Closeness	10.68	9.08	13.60	17.52	10.48
Maximum Closeness	8.76	6.00	10.60	37.96	8.64
Std Deviation Closeness	34.48	34.72	34.52	35.68	27.08
Average Eigenvector	31.28	31.28	31.28	31.28	24.00
Minimum Eigenvector	14.36	14.36	14.28	15.56	14.88
Maximum Eigenvector	5.24	5.40	5.80	7.52	4.00
Std. Dev Eigenvector	5.92	4.88	6.40	6.96	3.64

Statistical process control is a powerful statistical method for detecting the change. Figures 7 and 8 show the average betweenness score for a baseline simulation run (no change) and one of the simulation runs with the headquarters isolated. The difference between the figures is subtle. The difference is only apparent, because there are many observations of the network following the change. If there were only a few observed networks following the change, it would be more difficult to detect the network change. Figures 9 and 10 show the CUSUM statistic value for the baseline simulation run and the simulation run with the headquarters isolated. The dramatic difference in the plots can clearly be seen, paying attention to the values on the y-axis of the plots.

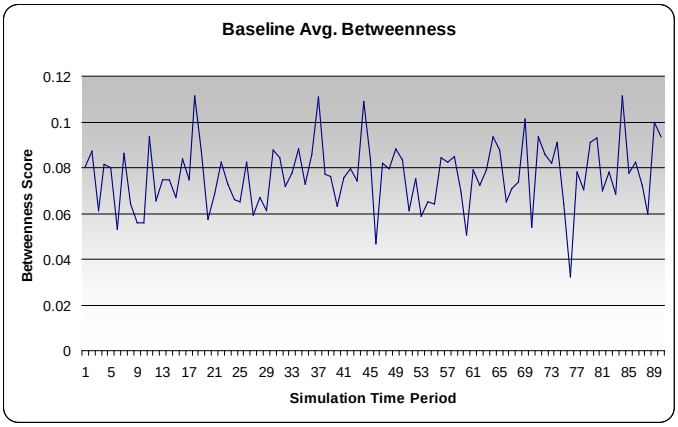


Figure 7. Baseline Betweenness Score.

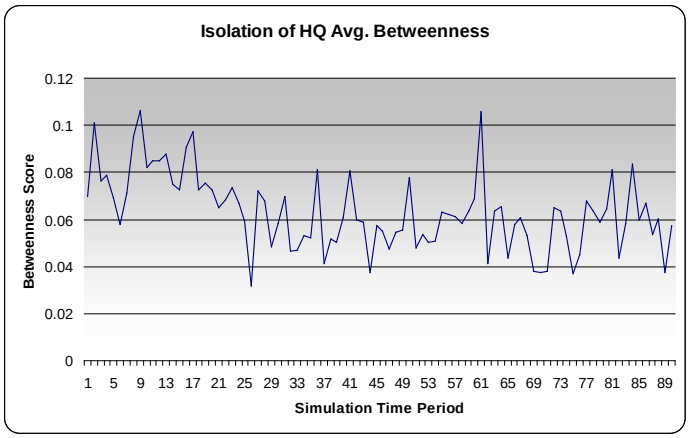


Figure 8. Isolation of HQ Betweenness Score.

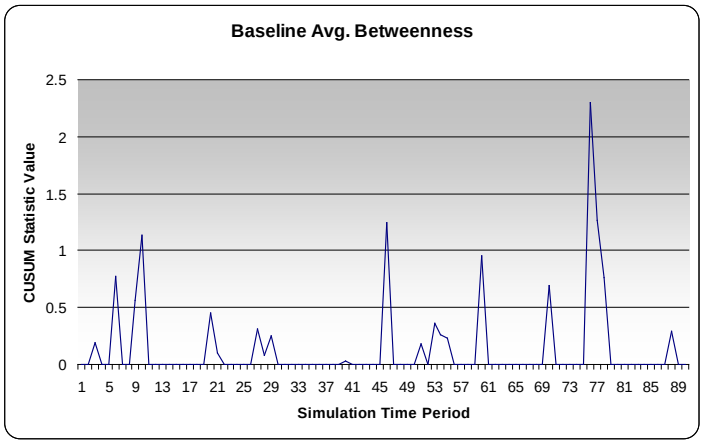


Figure 9. Baseline CUSUM Statistic Value.

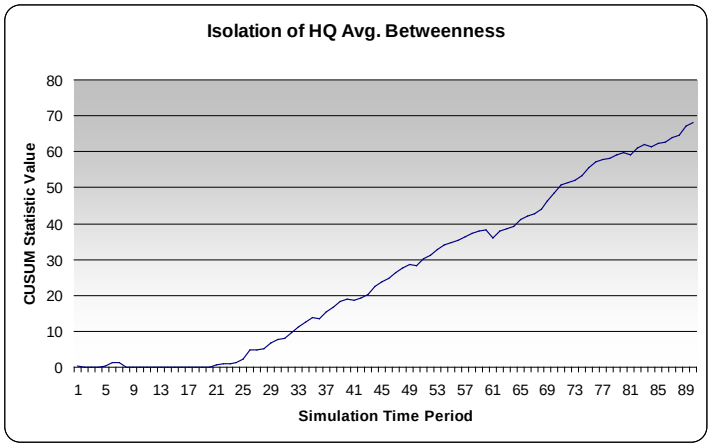


Figure 10. Isolation of HQ CUSUM Statistic Value.

The plot in Figure 10 clearly shows a sharp and sudden increase beginning at time period 30, which is when the isolation of the HQ element occurs. There is a similar

performance for other types of change imposed on the network, and other SNCD schemes that are used. The CUSUM is simply used to illustrate the power of the general change detection approach. Other magnitudes and types of change will be compared by simply reporting the ADL from when a change occurs until the SNCD scheme signals.

The isolation of the Company Headquarters was modeled by removing the 10 soldier headquarters section at time 30 for the remainder of the simulation. This is very similar to the platoon example, in that 10% of the organization is removed. Social network measures are again recorded for all time periods. Table 16 displays the ADL performance of each of the SNCD methods applied to the 100 node network. Again, it can be seen that the average of the betweenness is a more effective measure of change detection than the maximum or the standard deviation of betweenness. The performance of the closeness measures behave as they did in the case of platoon headquarters isolation. In this case, the maximum eigenvector centrality does not appear to be as effective of a measure for detecting change as does other measures. However, the standard deviation of eigenvector centrality still dominates all other measures for change detection performance.

Table 16. ADL Performance of SNCD on Isolation of Company Headquarters.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	11.16	11.08	10.20	13.48	6.96
Maximum Betweenness	17.32	17.76	18.20	20.12	13.72
Std Dev.Betweenness	18.08	19.40	20.88	22.52	17.36
Average Closeness	11.16	9.44	12.52	15.64	9.40
Maximum Closeness	10.44	9.72	12.64	51.76	9.60
Std Deviation Closeness	41.88	39.48	42.20	43.44	40.76
Average Eigenvecto	35.84	36.72	34.84	34.84	29.24
Minimum Eigenvector	16.00	17.96	17.88	16.76	13.60
Maximum Eigenvector	26.40	30.76	29.64	29.24	25.44
Std. Dev Eigenvector	10.40	10.72	9.36	9.48	6.44

The isolation of squad leadership was modeled by removing the squad leader at time 30 for the remainder of the simulation. This is also similar in that 11% of the organization is isolated. Table 17 shows the SNCD performance at the squad level, 9 node network. It is not clear that certain measures perform better than others for change detection in the 9 node network. It appears that the measures of average betweenness, average closeness, and the standard deviation of eigenvector centrality become better measures of network change as the size of the network increases. However, they do not necessarily perform worse on a small network. While an extensive study of the sensitivity of each measure to the network size is beyond the scope of this paper, it holds the promise of fruitful future research.

Table 17. ADL Performance of SNCD on Isolation of Squad Leader.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	16.12	15.76	16.32	17.92	12.32
Maximum Betweenness	16.64	17.40	19.52	18.56	11.56
Std Dev.Betweenness	17.68	17.76	18.20	18.72	12.08
Average Closeness	15.16	15.84	16.48	15.60	11.72
Maximum Closeness	18.72	19.60	18.68	23.80	14.32
Std Deviation Closeness	16.20	16.08	15.52	16.24	12.88
Average Eigenvector	24.12	24.12	24.12	24.12	15.12
Minimum Eigenvector	17.84	18.48	17.04	18.08	12.36
Maximum Eigenvector	19.36	21.56	20.56	20.56	13.84
Std. Dev Eigenvector	17.08	18.72	18.36	17.44	12.36

Loss of Subordinate Element

The loss of a subordinate element provides insight into how the magnitude of change affects change detection performance. For the 30 man platoon and the 100 man company, a nine man squad is isolated. This represents 30% of the platoon and 9% of the company. This change is obviously not feasible for the nine man squad, since it would involve removal of the entire organization.

The infantry platoon had one squad removed from the simulation at time period 30, for the remainder of the simulation. Social network measures were recorded for each time period. The ADL for each measure is reported in Table 18. Again, it can be seen that the average of the betweenness outperforms other betweenness measures. The closeness measures perform as in previously investigated cases. The minimum eigenvector centrality outperforms the maximum eigenvector centrality for most of the SNCD schemes for this particular type and magnitude of change. The standard deviation of eigenvector centrality still outperforms other eigenvector centrality measures, however, it is no longer dominates all other measures.

Table 18. ADL Performance for Loss of Subordinate Element in a Platoon.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	6.96	6.00	8.68	12.16	8.12
Maximum Betweenness	9.52	7.44	11.12	13.24	7.80
Std Dev.Betweenness	9.16	7.40	9.48	12.72	6.84
Average Closeness	9.64	8.36	12.72	19.28	11.40
Maximum Closeness	9.32	9.16	12.36	31.56	9.52
Std Deviation Closeness	18.96	16.44	19.40	26.24	17.04
Average Eigenvector	29.36	29.36	29.36	29.36	20.60
Minimum Eigenvector	10.08	9.64	12.24	12.60	10.28
Maximum Eigenvector	11.72	12.04	11.88	20.60	10.84
Std. Dev Eigenvector	8.48	6.28	9.80	10.44	6.88

The Infantry Company also had one squad removed at time 30 for the remainder of the simulation. The results for the Company network are shown in Table 19. It generally takes longer to detect the changes in the Company network. This was also observed in the isolation of the headquarters. This implies that the size of the network could impact the speed of change detection. The average betweenness, average closeness, and standard deviation of eigenvector centrality appear to outperform other measures for change detection performance. The maximum closeness measure dominates other measures in all cases except for the EWMA with $r = 0.3$.

Table 19. ADL Performance for Loss of Subordinate Element in a Company.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	13.64	11.72	13.80	20.60	12.68
Maximum Betweenness	23.80	19.64	23.80	30.72	25.44
Std Dev.Betweenness	24.84	18.12	24.96	25.52	22.04
Average Closeness	9.72	7.4	13.44	14.96	9.80
Maximum Closeness	6.92	4.92	7.48	53.16	6.32
Std Deviation Closeness	45.44	47.92	47.96	50.88	43.68
Average Eigenvector	34.72	36.60	34.72	34.72	30.64
Minimum Eigenvector	18.68	19.96	19.64	23.88	18.32
Maximum Eigenvector	18.28	25.80	25.00	27.20	25.88
Std. Dev Eigenvector	9.52	9.92	11.88	15.32	8.72

Addition of New Subordinate Element

Another type of change is the addition of a new subordinate element. A squad is added to both the 30 man platoon and the 100 man company.

The infantry platoon had one squad that was not present initially, and added at time period 30. Social network measures were calculated for each time period. SNCD methods were applied to the data. Results are shown in Table 20. Although the speed of change detection is much faster for this type of change, the same performance trends are seen as before. For betweenness measures, the average outperforms the maximum or the standard deviation. The average closeness and maximum closeness measure perform well, however, the maximum closeness does not perform well with an EWMA $r = 0.3$ scheme. The standard deviation of eigenvector centrality almost completely dominates other measures.

Table 20. ADL Performance for Addition of Subordinate Element in a Platoon.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	1.60	1.52	1.68	1.72	1.00
Maximum Betweenness	2.32	2.16	2.20	2.00	1.00
Std Dev.Betweenness	2.36	2.36	2.40	2.24	1.00
Average Closeness	1.48	1.52	1.56	1.52	1.00
Maximum Closeness	1.24	1.28	1.20	5.00	1.00
Std Deviation Closeness	3.44	4.60	4.20	3.48	2.64
Average Eigenvector	31.76	31.76	31.76	31.76	25.56
Minimum Eigenvector	6.24	5.6	6.16	6.80	4.20
Maximum Eigenvector	4.52	4.88	4.80	4.80	3.56
Std. Dev Eigenvector	1.16	1.60	1.24	1.24	1.00

The company model had a squad added at time period 30 for the remainder of the simulation. Again the platoon level performance is better than the company level performance, shown in Table 21. The average betweenness, average closeness, and maximum closeness all perform well at detecting the change. Surprisingly, the standard deviation of eigenvector centrality is not an effective measure for this type and magnitude of change.

Table 21. ADL Performance for Addition of Subordinate Element in a Company.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	9.64	9.52	9.84	10.28	5.04
Maximum Betweenness	14.52	16.96	15.80	17.44	12.16
Std Dev.Betweenness	12.88	13.16	13.32	14.56	8.92
Average Closeness	5.32	5.8	5.36	5.24	1.44
Maximum Closeness	4.24	5.12	4.48	6.04	1.04
Std Deviation Closeness	10.40	18.52	12.96	12.32	10.00
Average Eigenvector	35.56	37.04	38.64	37.60	30.24
Minimum Eigenvector	38.16	39.32	38.04	40.84	36.40
Maximum Eigenvector	30.20	33.48	34.44	29.52	30.92
Std. Dev Eigenvector	33.88	33.72	37.80	44.48	33.96

Sporadic Communication

Sporadic communication was modeled with a squad communicating from time period 30 to time period 40 only. It can be seen in Table 22 that the performance of different measures is much more similar than in previous types of change. It is also interesting that all of the ADL values are greater than 10, which means that the change was detected after the organization returned to its original state. This might be a result of the SNCD statistic being moved closer to the decision interval from time period 30 to

time period 40. When the organization returned to its original state, the statistic is much closer to the decision interval than it was before the change occurred. Therefore, the statistic is much more likely to signal a false positive after the sporadic change than it is to detect an actual change. This increased sensitivity can therefore provide an alert that a sporadic change may have occurred.

Table 22. ADL Performance for Sporadic Communication.

	CUSUM $k = 0.5$	EWMA $r = 0.1$	EWMA $r = 0.2$	EWMA $r = 0.3$	Scan Statistic
Average Betweenness	15.08	14.20	16.12	17.56	17.76
Maximum Betweenness	15.24	16.52	16.88	18.24	17.84
Std Dev.Betweenness	14.28	14.80	16.04	17.40	17.48
Average Closeness	13.72	13.68	16.84	16.80	17.52
Maximum Closeness	12.44	12.16	15.32	18.32	17.20
Std Deviation Closeness	23.16	19.96	21.76	21.36	17.24
Average Eigenvector	24.32	24.32	24.32	24.32	18.84
Minimum Eigenvector	12.76	14.32	11.92	12.80	14.56
Maximum Eigenvector	12.96	12.68	14.36	14.36	18.84
Std. Dev Eigenvector	12.88	14.20	16.80	16.48	21.28

All methods of SNCD were ineffective for detecting sporadic changes in the Company network. The sporadic change did not persist long enough to signal a possible change in most of the runs. The squad level network was not investigated for this type of change, due to a lack of context.

4.6 Discussion

Statistical process control is a critical quality-engineering tool that assists manufacturing firms in maintaining profitability (Montgomery, 1991; Ryan, 2000). The virtual experiments presented in this chapter demonstrate that SNCD could enable analysts to detect important changes in longitudinal network data. Furthermore, the most likely time that the change occurred can also be determined. This allows one to allocate minimal resources to tracking the general patterns of a network and then shift to full resources when changes are determined[2]. SNCD is therefore, an important analysis method for studying network dynamics.

This chapter describes three algorithms for change detection, and then demonstrates its ability to detect changes in simulated networks. No doubt other change

[2] Two social network change detection algorithms (Shewhart X-Bar and the Cumulative Sum) are available in the "Statistical Network Monitoring Report" in the software tool, Organizational Risk Analyzer (ORA) available through the Center for Computational Analysis of Social and Organizational Systems (CASOS), http://www.casos.cmu.edu.

detection methods will emerge. Our point is that it is critical to be able to detect change in networks over time and to determine when those changes are not simply the random fluctuations of chance. The strengths of the proposed method are its statistical approach, a wide range of social network metrics suitable for application, its ability to identify change points in organizational behavior, and its flexibility for various magnitudes of change. The proposed method requires the assumption of a period of dynamic equilibrium that is necessary to estimate the mean and standard deviation of social network measures for "typical" network observations.

Ideal social network measures to use appear to be standardized node level measures that are averaged over all nodes in the network. Examples, investigated in this paper are the average, maximum, minimum, and standard deviation of the closeness, betweenness, and eigenvector centrality. Future research will provide much greater insight into the strengths and limitations of this approach to the problem. These preliminary results indicate that the average of betweenness and closeness perform well, as does the standard deviation of eigenvector centrality. It is not important to determine which of the network measures is best, because each measures a different relation on the network. If a social scientist was interested in detecting changes in group cohesion, the average closeness measure would be best. However, if one were interested in the changes in informal leadership, the average betweenness would be more appropriate. The specific measures used for change detection should be based on some important set of group behaviors that may change over time. The remainder of this section will identify specific areas of caution when interpreting findings and identify areas for future research.

A limitation of this approach is that the derivation of the CUSUM assumes that network measures are normally distributed. Similar derivations for other distributions can be determined following the same algebraic steps outlined in this paper. Research on the distributions of network measures is needed however. Chapter 2 provides an initial look at some of the challenges in network statistics. Most important is that the context defining links in a network can have significant implications in the distributional properties of a network. Preliminary work on these distributions suggests that the assumption of normality appears to hold for human social networks of 20 or more nodes. The presence of power law distributions in networks is more common in networks that do not require a meaningful investment of time and resources to form a link (Alderson, 2008; McCulloh and Carley, 2009). It should also be pointed out that statistical process control is effectively used in manufacturing when the monitored process is non-normally distributed and when the observations are dependent (Montgomery, 1991; Ryan, 2000). While the CUSUM may be effective in detecting change in non-normally distributed measures in a practical sense, the false positive estimates, which are equivalent to the ADL will be biased. This limitation can be mitigated by checking the chosen measure to monitor using a normal probability plot in a similar fashion to residual analysis in regression[3]. In any case, the normality assumption can be easily verified. Future research will likely provide insight into the performance implications of such bias. It is

[3] A statistical distribution fitting feature is available in the Organizational Risk Analyzer (ORA) available through the Center for Computational Analysis of Social and Organizational Systems (CASOS), http://www.casos.cmu.edu.

important to point out that even with bias; SNCD still provides valuable insight into network dynamics as illustrated with the first three examples in this paper. Future work should consider these factors to determine the range of networks for which this approach will work. Clearly, if the network measures are normally distributed, the CUSUM control chart can be used to monitor network change. If they are not, a different control chart may be more appropriate, or a new approach might look to minimize the bias in false positives. Future work should address this issue.

Another limitation of this approach is that dependence assumptions are ignored. This is common in statistical process control. English (2001) points out that "the independence assumption is dramatically violated in processes subjected to process control." Many manufacturing processes include feedback control systems which create autocorrelation among factors affecting the process. This is similar to problems of dyadic dependence and ergodicity issues with networks. In practice however, statistical process control still provides a great deal of insight, identifying when a process changes. This is no different in a network application. Networks may even have less dependence issues than manufacturing processes. Most manufacturing processes are engineered with feedback and control in an attempt to optimize the process. This is not necessarily true with social networks. Robins and Pattison (2007) lay out several statistical tests involving dependence graphs that can be used to determine if dependence is a statistically significant problem in a network. Just like the issues of normality, the dyadic dependence in the network can be verified similar to residual analysis in regression. If dependence is an issue in the network, SNCD can still be used to determine that a change occurred, however, there may be bias and an increase in the probability of a false positive. Future research should investigate both the impact of dependence on ADL performance as well as methods to better handle the problem statistically.

Social networks may also exhibit periodicity over time. Intuitively, peoples' communication patterns may change in cycles over time. People tend to communicate with different people during the week, while at work, than on the weekends. People may communicate more frequently at certain times of the day. Even seasonal trends may affect observed social networks. The application of wavelet theory and Fourier analysis in particular may provide insight into the periodic behavior of network dynamics. These methods will be investigated in Chapter 5. This will allow SNCD to be more accurate in determining the time a change actually occurred and may reduce the ADL for certain changes.

Future research should also look at the sensitivity of the optimality constant, k and control limit values of the CUSUM Control Chart for network measure change detection. As stated earlier, these values are generally arbitrarily chosen and then optimized for the process. By using further Monte Carlo simulations, a researcher should determine which parameter value would be best in detecting certain types of changes such as sudden large changes or slow creeping shifts. Usage of control charts on comparing models and observations should also be studied to see what specific conclusions could be obtained.

Future work could also investigate hybrid approaches for change detection. For the changes investigated in the virtual experiments presented in this chapter, the scan statistic offered the best ADL. Unfortunately, only the CUSUM offers an estimate of when the change actually occurred. Perhaps, it would be wise to use the scan statistic to detect the presence of a change and then use the CUSUM to estimate when the change actually took place. It is also possible that different parameterizations of the CUSUM and scan statistic might offer different performance under different magnitudes and types of change. This opens many new directions for further research.

Multi-agent simulations provide valuable insight into the performance of control charts for social network change detection applications. Simulations allow an investigator to introduce various changes into a simulated organization and evaluate the time to detect for different algorithms. Simulations provide an efficient means of evaluating change detection on social networks. More importantly, however, is the ability to create more controlled experiments, by fixing certain variables, exploring others, and using many replications to estimate error. Simulation studies will continue to be extremely useful in exploring extensions of this methodology.

A more complex issue that multi-agent simulation can address is the issue of change detection on evolving networks. Perhaps some of the change detection approaches presented here will identify changes in evolutionary behavior. Alternately, the procedures may be modified such that instead of tracking a measure over time, the difference in the measure between regular time intervals is monitored. This would effectively be a rate of change in a network measure and may describe a rate of evolution. Chapter 7 offers a method for use with high variance measures or few node networks that could be modified to model evolutionary behavior. Changes in the residual could then be monitored for change.

Another extension to this work is an approach for detecting multiple changes in the same data set. A basic approach that is applied to the IkeNet data sets in Chapter 6, is to reset the detection procedure following a change. If a social network begins with a state of relationships among agents, I calculate the mean and variance of a network measure of interest to establish typical behavior in the network. I then monitor the network over subsequent time periods. When a change is detected an analyst or social scientist can investigate the group to confirm that there has been a change in the state of relationships between agents. At this point the mean and variance of subsequent measures of the network can be calculated to determine the new typical behavior. At this point, an additional change in the network, to include returning to the original network state may be observed in the change detection procedure.

Social network change detection is important for identifying significant shifts in organizational behavior. This provides insight into policy decisions that drive the underlying change. It also shows the promise of enabling predictive analysis for social networks and providing early warning of potential problems. In the same way that manufacturing firms save millions of dollars each year by quickly responding to changes in their manufacturing process, social network change detection can allow senior leaders

and military analysts to quickly respond to changes in the organizational behavior of the socially connected groups they observe. The combination of statistical process control and social network analysis is likely to produce significant insight into organizational behavior and social dynamics. Immediate applications to counter terrorism and organizational behavior are obvious. As a scientific community we can hope to see more research in this area as network statistics continue to improve.

5 Spectral Analysis of Social Networks to Identify Periodicity

Longitudinal social networks are an important area of study in social network analysis. Stan Wasserman describes "the analysis of social networks over time has long been recognized as something of a Holy Grail for network researchers" (Carrington, Scott and Wasserman, 2007). Pat Doreian (1997) has described the concept of "Network Dynamics" which assumes that there is some underlying stochastic process that drives network behavior over time. McCulloh and Carley (2008) extend this concept to describe four network dynamic behaviors that a network can exhibit over time. A network can remain *stable*, which means that the underlying relationships between agents in a network remain the same, although there may exist some fluctuation in observed links within the network due to measurement error or weak relationship. This type of network can be analyzed as a static network (McCulloh, Lospinoso and Carly, 2007; Wasserman and Faust, 1994). A network can *evolve*, which occurs when relationships between agents change as a result of agent interaction, exchange of beliefs and ideas, or as agents gain a greater knowledge of the traits and resources other agents have in the network. Network evolution has been explored through multi-agent simulation (Banks and Carley, 1996; Sanil, Banks and Carley, 1994; Carley, 1996; Carley, 1999; Doreian and Stokman, 1997). Network evolution has also been explored through Markov chains (Leenders, 1995; Snijders, 1995, 1996, 2001, 2007; Wasserman and Pattison, 1996). A network can exhibit a *shock*, which occurs when some exogenous impact to the network causes relationships to change (McCulloh and Carley, 2008). Finally, a network can experience a *mutation* if a shock initiates evolutionary change (Doreian, 2008). Distinguishing between these four different types of network behavior over time is important for understanding the social mechanisms that drive over-time behavior in social groups.

Social network change detection (McCulloh and Carley, 2008) applies statistical process control to graph level measures within a social network to detect statistically significant changes in a network over time. This has been found to be effective in several different data sets ranging from terrorist networks (McCulloh, Webb and Carley 2007) to email networks (McCulloh et al, 2007). Social network change detection estimates the mean and variance of a graph level measure within a longitudinal set of social networks. Sequential observations of the graph level measure are standardized using the estimated mean and variance and used to calculate some statistic on the network. The test statistic is compared to some decision interval. If the statistic exceeds the decision interval, then the procedure indicates that there may have been a change in the network. The network analysts can use certain change statistics to estimate the point in time when the change most likely occurred. This change may have been evolutionary in nature, or it may have been caused by some exogenous source; a shock. Identifying that the change occurred and when the change occurred are the first two steps in understanding the network dynamics affecting empirical data.

One major obstacle to the study of network dynamics is periodicity or over-time dependence in longitudinal network data. For example, if I define a social network link as an agent sending an email to another, I have continuous time stamped data. Intuitively, we can imagine that individuals are more likely to email each other at certain

times of the day, days of the week, etc. If the individuals in the network are students, then their email traffic might follow the school's academic calendar. Seasonal trends in data are common in a variety of other applications as well. When these periodic changes occur in the relationships that define social network links, social network change detection methods are more likely to signal a *false positive*. A false positive occurs when the social network change detection method indicates that a change in the network may have occurred, when in fact there has been no change. To illustrate, assume that we are monitoring the density of the network for change in hourly intervals. The density of the network measured for the interval between 3 A.M. and 4 A.M. might be significantly less than the network measured from 3 P.M. to 4 P.M. because most of the people in the network are asleep and not communicating between 3 A.M. and 4 A.M. This behavior is to be expected, however, and it is not desirable for the change detection algorithm to signal a potential change at this point. Rather, it would be ideal to control for this phenomenon by accounting for the time periodicity in the density measure. Only then can real change be identified quickly in a background of noise.

Periodicity can occur in many kinds of longitudinal data. Organizations may experience periodicity as a result of scheduled events, such as a weekly meeting or monthly social event. Social networks collected on college students are likely to have periodicity driven by both the semester schedule and academic year. Even the weather may introduce periodicity in social network data, as people are more or less likely to email, or interact face-to-face. At the U.S. Military Academy, people tend to run outside in warm weather in small groups of two or three. During the winter, people go to the gym, where they are likely to see many people. This causes an increase in face-to-face interaction as people stay inside. In a similar fashion, during the Spring and Fall, many people participate in inter-unit sporting events such as soccer, or Frisbee football. This can also affect people' face-to-face interaction and thus the social network data collected on them.

Spectral analysis provides a framework to understand periodicity. Spectral analysis is mathematical tool used to analyze functions or signals in the frequency domain as opposed to the time domain. If we look at some measure of a social group over time, we are conducting analysis in the time domain. The frequency domain allows us to investigate how much of the given measure lies within each frequency band over a range of frequencies. For example, Figure 11 shows a notional measure on some made-up group in the time domain. It can be seen that the measure is larger at points B and D corresponding to the middle of the week. The measure is smaller at points A, C, and E.

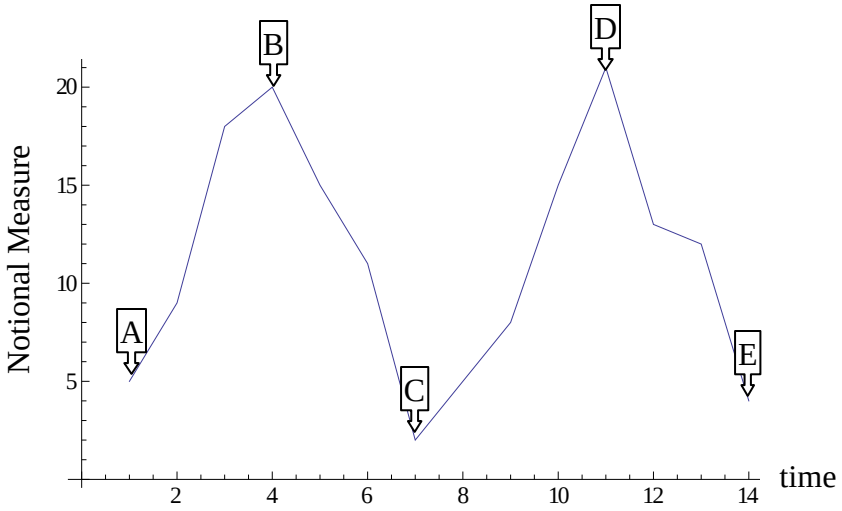


Figure 11 Notional Measure in Time Domain

If the signal in Figure 11 is converted to the frequency domain as shown in Figure 12, we can see how much of the measure lies within certain frequency bands. The negative spike in Figure 12 corresponds to 7 days, which is the weekly periodicity in the notional signal. The actual frequency signal only runs to a value of 8 on the x-axis in Figure 12. The frequency domain signal after a value of 8 is a mirror image, or harmonic of the actual frequency signal.

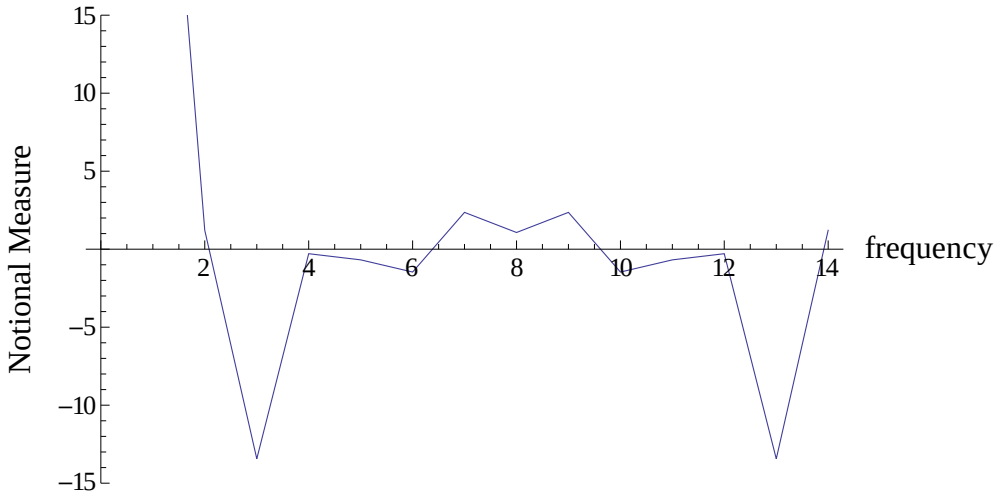


Figure 12. Notional Measure in Frequency Domain

The frequency domain representation of a signal also includes the phase shift that must be applied to a summation of sine functions to reconstruct the original over-time signal. In other words, we can combine daily, weekly, monthly, semester, and annual periodicity to recover the expected signal over-time due to periodicity. For example, Figures 13-15 represent monthly, weekly, and sub-weekly periodicities. If these signals are added together, meaning that the observed social network exhibits all three of these periodic behaviors, the resulting signal is shown in Figure 16.

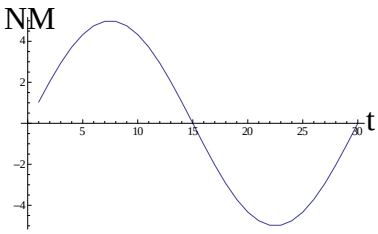


Figure 13. Monthly Period

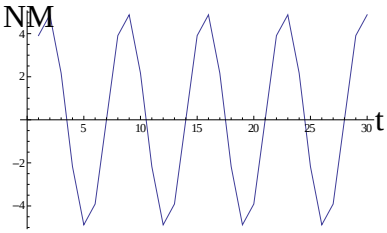


Figure 14. Weekly Period

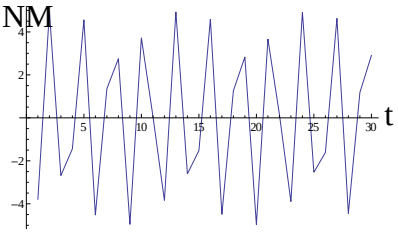


Figure 15. Sub-weekly Period

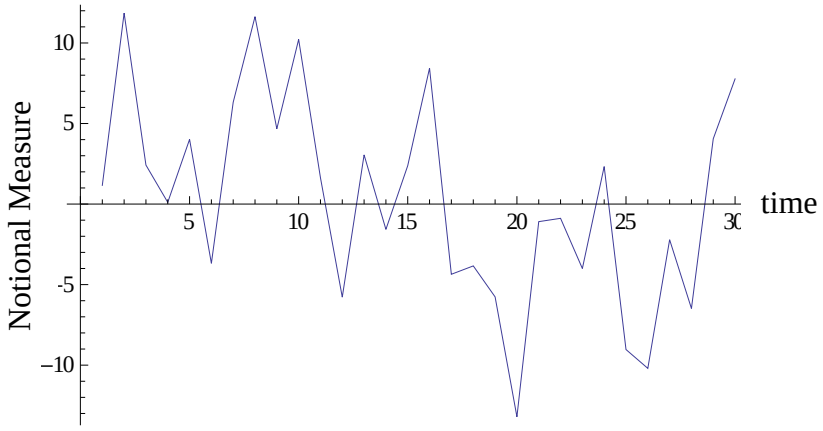


Figure 16. Sum of the Signal in Figures 13-15

If the periodicity in the signal shown in Figure 16 is not accounted for, it appears that there may be a change in behavior around time period 20, where the signal is negatively spiked. In reality, this behavior is caused by periodicity. If we transform the signal to the frequency domain as shown in Figure 17, we can see the weekly periodicity at point B and the sub-weekly periodicity at point A.

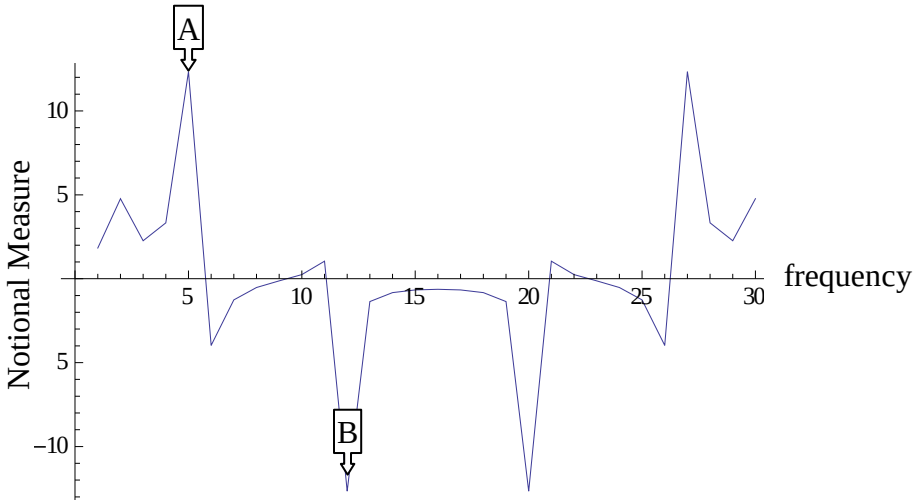


Figure 17. Transformation of Figure 16 to the Frequency Domain

I propose that spectral analysis applied to social network measures over time will identify periodicity in the network. I will transform an over time network measure from

the time domain to the frequency domain using a Fourier transform. I will then identify significant periodicity in the over-time network and present two methods for handling the periodicity. This newly proposed method will be demonstrated on real-world data sets as well as simulated data sets.

Handling periodicity is a very important problem. For social scientists to gain insight into the evolution of social networks, they must be able to distinguish between shock, evolutionary change, and typical periodic behavior. This chapter will lay out a method for identifying and removing the periodic behavior of a signal so that change detection can be performed more accurately.

5.1 Background

Networks can be described by a number of different measures. Measures can be defined for individual nodes or for the network as a whole. In this thesis, I will restrict our attention to network level measures; although I point out that there is no reason that the methodology presented could not be applied to node level measures. Common network level measures include density, the number of nodes in the network, and the average path length. In addition, node level measures such as betweenness, closeness, and eigenvector centrality can be averaged over all nodes in a network to create network level measures.

These measures may fluctuate in a periodic fashion over time. As agents in a network change their relationships to other agents based on seasonal trends, these fluctuations may be noticed in the network measures of those relationships. For example, during the workweek, one might expect more email communication within an office than during the weekend. This could be observed by a greater network density (percentage of possible relationships) during the week than during the weekend. The social network measures therefore provide a measure of the group as a whole.

Spectral analysis can be used to detect periodicity within social network measures over time. Periodicity in the social network measure provides some insight into the periodicity in the underlying social organization. Spectral analysis can therefore be used to either filter out periodicity in over-time measures or provide insight into how data should be aggregated to best represent a social group.

Spectral analysis is a mathematical process of converting some function or series of numbers which I call a signal, from the time domain into the frequency domain. A function or signal can be converted from the time domain to the frequency domains with a transformation. A common transformation is the Fourier transform, which decomposes a signal into a sum of sine waves having different phase shifts and amplitudes. The Fourier transform is given by,

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi f t} dt.$$

A convenient property of the Fourier transform is that the inverse of the Fourier transform is also a Fourier transform. This property makes it convenient to convert back and forth between the time and frequency domains. I will use this property to convert a signal from the time domain to the frequency domain; identify significant frequencies; and convert those frequencies back into the time domain to provide an understanding of the periodicity inherent in longitudinal social network measures.

5.2 Data

The approaches for handling periodicity in network data are demonstrated on a longitudinal data set of email traffic collected at the U.S. Military Academy at West Point, NY. This data set was collected as part of this thesis to demonstrate longitudinal network analysis. The participants consist of 25 undergraduate cadets at the U.S. Military Academy serving in military leadership positions in one of four cadet Regiments. All participants volunteered to allow me to monitor the header information of their email traffic for the Fall semester of 2008. This study was approved for ethics by the West Point institutional review board. The email header information was used to create social networks by assigning a directed link from node *i* to node *j* if node *i* send node *j* an email sometime during the designated time period. This unique data set allowed me to investigate the periodicity of the data for many hourly networks, or a few monthly networks. In addition, I was able to interview the participants to investigate potential causes of periodicity in the email communication networks.

While the West Point Cadet data is sufficient to demonstrate spectral analysis of networks, I use a simulated periodic signal to demonstrate the importance of spectral analysis for change detection. The simulated data consists of a simulated sine wave representing some measure of interest, where a change in the mean of the wave is introduced at a known point in time. Random uniform error between 0 and the amplitude of the sine wave is added to the signal. The accuracy of the CUSUM change point identification against a background of noise is then compared between whether spectral analysis is applied or not.

5.3 Method

The spectral analysis approach proposed in this thesis consists of five steps to determine the significant periodicity and then suggests two methods of handling the periodicity in the data. I list these steps here and demonstrate them on the West Point Cadet data in the next section.

Step 1: Plot the measure of interest. This first step is to determine network measures of interest. These can be network level measures or node level measures. In this thesis I have restricted my attention to network level measures. For the purpose of demonstration, I will use the average betweenness of nodes in the network as a network level measure. Another issue in this step is number and length of time periods. In this

example, I investigate daily networks, with the hope of determining weekly or monthly periodicity. I could measure hourly networks, or even networks corresponding to each second of the day. Intuitively, smaller time periods will result in sparser networks. Some amount of judgment will be required by the analyst to select an aggregation level where most of the nodes in the network are connected, but every node is not necessarily connected to every other node.

Step 2: Fast Fourier Transform. The second step is to transform the network measure of interest from the time domain to the frequency domain. Since the network measures correspond to discrete time periods and the measure is not continuous, the Fourier transformation cannot be applied directly. A discrete version of the Fourier transform is used. The discrete version is known as the Fast Fourier Transform. This operation is standard in many mathematical software packages such as MATLAB and Mathematica. It is also available in the Organizational Risk Analyzer (ORA).

Step 3: Determine normal frequencies. The third step is to determine the normal range of frequencies for the signal. The Fourier transform makes use of the normal distribution function for its transformation. Therefore, we may assume that the frequencies of the transformed signal approximate a normal distribution. In fitting a normal distribution to the frequencies, we will be able to determine statistically anomalous or significant frequencies.

Step 4: Identify significant frequencies. This step requires that the analyst determine a confidence level for detecting periodicity. The 95% confidence level is approximately equal to ±2 standard deviations from the mean frequency. Therefore, all frequencies within 2 standard deviations from the mean are set to equal 0. This creates a new discrete signal in the frequency domain of only statistically significant signals.

Step 5: Identify significant periods. Recall that the Fourier transform is also its inverse. Therefore, the Fast Fourier Transform is applied to the discrete signal in Step 4 to determine the significant periodicity.

At this point the analyst has two options for handling the periodicity in the data. The simplest method is to aggregate over the period. For example, the analyst may find weekly periodicity. People may have different email behavior on the weekend than they do during the weekday. The analyst could then aggregate over the daily networks to create weekly networks. Then the weekly periodicity would be controlled within each weekly network. If the network becomes too dense by establishing a link between nodes for a single weekly email, the analyst is free to require more than one email message to define a link. This is an important decision that was discussed in greater detail in Chapter 2.

The analyst can also choose to keep using the weekly networks, but control for the periodicity. The discrete signal in Step 5 is really the expected value of the chosen social network measure from Step 1 for each point in time. The analyst can create a filtered network measure by taking the difference between the original signal from Step 1

and the signal from Step 5. This new signal is then a filtered signal that can improve the performance of social network change detection.

This second approach for handling periodicity is investigated through simulation. A periodic signal is simulated in Mathematica, a mathematical software environment. The signal is shifted at a particular point in time. Uniform random noise is added to the signal where the range of error is equal to the amplitude of the signal. The CUSUM change detection algorithm is applied to the periodic signal as well as a signal filtered in the manner described above. The change point identification of the CUSUM applied to each signal is compared.

5.4 Results

The West Point Cadet data average betweenness is displayed in Figure 18 for a one month period during the Fall 2008 semester. If an analyst were just looking at this data, it may appear that the average betweenness is unusually high around day 15. There also appears to be moderately high values around day 8 and day 22.

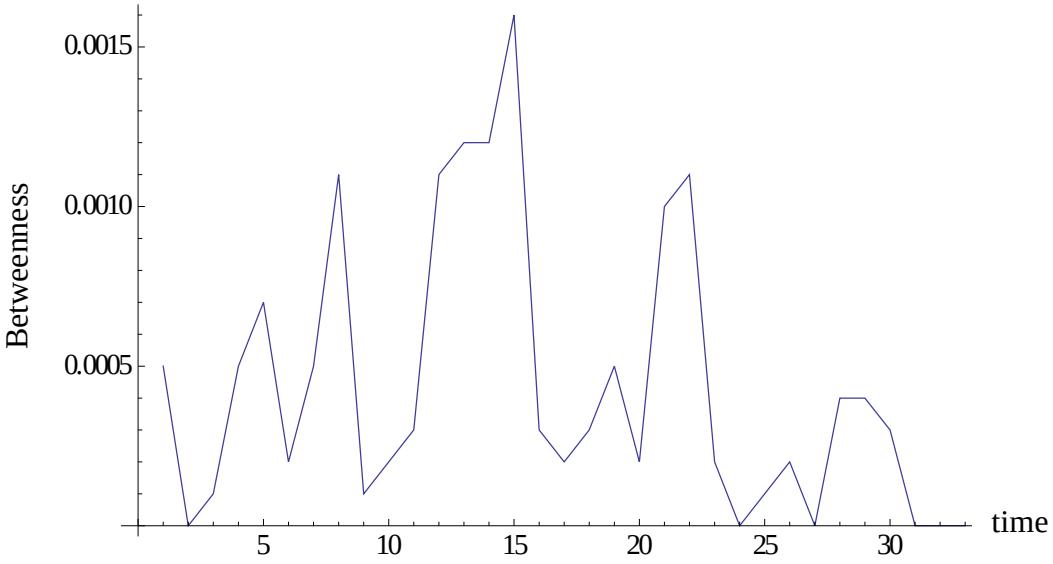


Figure 18. West Point Cadet Data Average Betweenness

The Fast Fourier Transform is applied to the average betweenness scores, transforming these values from the time domain to the frequency domain. A plot of the transformed values is shown in Figure 19. It appears that there may exist significant periodicity in the over-time measure.

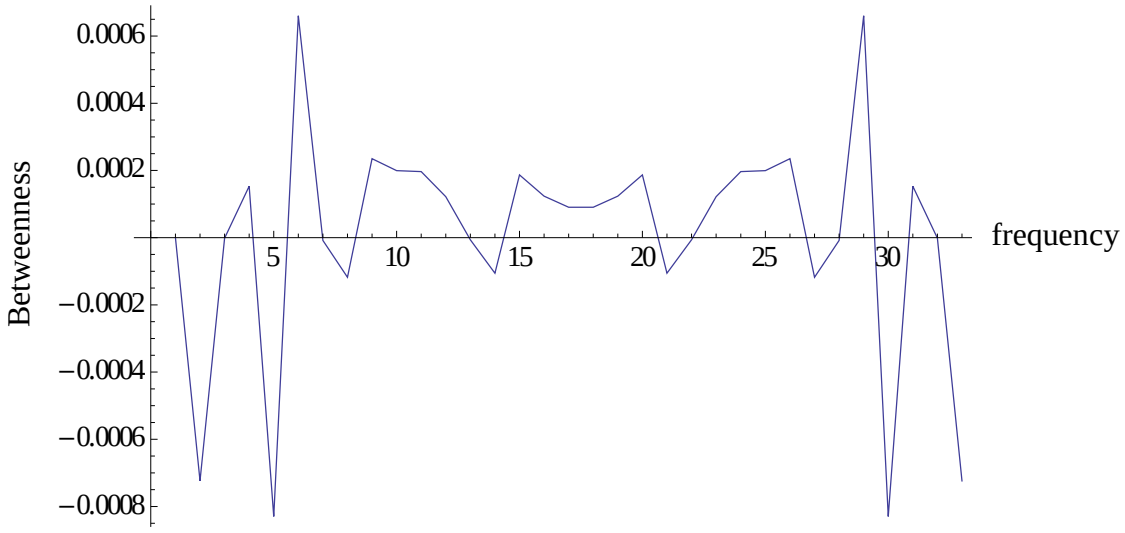


Figure 19. Fast Fourier Transform of West Point Cadet Data Avg. Btwn.

A normal distribution is fit to the discrete frequency signal and values within two standard deviations of the mean are set equal to 0. Figure 20 shows the significant frequencies.

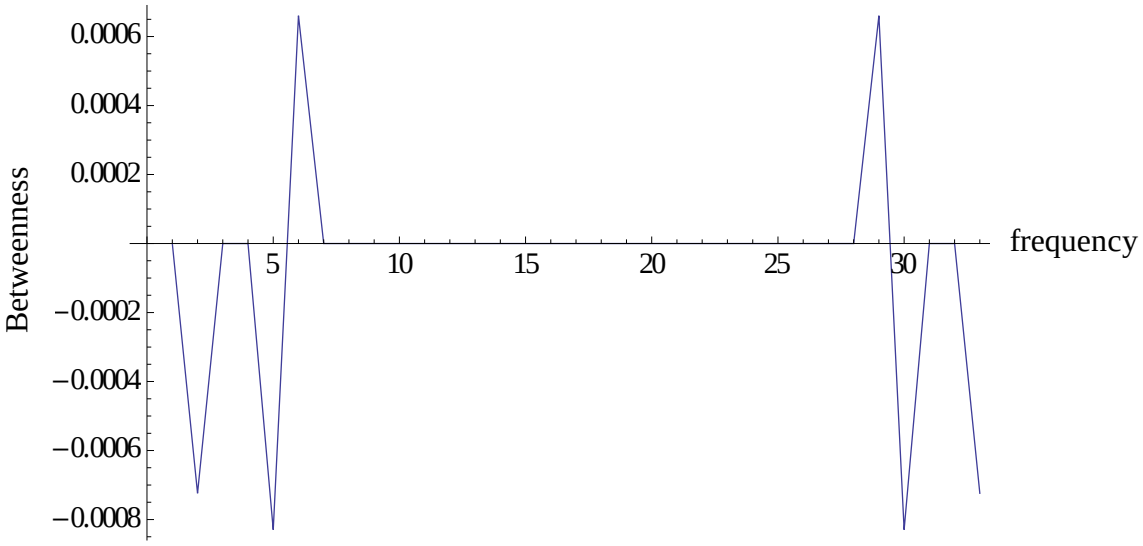


Figure 20. Significant Frequencies in West Point Cadet Data

The significant frequencies are transformed back into the time domain using the Fast Fourier Transform. This is known as taking an inverse transform of the signal. The resulting plot in the time domain can be interpreted as the significant periodicity in the measure, since only the signficant frequencies were transformed back into the time domain. The significant frequencies are plotted in the frequency domain. The significant periodicity, on the other hand, is plotted in the time domain. Figure 21 displays a plot of the significant periodicity in the average betweenness signal.

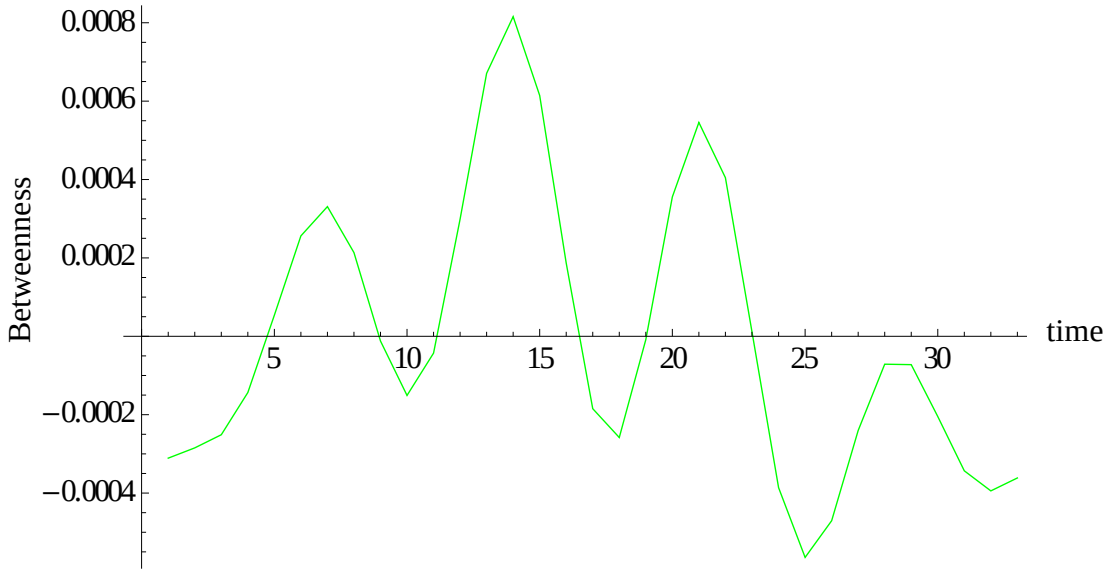


Figure 21. Significant Periodicity in the West Point Cadet Data

It can be seen in Figure 21 that there is a spike in significant periodicity corresponding to days 7, 14, 21, and 28. This is perfect weekly periodicity. An interview with the regimental commander of the participants in the study revealed that the participants have a weekly meeting every Sunday. During this meeting, important information is given to the group regarding events and activities for the week. In addition, subordinate leaders are required to account for the whereabouts of all of the cadets within their subordinate units and report the information up the chain of command. This process of sending information up and down the chain of command will significantly affect the average betweenness of the network on Sundays. Failing to account for this behavior may in turn affect an analyst's ability to detect real organizational change within this group.

At this point, an analyst can choose to monitor weekly networks, or continue to monitor daily networks and filter out some of the periodicity. Figure 22 shows a filtered signal in the time domain. This signal was obtained by taking the original signal found in Figure 18 and subtracting the periodicity found in Figure 21 for each time period. In effect, the new figure shown in Figure 22, displays the deviation from what is expected in the signal due to the time of week.

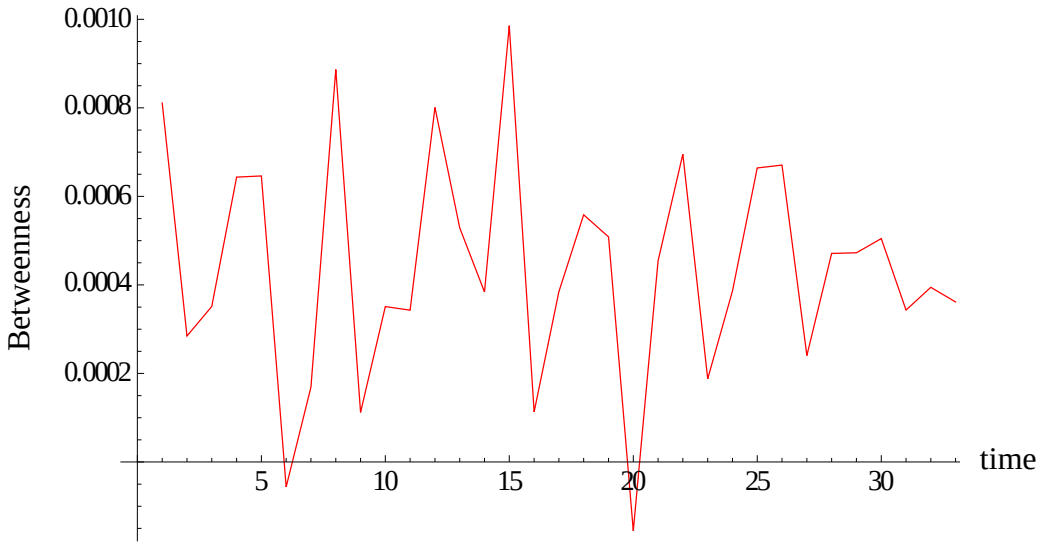


Figure 22. Filtered Plot of Average Betweenness in the West Point Cadet Data

Figure 23 shows the original and filtered signals together. It can be seen that the extreme values of average betweenness detected in our first observation of the network do not appear as extreme in the filtered signal. Therefore, the filtered signal is less likely to cause a false alarm in change detection.

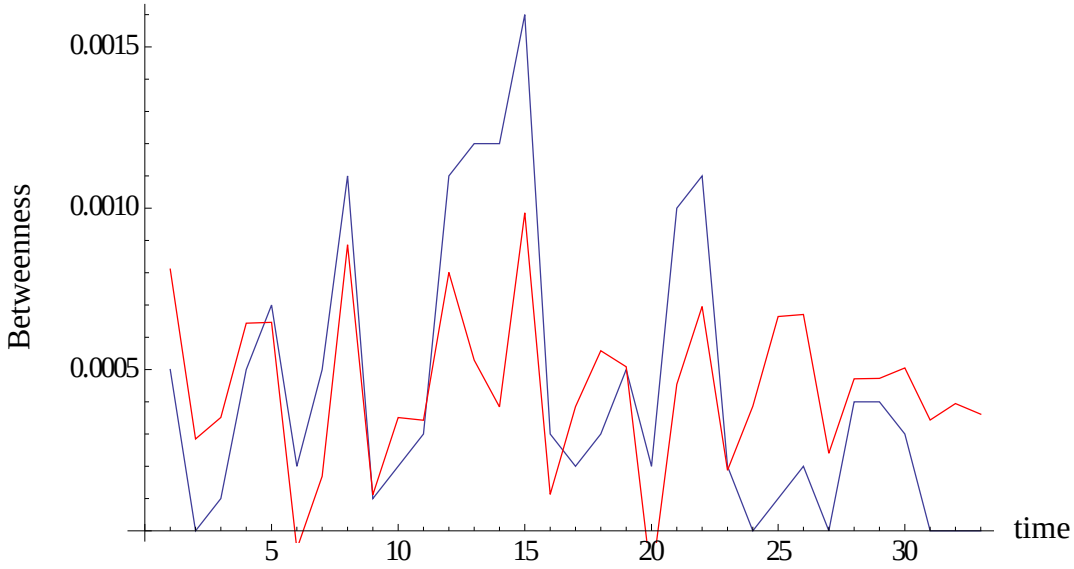


Figure 23. Original and Filtered Plots of Average Betweenness

To further illustrate the importance of accounting for periodicity, we turn our attention to an extreme case. Figure 24 displays a sine wave, where a change in the mean of the signal occurs at time period 40. In addition to the periodicity, noise is added to the signal in the form of uniform random error with a range equal to the amplitude of the sine wave. A random instance of this signal is displayed in Figure 25. It can be seen that identifying the change at time period 40 may be difficult with the combination of periodicity and noise.

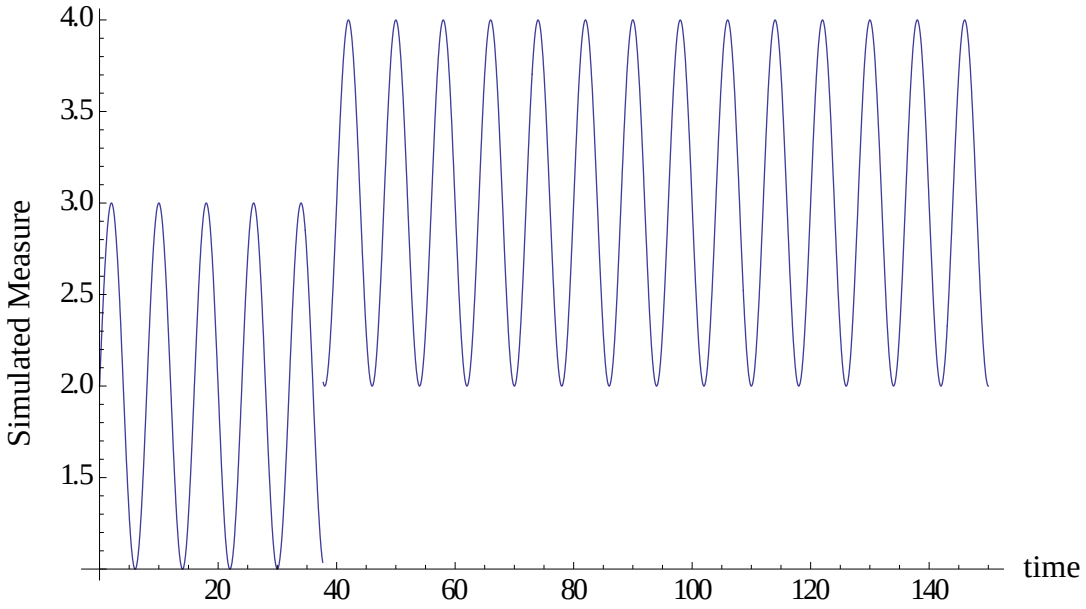


Figure 24. Sine Wave with Change at Time 40

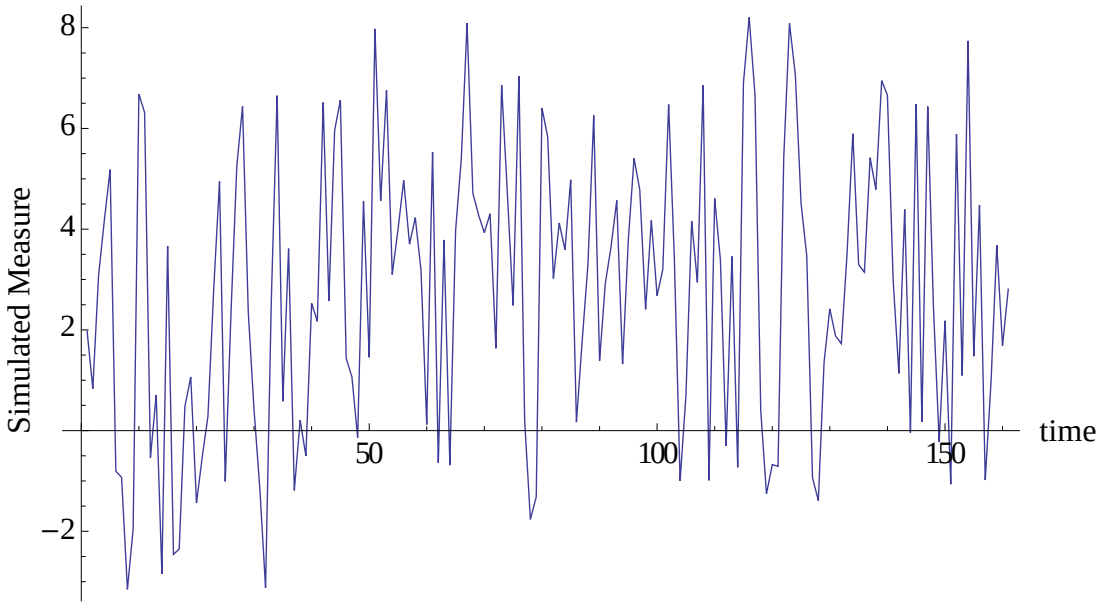


Figure 25. Sine Wave with Random Error and Change at Time 40

The CUSUM change detection algorithm is applied to the noisy signal in Figure 25. Figure 26 shows a plot of the CUSUM statistic. It can be seen that the CUSUM statistic can be powerful in illuminating subtle change in a background of noise. It also appears that the algorithm may have signalled false alarms around time pionts 10 and 30. It is not clear that there is a good solid indication of change until after time point 50.

The filtering approach can be extremely useful in improving the performance of the change detection approach. Figure 27 shows a plot of the CUSUM statistic on the same signal as Figures 25 and 26, where the signal was first filtered for periodicity using

the steps outlined above. It can be seen in Figure 27 that the signal may more acurately identify the correct change point in the signal and is less prone to false signal.

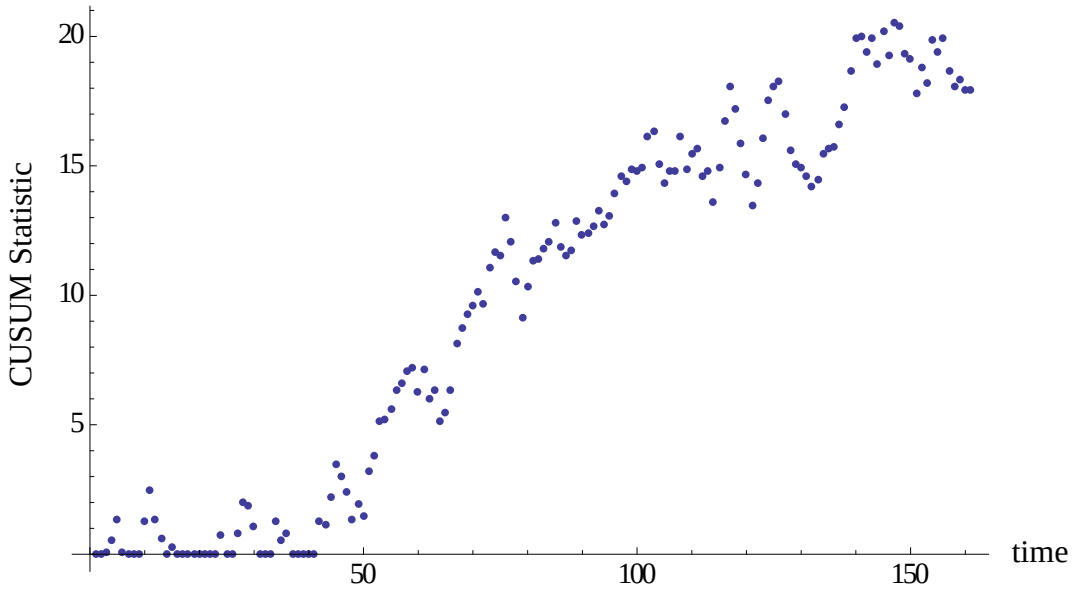


Figure 26. CUSUM Statistic Applied to Noisy Sine Wave

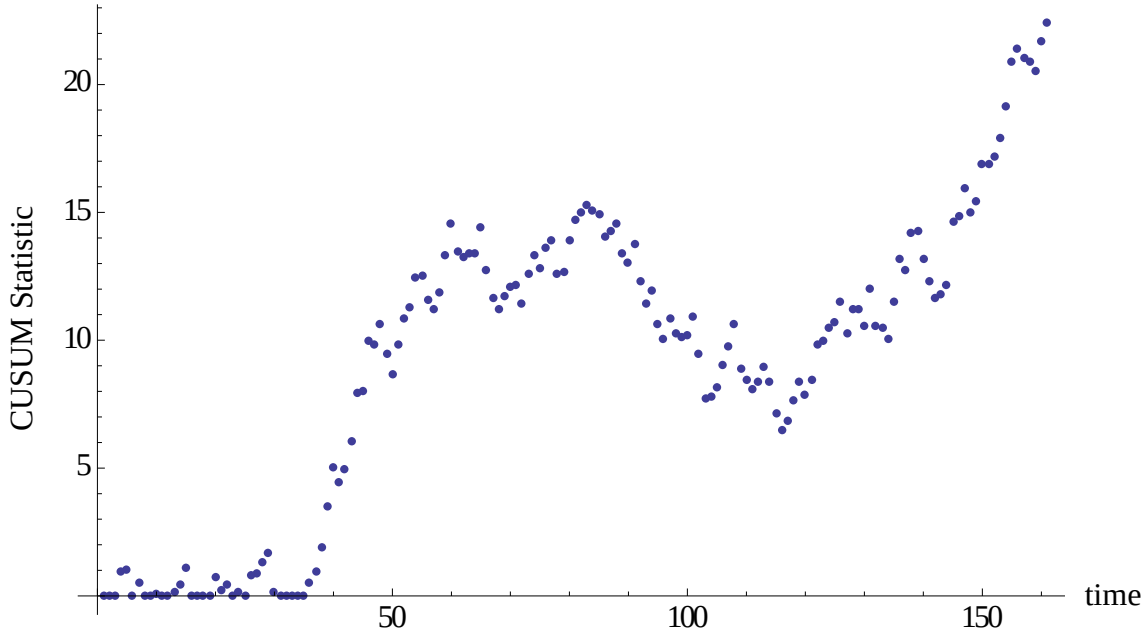


Figure 27. CUSUM Statistic Applied to Filtered Signal

The simulation was repeated with four different levels of uniform random noise. The level of random noise was set as a percentage of the amplitude of the sine wave at 30%, 50%, 67%, and 100%. The change occurred at time 40 and the size of the change was the amplitude. The average time to detect the change was compared across the four levels of noise. For each simulation run, the CUSUM was applied to both the original

signal and the filtered signal. A pairwise t-test for the time to detect change was conducted between the original and filtered signals for 100 independently seeded instances of the noisy sine wave. The null hypothesis was that there was no difference between detection performance between the original and filtered signals. The p-values for this null hypothesis are 0.05, 0.04, 0.72, and 0.88 respectively for noise levels of 30%, 50%, 67%, and 100% of amplitude. The p-values for the error that was less than or equal to 50% of amplitude are significant, indicating that the filtering improves the time to detect a change. The p-values for the error that was greater than 50% of the amplitude are not significant, meaning we have no reason to reject the null hypothesis that filtering does not improve change detection.

This behavior in performance appears reasonable. If the periodicity in the over-time measure is greater than the level of observation error, then filtering the signal is likely to improve change detection performance. If on the other hand, the level of error in the observed over-time measure is greater than the periodicity, then spikes in error may appear as a significant frequency, which may adversely bias the change detection algorithm. It is possible that if the error is much greater than periodicity, the spectral analysis may even mask true change. Future work should investigate the impacts of spectral analysis on change detection performance.

5.5 Conclusion

Periodicity is an important issue in the longitudinal analysis of social networks. Intuitively, peoples' observable relationships may change with the time of day, week, month, year, etc. Accurate modeling of social network relations therefore requires a way to account for and control for this periodicity. This issue is especially important for any longitudinal analysis.

Fourier analysis can detect periodicity and provide insight to control for its effect. The success of this approach has been demonstrated on both real-world and simulated data sets. More research is needed to investigate how observation error and organizational dynamics might affect the periodicity. It is expected that if the random error in the signal is much higher than the amplitude, the filtering techniques proposed here may not be effective. Likewise, if there is very little error, filtering may be unnecessary. For most longitudinal analysis, however, I propose that applying the approach laid out in this Chapter may detect significant periodicity and therefore improve the performance of change detection.

The spectral analysis has only been investigated for filtering and detecting trigonometric cycles in an over-time signal. It is conceivable that some forms of periodicity may not follow a trigonometric cycle. For example, major holidays in the U.S. are likely to affect communication patterns between individuals; however, they do not occur on the calendar with regular trigonometric frequency. In addition, changes in relations may taper off suddenly as in the case of an organization that has a prescribed start and stop time to the work day. In this situation a sine wave may not appropriately

capture the periodic behavior of the group. More research into Wavelets that consider different periodic signals is warranted. While the same general approach laid out here may apply, the choice of transformation may differ.

The success of spectral analysis will be related to the number of available time periods with network data. This approach requires continuous data with many time periods. This type of data may be difficult to obtain. In some cases the number of longitudinal networks may be already aggregated over some period of time. I recommend that a prospective analyst apply this approach when looking at longitudinal data, but be aware of the potential problems when investigating fewer than 10 longitudinal networks.

Spectral analysis of longitudinal network measures appears to be a powerful technique for understanding periodicity in over-time data. While an entire thesis could be devoted to this topic alone, I have shown how it can be effective on one real-world data set. I have further demonstrated how spectral analysis can improve the performance of the CUSUM algorithm using a simulated noisy sine wave. In addition to the change detection performance implications, this approach also leads to interesting insights into organizational behavior. The spectral analysis of the West Point cadet data for example, revealed the organization's weekly meeting time. Whether used for change detection or simply organizational insight, spectral analysis represents a major contribution to the study of longitudinal network data.

6 Real World Examples

Social network change detection (SNCD) is demonstrated on eight different real world data sets in this chapter. These examples will serve to demonstrate the applicability and promise of this novel approach to longitudinal network analysis. In addition, the examples will hopefully make the change detection process more clear with tangible examples.

The CUSUM procedure is used to demonstrate SNCD on the eight data sets. Recall that there are two important things to detect in longitudinal analysis. It is important to detect *that* a change occurred and *when* a change occurred. As explained in Chapter 4, the CUSUM, EWMA, and Scan statistic all detect that a change occurred, but only the CUSUM provides an estimate of when a change occurred.

For exposition purposes the CUSUM procedure used the same parameterization for all example data sets. The optimality constant was set to $k = 0.5$, which corresponds to a one standard deviation shift in the monitored measure. The decision interval was set to $h = 3.5$, which corresponds to a false positive rate of 1%. The sensitivity of the parameters is investigated in the last section by using a decision interval of $h = 2.5$, 3.0, and 4.2, which corresponds to false positive rates of 5%, 2%, and 0.5% respectively. For each of these parameters the time *that* a change occurred and the estimate of *when* the change occurred are presented in separate tables and compared to the $h = 3.5$ situation.

The data sets were chosen to represent a range of potential social network data. They vary in size from 17 nodes to 260. They include various data collection schemes from surveys to monitored email communication. Some of the data sets have been well established in the social network literature. Many of the data sets have a known point in time when a change occurred in the network. Finally, all the networks have at least eight time periods, allowing typical network behavior to be assessed so that change detection can be run. As a rule of thumb, I use the first 20% of the data to assess typical behavior of the network in these examples. Table 23 provides a comparison of the data sets used in this chapter for exposition purposes.

Table 23. Comparison of Real World Data.

	No Nodes	Time Periods	Method of Collection	Type of Relation	Design	Known Change
Fraternity	17	15	Survey	Ranking	Fixed	Yes
Leav 07	68	8	Survey	Rating	Free	Yes
Leav 05	158	9	Survey	Rating	Free	None
Al-Qaeda	62-260	17	Text	Rating	Free	Yes
Winter C	22	9	Observation & Survey	Rating	Fixed	Yes
Winter A	28	9	Observation & Survey	Rating	Fixed	Yes
IkeNet 2	22	46	Email	Count Msg	Free	Yes
IkeNet 3	68	121	Email	Count Msg	Free	Yes

6.1 Newcomb Fraternity Data

The first data set was collected by Theodore Newcomb (1961) at the University of Michigan. The participants included 17 incoming transfer students, with no prior acquaintance, who were housed together in fraternity housing. The participants were asked to rank their preference of individuals in the house from 1 to 16, where 1 is their first choice. Data was collected each week for 15 weeks, except for week number 9. David Krackhardt (1998) dichotomized the network data by assigning a link to preference ratings of 1-8 and having no link for ratings of 9-16. A visualization of the Newcomb Fraternity network for time period 8 is shown in Figure 28. The mean and standard deviation of the density, average betweenness, and average closeness was estimated from the first five networks to determine typical behavior. The CUSUM statistic was then calculated for all time periods.

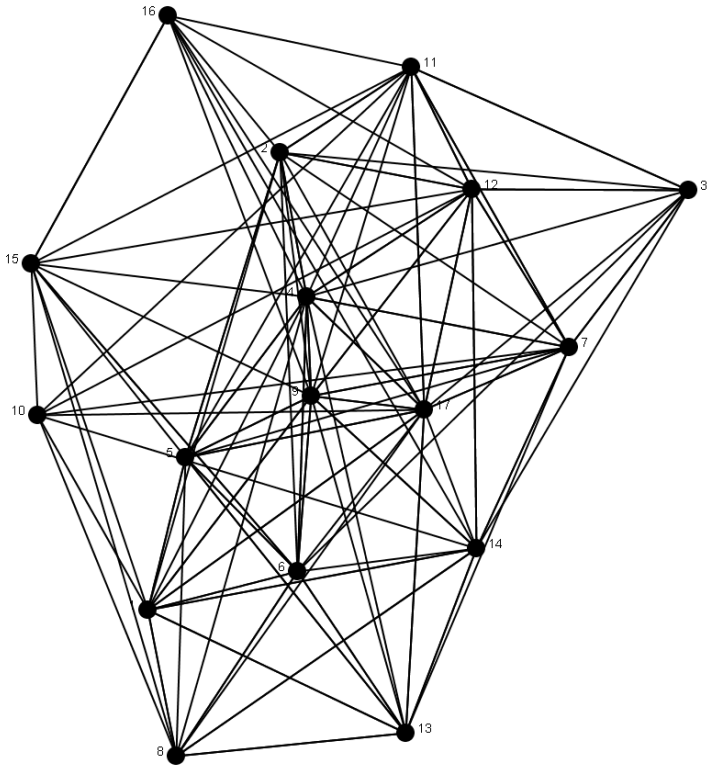


Figure 28. Dichotomized Newcomb Fraternity Network for Time Period 8.

The approach proposed in this paper was found to be successful at detecting significant events in the Fraternity data. Figure 29 displays a plot of the *C* statistics for average betweenness over time for the Newcomb Fraternity data. Recall that the CUSUM will detect either increases or decreases in a measure, but not both. Therefore, two control charts must be run for each social network measure monitored. In the figure, the two lines correspond to the chart for detecting increases in the measure and the chart for detecting decreases in the measure over time. The trends in the data for the betweenness measure are similar to the closeness measure. The density measure is not effective for change detection since the network is fixed-choice and the density remains 0.5 for every network.

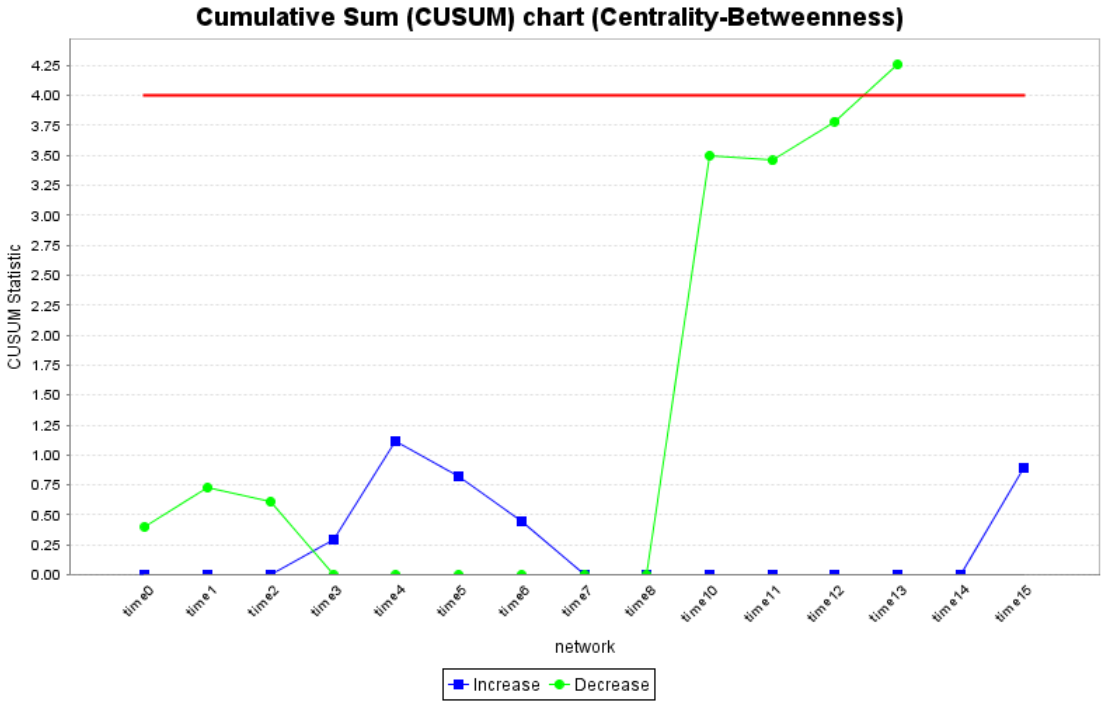


Figure 29. Plot of the CUSUM C Statistic Over Time for the Newcomb Fraternity Data..

According to Figure 29, the control chart for average betweenness signals at time period 13 that a change may have occurred in the social network of the fraternity members. The most likely time that the change actually occurred is the last time period that the *C* statistic was equal to 0. This change point corresponds to time period 8 in the Newcomb Fraternity data, which was the week before a mid-semester Break. It is not unreasonable that social relationships may have changed over a Break as participants possibly vacationed together. Unfortunately, the exact activities and dynamics of the group are not completely known. However, this data does provide evidence of the importance of the proposed method in analyzing network dynamics.

6.2 Leavenworth 2007 Data

The second data set was collected from an Army war fighting simulation at Fort Leavenworth, Kansas in April 2007, by Craig Schreiber. The participants were mid-career U.S. Army officers taking part in a brigade level staff training exercise. There were 68 participants in this data set, who served as staff members in the headquarters of the brigade conducting a simulated training exercise. The data contains the communication between agents in the network which were collected through self reported communications surveys. Data was collected over a period of four days, twice per day. Thus, there were 8 time periods. Half way through the second day (after time period 3), the Brigade Commander was displeased at the lack of coordination between the officers in the exercise. He brought all 68 participants together and chastised them for their performance and told them that they were expected to perform better. Therefore,

SNCD might be able to indicate a significant change in the network corresponding to the Brigade Commander's interaction with the participants. This data set is unique in that it contains a known change point in time that can be used to validate the proposed method. Figure 30 shows the social network for time period 4 from the Leavenworth data set. The mean and standard deviation of the density, average betweenness, and average closeness was estimated from the first three networks to determine typical behavior. The CUSUM statistic was then calculated for all time periods. Three time periods were used because that represents about 30% of the time periods and is comparable to the number used with the Newcomb Fraternity data. Ideally, more networks will allow a more accurate estimate of typical behavior. The reader is reminded that these examples are used to illustrate the proposed methodology, while the performance of the method is evaluated using a simulated data set.

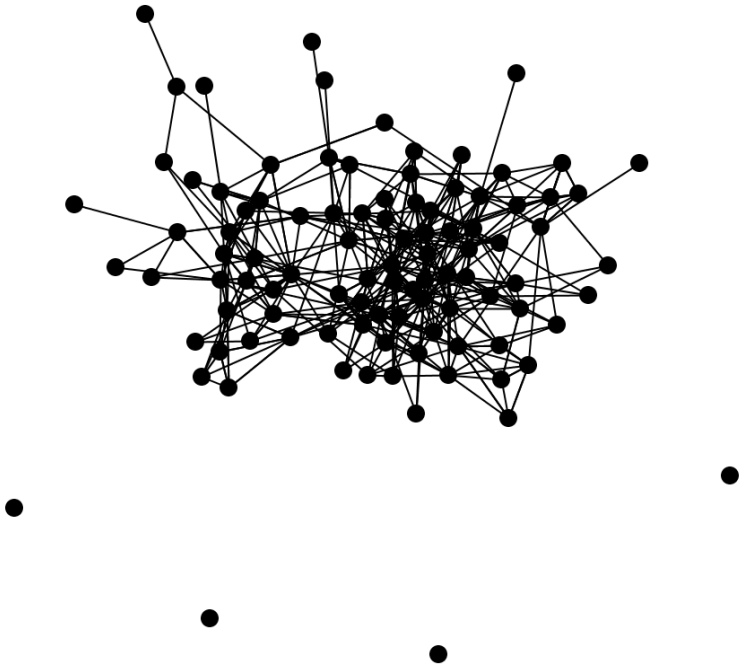


Figure 30. Leavenworth Network for Time Period 4.

The Leavenworth data perhaps provides more compelling support for SNCD. Figure 31 illustrates the *C* statistics for average betweenness over time. The chart in Figure 31 signals at time period 5 that a change in the network may have occurred. The likely time the change actually took place is time period 3 which coincides with the Brigade Commander chastising the members of the group.

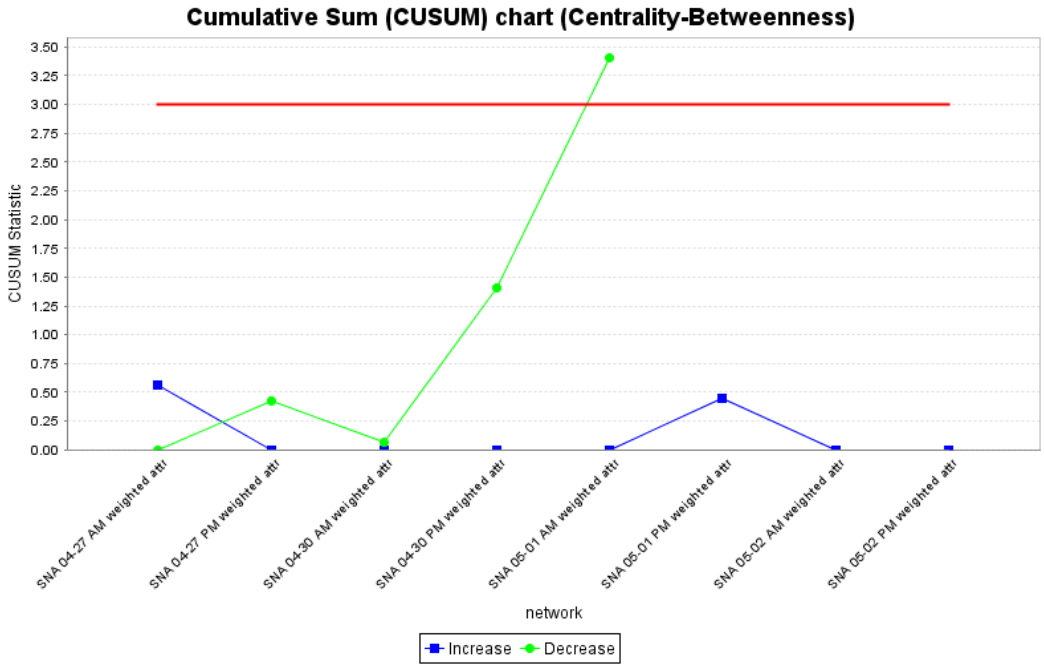


Figure 31. Plot of the CUSUM C Statistic Over Time for the Leavenworth Data.

6.3 Leavenworth 2005 Data

The third data set is very similar to the Leavenworth 2007 Data and was also collected from a war fighting simulation in FT Leavenworth, KS, this time in 2005. The data was collected by Craig Schreiber and Lieutenant Colonel John Graham. This data set contains 156 mid-career Army officers that were monitored over the course of nine iterations of a military command and control exercise conducted over the course of 5 days. This data set displays the communication of all agents in the network based on self reported communications surveys. Figure 32 shows the network for time period 4 for the Leavenworth 2005 Data.

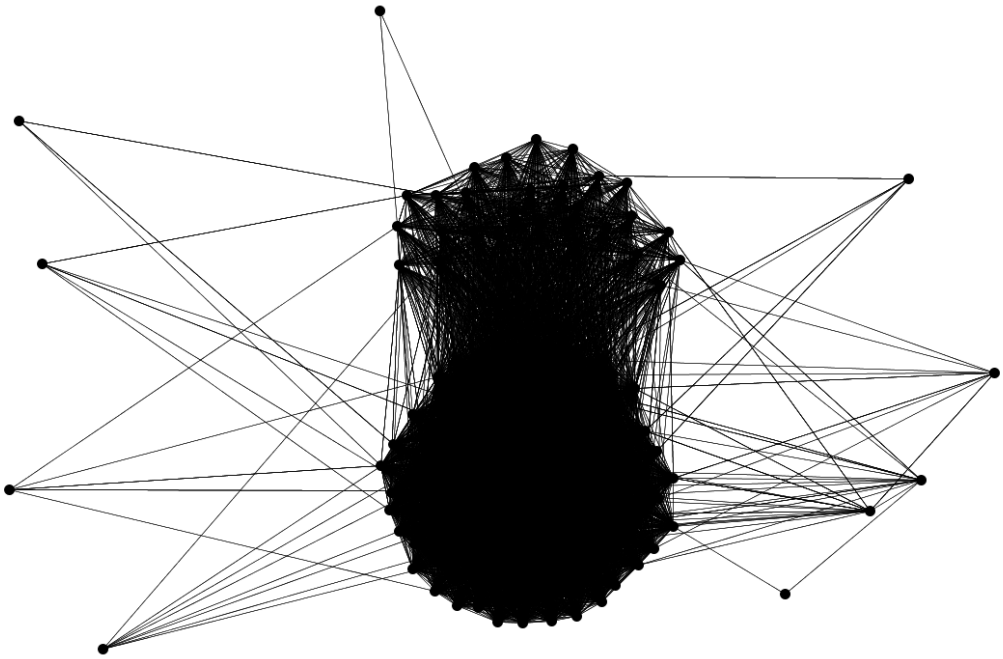


Figure 32. Time Period 4, Leavenworth 2005 Data

Unlike the Leavenworth 2007 Data there is no known shock to the network. Furthermore, the officers in the network have been working together for several months and as a result, network evolution is unlikely to play a role in network dynamics over the course of a single week. It is therefore expected that there will be no identifiable change point in this data set.

The CUSUM control chart was run on the Leavenworth 2005 Data. The risk of false positive was set to 0.01, which would correspond to a decision interval of approximately $h = 3.5$. The optimality constant was set to detect a one standard deviation shift in the mean of a measure. Figure 33 shows the CUSUM procedure for the average betweenness value of nodes in the network. The maximum value of the change statistic is 0.91 at time period 1, which does not exceed the decision interval. Therefore, there is no statistical evidence that any changes are likely to have occurred.

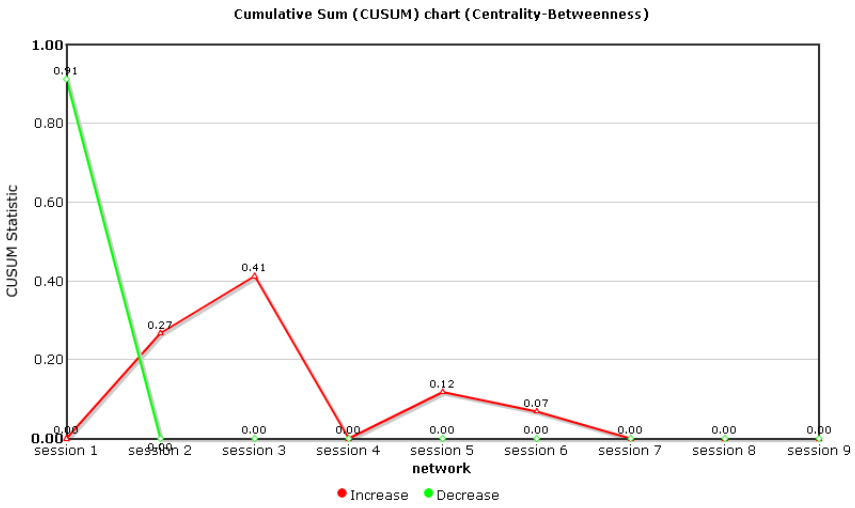


Figure 33. CUSUM of Average Betweenness for Leavenworth 2005 Data.

This data is especially interesting in comparison with the Leavenworth 2007 Data. Both data sets are collected on an equivalent demographic of individuals performing the identical tasks over a similar length of time. In one situation, there was a known change introduced to the group, which I classify as a shock. In the other situation, the 2005 case, there was no identified shock to the network. The change is detected in the 2007 data set, where a change was known to exist, whereas no change was detected in the 2005 data set. This evidence suggests that the CUSUM procedure appears effective at detecting real change (Leavenworth 2007) as opposed to typical random variations between time periods (Leavenworth 2005).

6.4 Al-Qaeda

The Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University created snapshots of the annual communication between members of the al Qaeda organization from its founding in 1988 until 2004 from open source data (Carley, 2006). The data is limited in that I do not know the type, frequency, or substance of the communication and all links are non-directional, meaning I do not know who initiated communication with whom. Finally, the completeness of the data is uncertain since it only contains information available from open sources. The number of nodes in the network range from 62 in 2004 to as many as 260 in 2001. The data is unique in that it provides a network picture of a robust network over standard time-periods of one year.

Using the network snapshots for each year time-period, the average social network measures were calculated and plotted for betweenness, closeness, and density. Each of these measures increased from 1988 until 1994, and then leveled off. There are many possible reasons for this burn-in period, such as the quality of our intelligence gathering on al Qaeda and the rapid development and reorganization of a fast growing

organization. In al Qaeda's early years, access to the infant organization may have been limited, as well as the resources devoted to tracking a small, new, and relatively unaccomplished terrorist network. The organization itself may have also been changing drastically during its first years by actively recruiting new members, and shifting its structure to accommodate new resources and infrastructure. For this reason, the averages for each measure and standard deviation were calculated over the five years that follow the burn-in period that ended in 1994. The CUSUM control chart was then used to monitor the network from 1994 to 2004. Figure 34 is a snapshot of the Al-Qaeda social network.

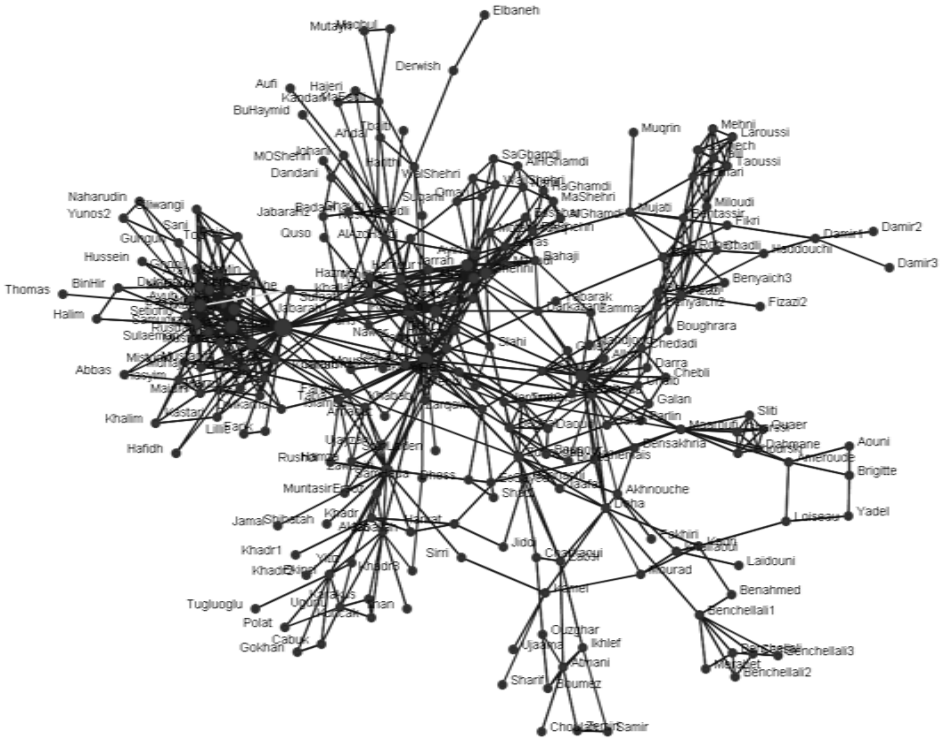


Figure 34. Monitored al Qaeda Communication Network for Year 2001.

The Al-Qaeda data set offered data with more nodes that were aggregated over a much larger time period. At the same time, I was able to identify at least one major event in Al-Qaeda's history. The question was asked, "can we identify September 11 from the social network?" Perhaps more importantly, "can we identify the point in time when the organization changed and began to plan the attacks?" Figure 35 shows the CUSUM statistic for the average betweenness of the Al-Qaeda network.

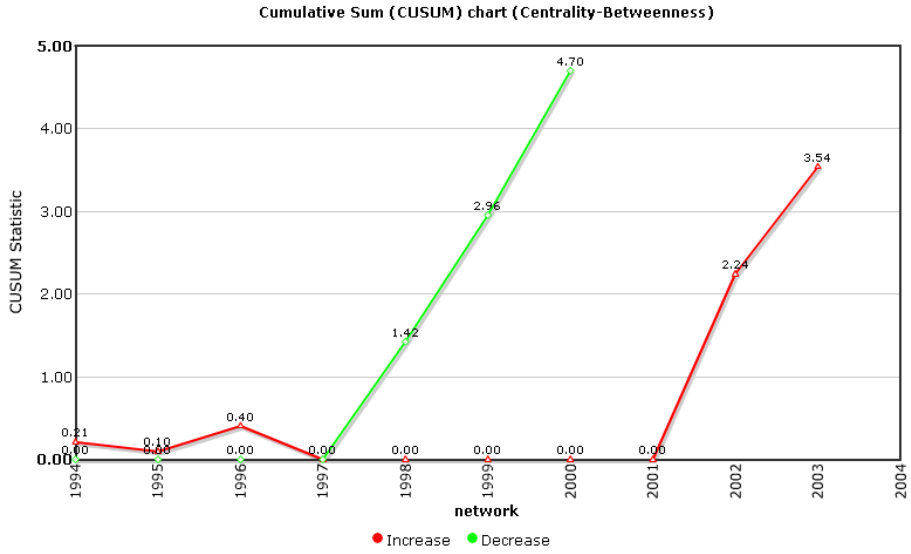


Figure 35. Plot of Betweenness CUSUM Statistic of al Qaeda.

The reference value, *k*, and the control limit, *h*, were set at 0.5 and 3.5 respectively for all of the social network control charts. The value of $k = 0.5$ is set to optimize the chart to be the most statistically powerful in detecting a one standard deviation shift in the mean value of the measure of interest. The value of $h = 3.5$ corresponds to a false positive rate of 1% (see Appendix C). This would be equivalent to a false positive once per century on average.

The most likely time that the change occurred is 1997. To understand the cause of the change in the Al-Qaeda network, an analyst should look at the events occurring in Al-Qaeda's internal organization and external operating environment in 1997.

Several very interesting events related to Al-Qaeda and Islamic extremism occurred in 1997. Six Islamic militants massacred 58 foreign tourists and at least four Egyptians in Luxor, Egypt (Jehl, 1997). United States and coalition forces deployed to Egypt in 1997 for a bi-annual training exercise were repeatedly attacked by Islamic militants. The coalition suffered numerous casualties and shortened their deployment. In early 1998, Zawahiri and Bin Laden were publicly reunited, although based on press release timing, they must have been working throughout 1997 planning future terrorist operations. In February of 1998, an Arab newspaper introduced the "International Islamic Front for Combating Crusaders and Jews." This organization established in 1997, was founded by Bin Laden, Zawahiri, leaders of the Egyptian Islamic Group, the Jamiat-ul-Ulema-e-Pakistan, and the Jihad Movement in Bangladesh, among others. The Front condemned the sins of American foreign policy and called on every Muslim to comply with God's order to kill the Americans and plunder their money. Six months later the US embassies in Tanzania and Kenya were bombed by Al-Qaeda. Thus, 1997 was possibly the most critical year in uniting Islamic militants and organizing Al-Qaeda for offensive terrorist attacks against the United States.

These findings should be interpreted with caution. The data was largely collected retrospectively and is most likely incomplete. The type of findings demonstrated in this example, however, show how SNCD can enable an analyst to look inside of the decision cycle of an organization. If intelligence analysts were able to collect social network data on an organization, the detection of changes in the organizational behavior of the group might provide early warning of some action carried out by the group.

6.5 Johnson Year C Wintering Over Data

The fifth longitudinal network data set was collected at the Amundsen-Scott South Pole Station (Johnson, Boster, and Palinkas, 2003). This is an American polar station run by the National Science Foundation (NSF), located at 90^{o} south latitude in the Ant-arctic. The station is used to conduct scientific research in several fields. Data were collected on the social interaction between crew members over each of three wintering over periods. During the wintering over months the station is completely isolated. This creates a well-bounded group where the interactions are largely free from outside influence.

There are 22 individuals in the wintering over group. These individuals include contractors that support the facilities and NSF scientists. There were four females and 18 males. They began training as a group in August in the U.S. and actually arrived at the station the following October. They remain at the station until November of the following year. They are not permitted to remain at the station for two consecutive winters. The data was collected in the 1990s. The actual year of the data is not reported in order to protect the identities of the respondents. During the winter temperatures can reach as low as -119^{o}F, making flights to the station next to impossible. For eight and a half months between mid-February and the end of October, the station is completely isolated. The nearest American base is McMurdo Station, 800 miles away. On the 15^{th} day of each month beginning in March, the station physician collected social network data by questionnaire. Respondents were asked to self rate their social interactions with each of the 21 other group members on a scale of 0 to 10, with 0 representing no interaction.

During the wintering over period a change was introduced into the network. Johnson reports that "events transpired in Year C to undermine the ability of the formal leader to maintain his/her informal leadership role." He also reports that the only expressive leader disappeared sometime in the middle of the winter "due in part to harassment by a marginalized crewmember." Figure 36 shows the first recorded network in the data set, which was collected on 15 March. Figure 37 shows the last recorded network, which was collected on 15 October. The nodes are sized by betweenness in both of the figures. Betweenness is a network measure of power (Krackhardt, 2008). The change is a shift in power and influence among the agents as Agent 8 gives up power, and influence in the organization to nodes A18, A19, A11, and A14, who were on the periphery of the network in March.

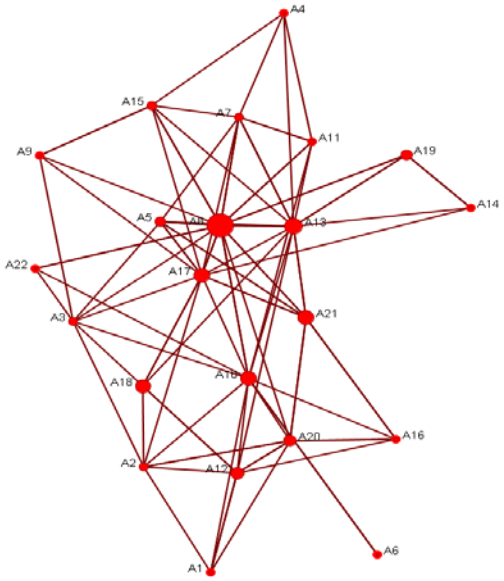


Figure 36. March (Time 1), Year C Winter-over Data.

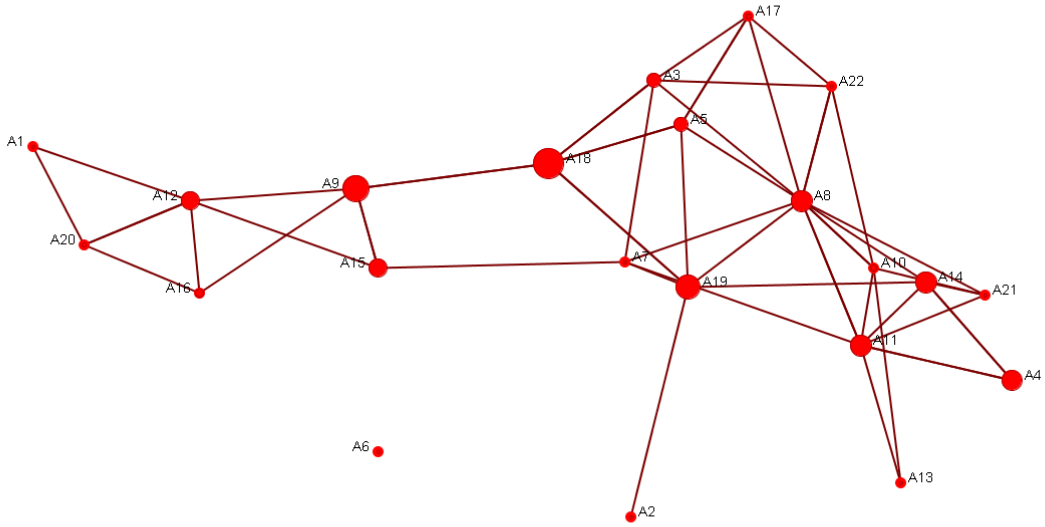


Figure 37. October (Time 8), Year C Winter-over Data.

SNCD should be able to detect the change in organizational behavior in this network. In March, there is only one node with a moderately high betweenness score, A8. By October, Nodes 9, 18, and 19 have higher betweenness within the network than Node 8. When did this change occur? The CUSUM algorithm with $k = 0.5$ and $h = 3.5$ is used. Figure 38 shows a plot of the CUSUM statistic over the eight time periods. The change is not noticed until October in this data.

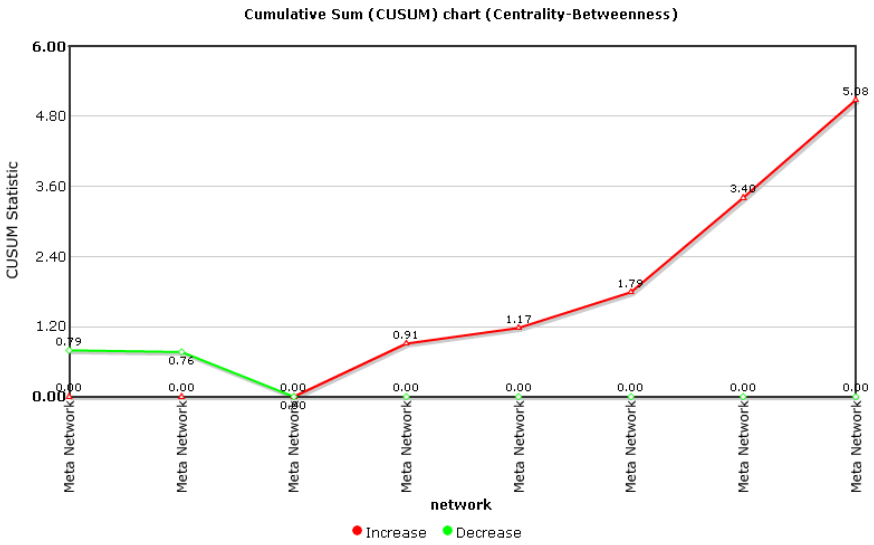


Figure 38. CUSUM Statistic for Winter-over Year C Data.

The most likely time the change occurred was the last time the statistic was 0, so sometime between time 3 and 4 above, which corresponds to May-June. If we look at the networks in May and in June we can see the change. Figures 39 and 40 display the May and June networks respectively.

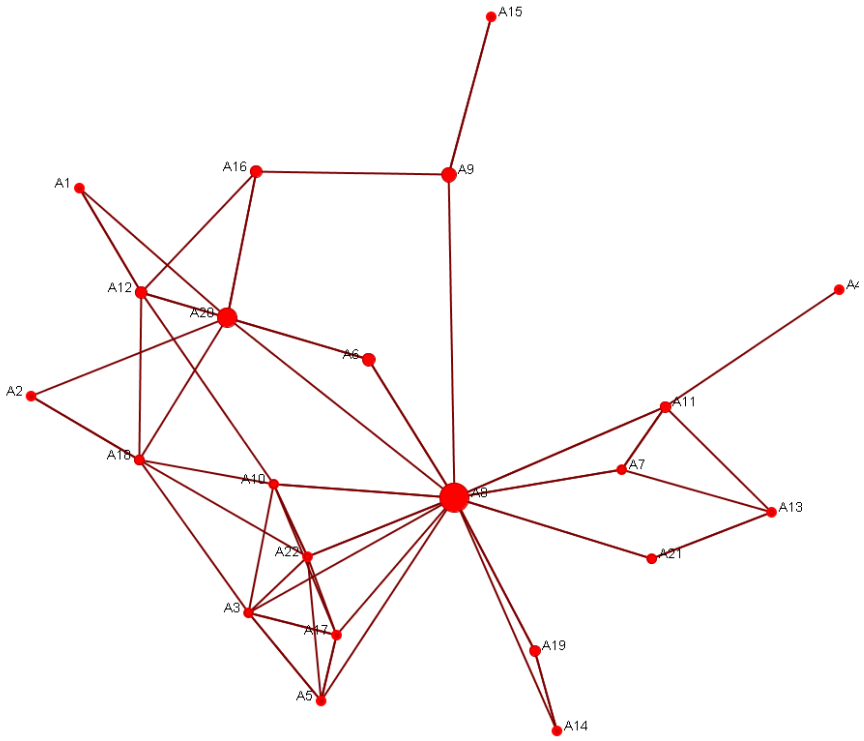


Figure 39. May (Time 3), Year C Winter-over Data

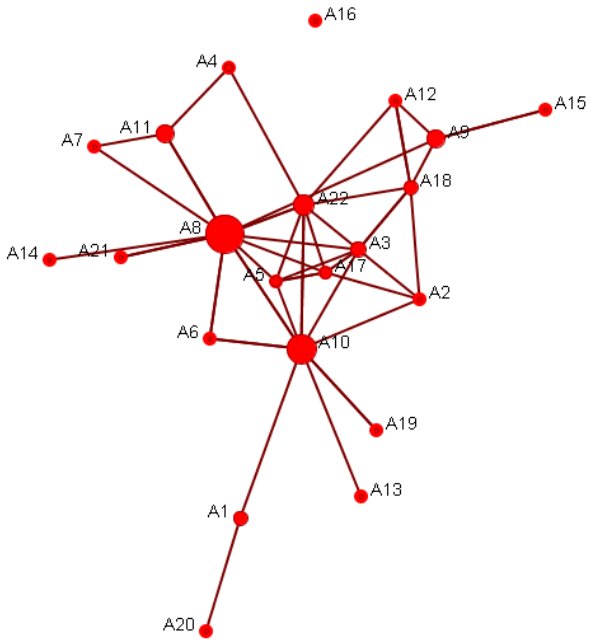


Figure 40. June (Time 4), Year C Winter-over Data

It can be seen that it is between these two networks that other nodes (A10, A22) take significantly more leadership within the network. A8 still maintains the most power within the network. We can see that the other nodes that hold power within the network shift from month to month, however A8 no longer maintains a monopoly on betweenness following the change point. Figure 41 shows the network of the 5th time point.

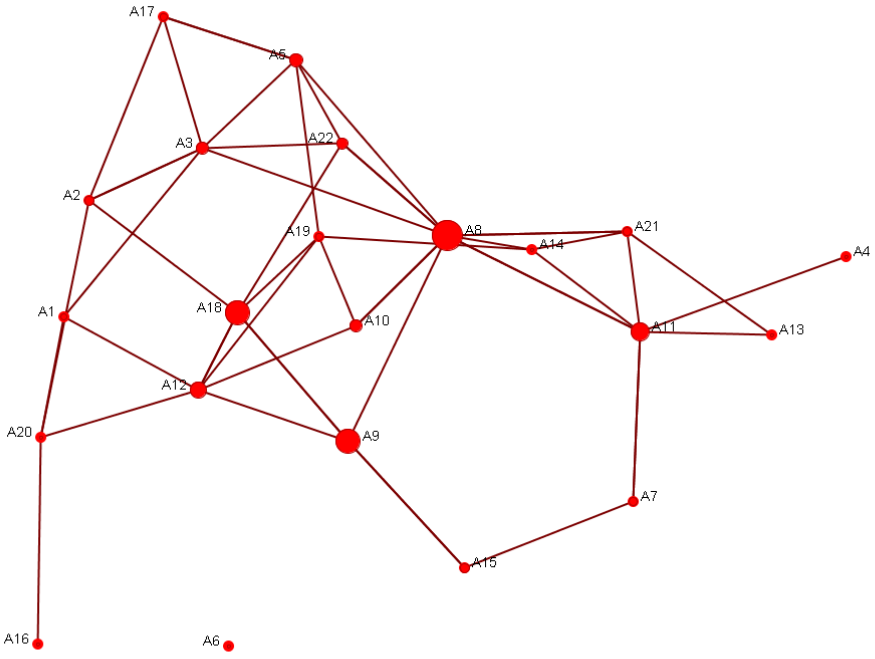


Figure 41. July (Time 5), Year C Winter-over Data.

It is not until September that A18 surpasses A8 in betweenness. This corresponds to the time just before the change is detected. It should be noted however, that the cause of this change in the organization is more likely to have occurred between May and June, however, the real effects of the change are not realized for some time.

The important concept illustrated in this example is that SNCD can be used by social scientists to study network change and evolution. The social network change detection provided the statistically based insight of where to look for change. This makes isolating the cause of real change more objective and scientific. In this case the change was likely due to the harassment and subsequent isolation of an expressive leader in the group. This incident is likely to have occurred between 15 May and 15 June. A social scientist recording the data may not be aware of such an event if the respondents do not report the harassment. Using SNCD, the social scientist would be aware of the change and be able to ask the respondents more questions about what occurred during the estimated change point, in this case 15 May to 15 June. Therefore, SNCD is not only useful for identify *that* a change occurred. It is equally useful to detect *when* a change occurred.

6.6 Johnson Year A Wintering Over Data

The sixth longitudinal network data set was also collected at the Amundsen-Scott South Pole Station in the 1990s (Johnson, Boster, and Palinkas, 2003), in a different year than the fifth data set. Data were again collected on the social interaction between crew members over a wintering over period.

There are 28 individuals in the Year A wintering over group. There were nine females and 19 males. They also began training as a group in August and arrived at the station the following October. They remain at the station until November of the following year. On the 15th day of each month beginning in March, the station physician collected social network data by questionnaire. Respondents were asked to self rate their social interactions with each of the 27 other group members on a scale of 0 to 10, with 0 representing no interaction. The first (March) and last (October) networks are shown in Figures 42 and 43 respectively. The nodes are again sized based on the betweenness value of the node.

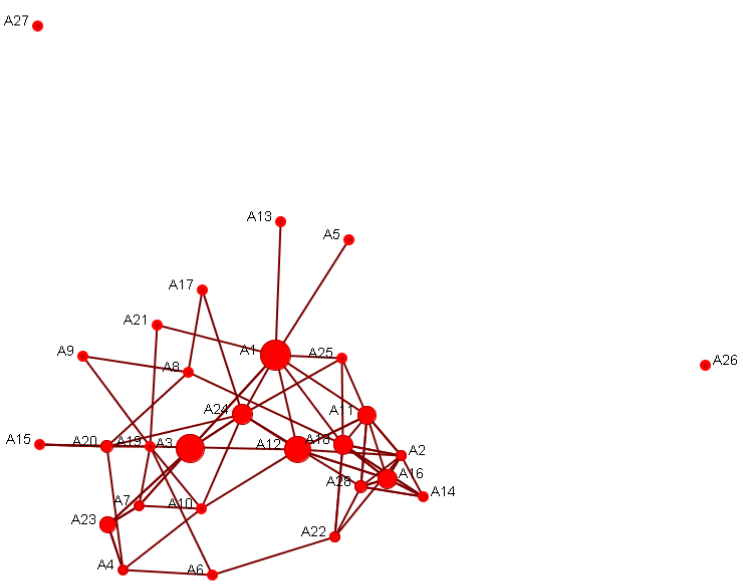


Figure 42. March (Time 1), Year A Winter-over Data

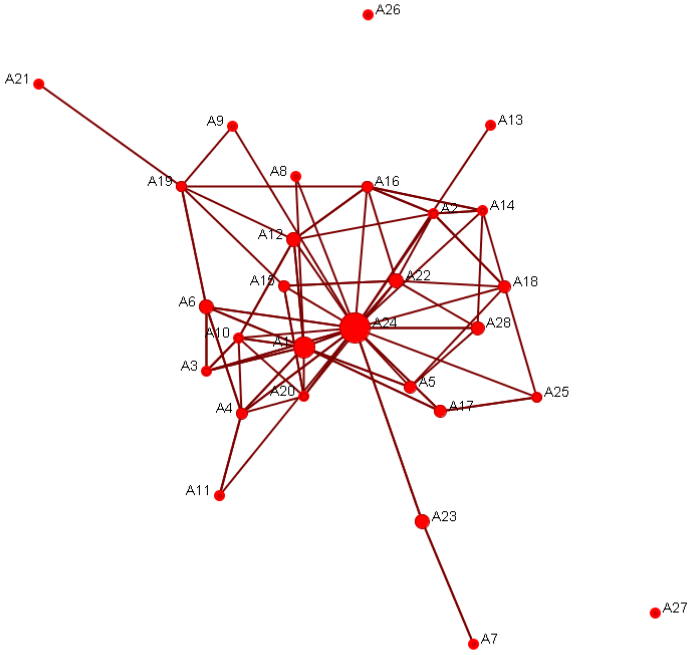


Figure 43. October (Time 8), Year A Winter-over Data

In the Year A data, a statistically significant change occurs between 15 May and 15 June. The change is a shift in power and influence among the agents as Agent 24 assumes more leadership, power, and influence in the organization. In March, there are several nodes (6-8) with a moderately high betweenness score. By October, Node 24 completely dominates the betweenness within the network. Node 24 was not one of the top three nodes highest in betweenness in March, but clearly is an influential member of

the organization by October. For this data set, the goal of SNCD is to detect when this shift in power occurs.

The CUSUM procedure is applied to the average betweenness value of the Year A data using parameters $k = 0.5$ and $h = 3.5$, corresponding to a risk of false positive of 1%. Figure 44 is a plot of the CUSUM statistic over time. The most likely time the change occurred was the last time the statistic was 0, so sometime between time 3 and 4, which corresponds to May-June. If we look at the networks in May and in June, Figures 45 and 46 respectively, the change can be seen. In the networks between March and May, several nodes compete for influence in the network. Beginning with the June network, Agent 24 dominates the betweenness centrality measure, which is a measure of power in the network. Agent 24 continues to dominate the network for the remainder of the wintering over period.

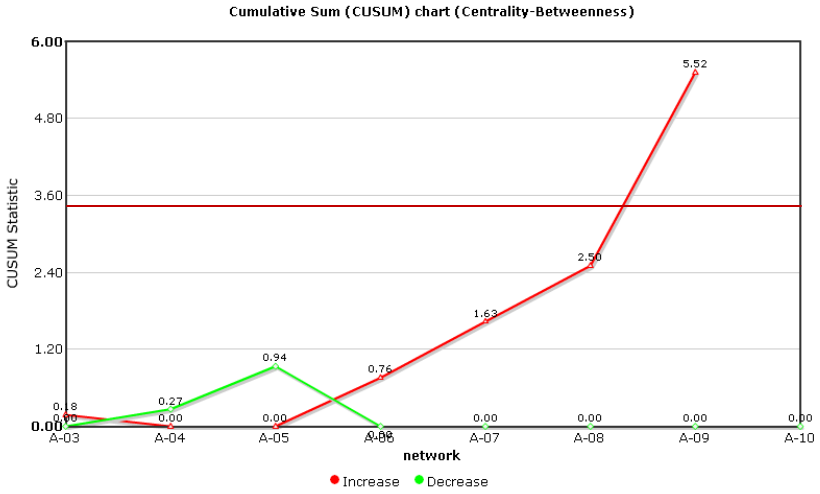


Figure 44. CUSUM Statistic for Winter-over Year A Data.

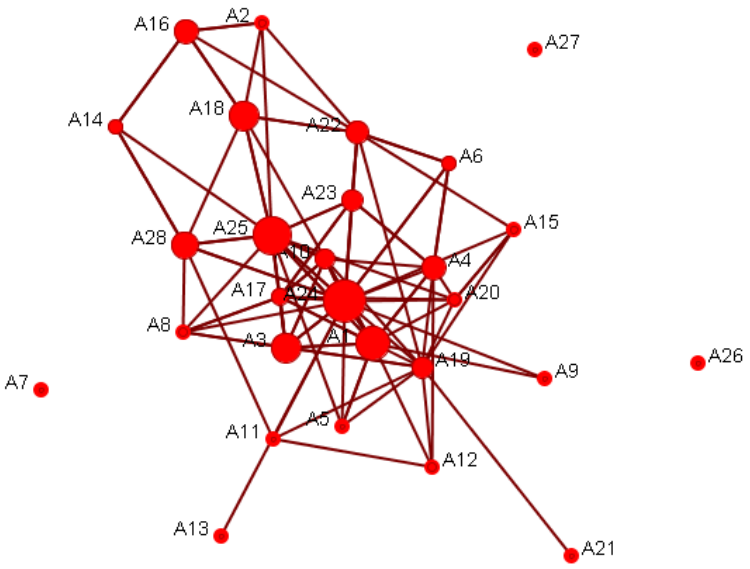


Figure 45. May (Time 3), Year A Winter-over Data

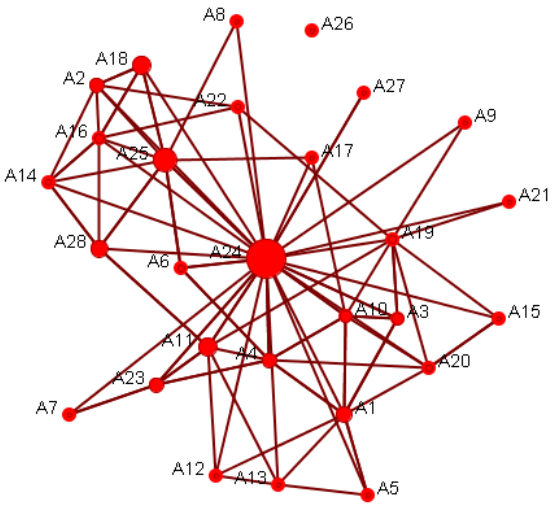


Figure 46. June (Time 4), Year A Winter-over Data

Johnson et al. (2003) identified this change as one leading to a greater consensus among individuals in the group on who played an instrumental leadership role. They speculate that positive deviant roles, such as comedians contribute to this change as the comedians play pranks on group members. Unfortunately, Johnson et al. lacked a statistical approach for estimating the time change occurred. Although they discuss the change in the group between the first and last time periods, they do not offer any insight into when the change occurred or what specific events may have contributed more significantly to the emergent group dynamics. Perhaps if they applied this approach and looked for a potential cause of change, they might find that practical jokes played

between 15 May and 15 June led to a change in the group dynamics. This example, again illustrates the importance of estimating *when* a change occurs in addition to determining *that* a change occurs in longitudinal social network analysis.

6.7 Johnson Year B Wintering Over Data

The seventh longitudinal network data set was also collected at the Amundsen-Scott South Pole Station in the 1990s (Johnson, Boster, and Palinkas, 2003), in a different year than the fifth and sixth data sets. Data were again collected on the social interaction between crew members over a wintering over period.

There are 27 individuals in the Year B wintering over group. There were seven females and 20 males. They also began training as a group in August and arrived at the station the following October. They remain at the station until November of the following year. On the 15th day of each month beginning in March, the station physician collected social network data by questionnaire. Respondents were asked to self rate their social interactions with each of the 26 other group members on a scale of 0 to 10, with 0 representing no interaction. The first (March) and last (October) networks are shown in Figures 47 and 48 respectively. The nodes are again sized based on the betweenness value of the node.

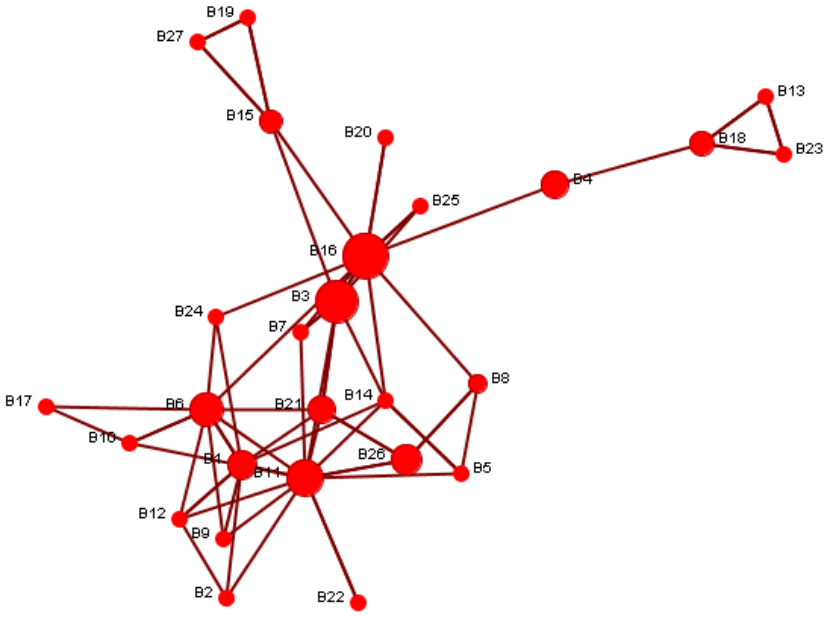


Figure 47. March (Time 1), Year B Winter-over Data

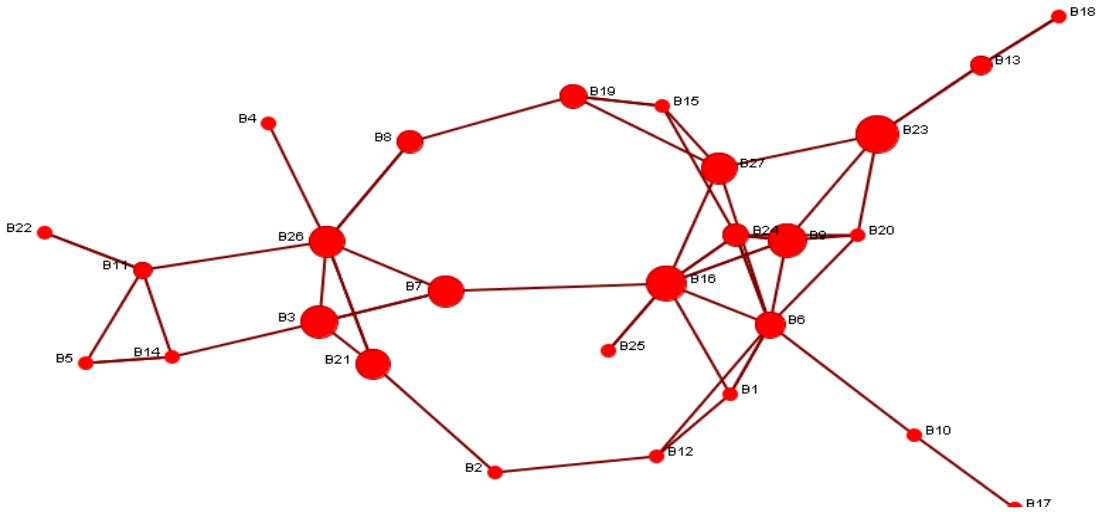


Figure 48. October (Time 8), Year B Winter-over Data

A statistically significant change occurs in the year B data between 15 May and 15 June. The change is where the respondents form the three separate subgroups identified by Johnson et al (2003).

The CUSUM procedure is applied to the average betweenness value of the Year B data, again using parameters $k = 0.5$ and $h = 3.5$, corresponding to a risk of false positive of 1%. Figure 49 is a plot of the CUSUM statistic over time. The most likely time the change occurred was the last time the statistic was 0, so sometime between time B05 and B06, which corresponds to May-June.

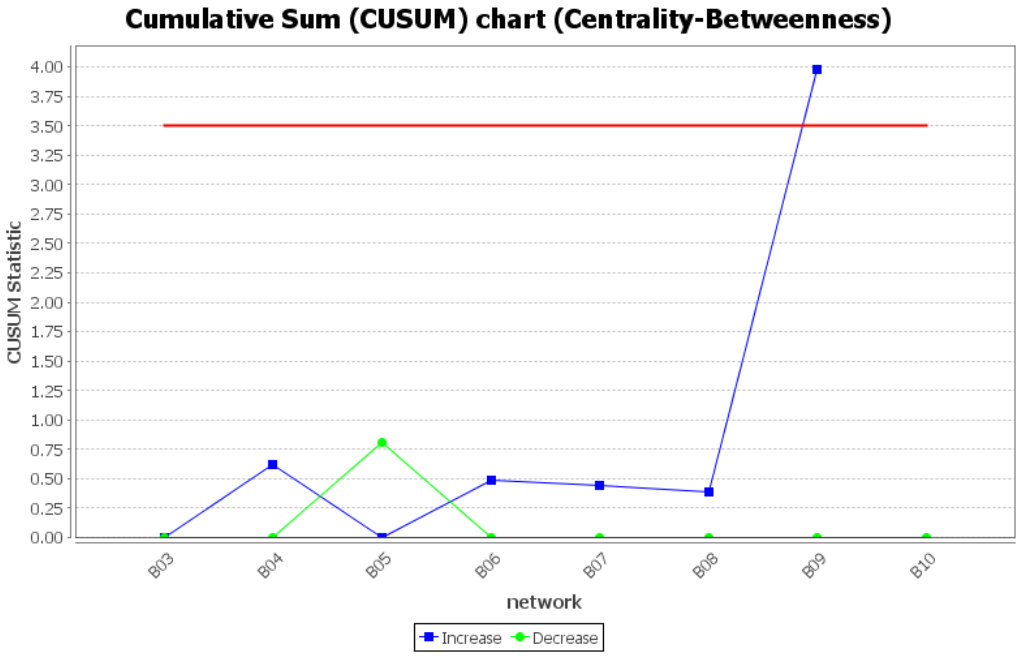


Figure 49. CUSUM Statistic for Winter-over Year B Data.

6.8 IkeNet 2

IkeNet is a five year research project funded by the U.S. Army Institute for the Behavioral and Social Sciences to collect longitudinal network data on email activity. The participants are mid-career Army officers in a one-year graduate program at Columbia University. The participants all live on the West Point military installation and attend most of their courses at West Point. Following graduation, most of these officers assume duties as tactical officers, responsible for military training and discipline at the U.S. Military Academy (USMA).

The participants were all given a BlackBerry. They consented to allow me to monitor the header information of their sent email traffic. The header information includes the TO, FROM, CC, BCC, Subject, and Date-Time. The data was collected by installing a client side visual basic patch in their Microsoft Outlook, that would compile a spreadsheet from their sent mail folder and email to me daily. Another custom software plug-in to my Outlook allowed me to compile and parse the data efficiently. The email activity was divided into calendar weeks from Sunday through Saturday. Networks were constructed where the nodes were the officers in the ELDP program and the links connected a source node to a target node, weighted by the number of sent emails. A total of 46 networks were collected beginning 20 May 2007 and ending 4 April 2008. Unfortunately, the first 13 networks contained missing data due to technical difficulties in implementing the Outlook patch. Therefore, we begin our analysis with week 14 data.

There were 22 participants who volunteered to participate in the research. There were 3 females and 19 males. This experiment was approved by the West Point Institutional Review Board for Human Subject Experimentation. Figure 50 shows the IkeNet2 network for week 14.

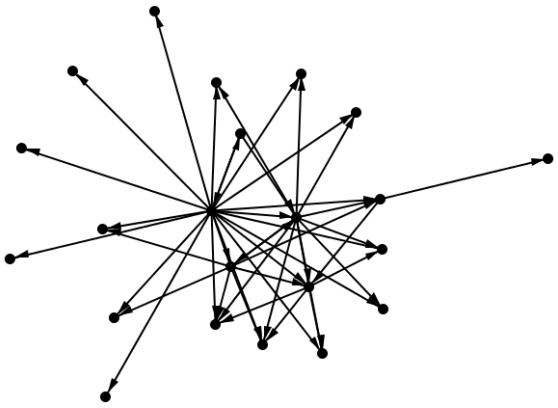


Figure 50. IkeNet 2, Week 14

The CUSUM procedure is applied to the average betweenness of the networks over time. The value of $k = 0.5$ and $h = 3.5$, which corresponds to a false positive rate of 1%. Figure 51 shows the CUSUM statistic plotted for the IkeNet 2 data. It can be seen in the figure that the CUSUM chart indicates that there may be a change at Week 26, and that the likely time the change occurred was Week 25.

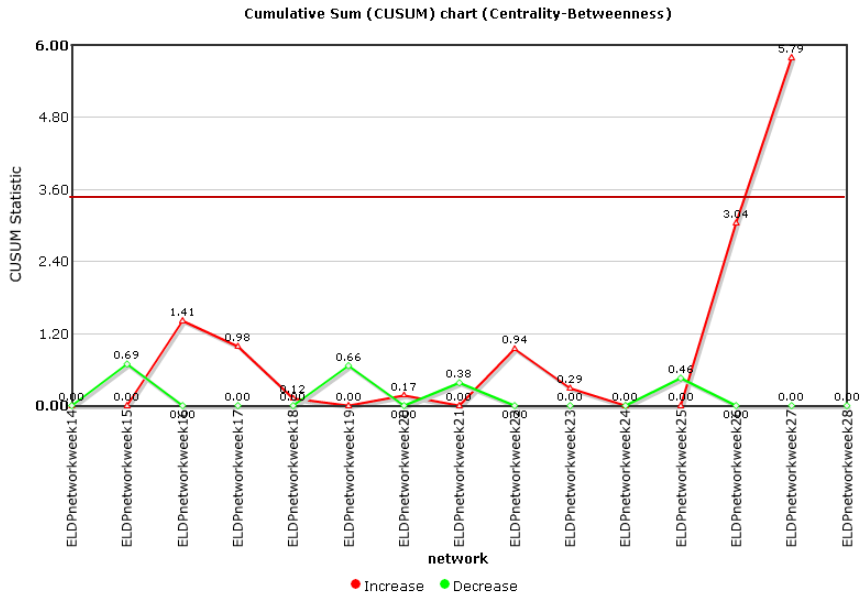


Figure 51. IkeNet 2, CUSUM applied to Average Betweenness.

Week 25 was Army-Air Force week at USMA. This is perhaps the busiest week at USMA next to graduation. The U.S. Air Force Academy played football in Michie Stadium at West Point on the Friday of Week 25. The ELDP officers were involved in planning a large tailgate event this week. In addition, there were an unusually large number of academic requirements due earlier in the week. It is reasonable that this is a significant change point in the ELDP network. Figure 52 is the network diagram for Week 25. It can be seen that the Week 25 network is more distributed than the Week 14 network, which had a couple key influential people with others having a peripheral role.

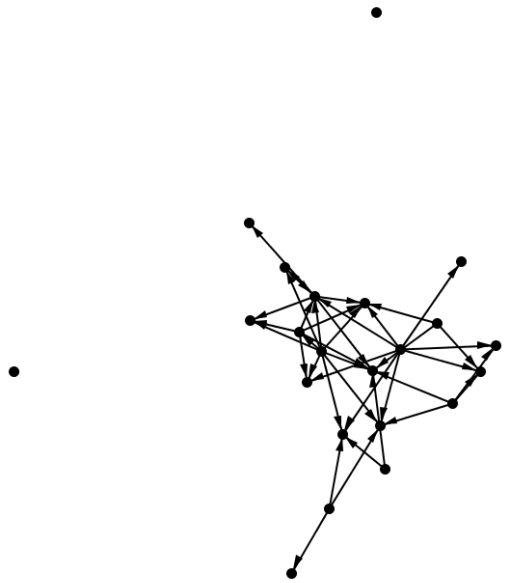


Figure 52. IkeNet 2, Week 25.

The CUSUM procedure is reset using the same parameters and run on subsequent data. Figure 53 shows a plot of the CUSUM applied to the average betweenness, beginning in Week 26. The CUSUM detects another change that likely occurred beginning Week 28, which corresponds to Thanksgiving week. This was also the last week that the ELDP students had any graded assignments due for the term.

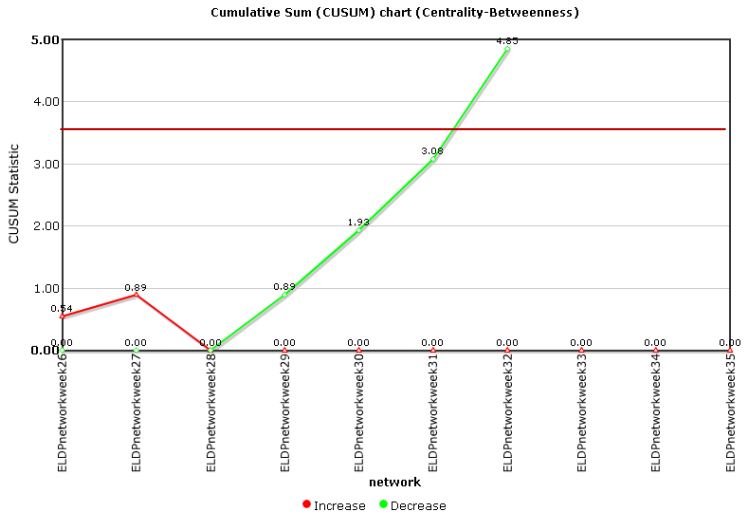


Figure 53. IkeNet 2, CUSUM Week 26 Week 35

The ELDP students began Christmas break during Week 32. Figure 54 is a network image for the first full week of Christmas break. The network is very sparse. This is expected while the participants are on vacation. We therefore, wait until they are back in school to continue change detection.

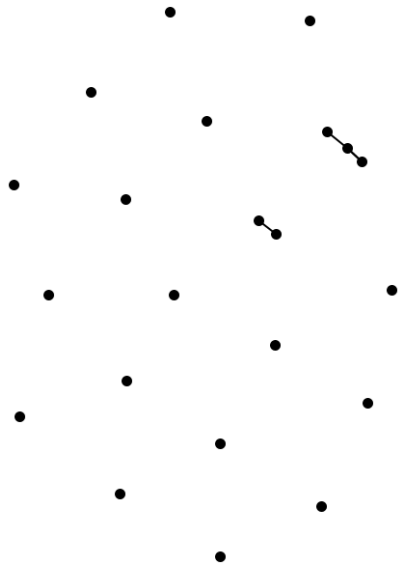


Figure 54. IkeNet 2, Week 33.

The CUSUM is again applied to the IkeNet 2 data, beginning with week 36, which is the first week back after Christmas break. Figure 55 shows a plot of the CUSUM statistic over time. Although the chart does not signal a potential change, it appears that it is heading toward a signal if there were another week or two of data. The likely change point if this chart were to signal would be the last full week before they took their comprehensive exam for their graduate program.

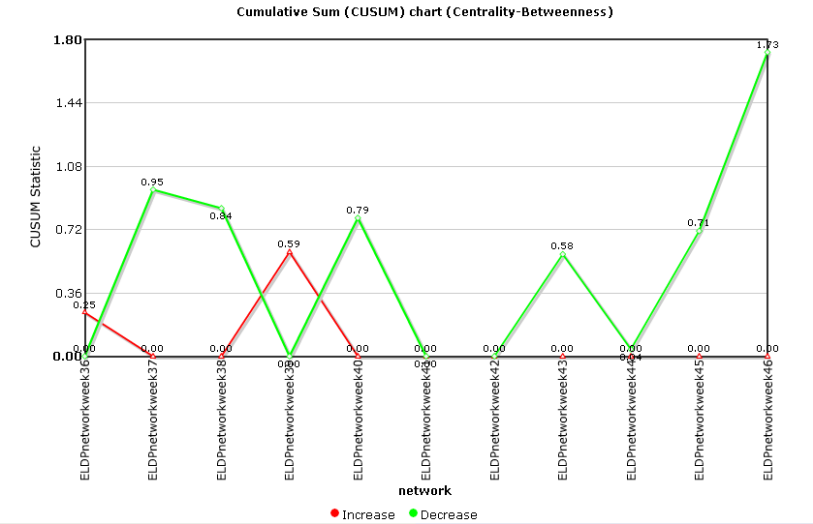


Figure 55. IkeNet 2, CUSUM for Week 36-Week 46.

IkeNet 2 is another successful example of SNCD. The events during Army-Air Force week were detected, as well as the change in network behavior for Christmas break. The final change detected was for their completion of the comprehensive exam for their graduate program.

6.9 IkeNet 3

The IkeNet 3 data are data collected on the email interaction between three distinct groups at the U.S. Military Academy (USMA) at West Point. This is the third in a five year research project funded by the U.S. Army Institute for the Behavioral and Social Sciences to collect longitudinal network data on email activity. The first sub-group consists of undergraduate cadets in a regimental chain of command. The chain of command is comprised of 5 females and 19 males between the ages of 21 to 25. The chain of command is also comprised of many ethnicities to include African Americans, Hispanics, Pacific Islanders, and Asian Americans.

The second group is composed of 14 mid-career Army officers in a one year graduate program run jointly by Columbia University and the USMA. This program is called the Eisenhower Leadership Development Program (ELDP). Following their graduation in May, they begin duties as tactical officers at the US Military Academy, responsible for the military training and discipline at the Academy. None of these officers are acquainted prior to entering the program. They have no formal chain of command assigned to them. There are 2 females and 12 males. There are varying ethnicities, representative of the demographics at the USMA. Some of these officers may interact with cadets in the regimental chain of command, serving as mentors to them or serving as a faculty representative to one of the many extra-curricular clubs at USMA

The third group consists of faculty and staff at the USMA that conduct research in network science. Several faculty and staff have begun conducting active research in the area of network science. 30 of the staff and faculty involved with network science were given a BlackBerry. There were 7 females, 23 males all of varying ethnicities. These faculty had limited contact with ELDP members. Several of the cadets from the regimental chain of command were working on network science related senior research theses and had regular communication with the faculty.

E-mail traffic was collected to create networks that describe communication between members of the three different groups. All individuals consented to participate in this experiment, which involved monitoring their email, cell phone, and text messaging data. This experiment was approved by the West Point Institutional Review Board for Human Subject Experimentation. Only the header information (header information includes the to, from, cc, bcc, subject, date, and message id) of emails and phone traffic sent and received by the participants were monitored and collected. The message body and attachments of emails and text messages were not collected due to privacy concerns. With the help of the Directorate of Information Management (DOIM) at West Point the Microsoft Email Exchange server was synced with the BlackBerries so that a participant could receive and send emails from their BlackBerry. Email data was collected directly from the server instead of from each clients' computers and BlackBerries due to the greater quality of data found at the server (McCulloh et al, 2008; Appendix D). Data was parsed from the email log file using the Organizational Risk Analyzer (ORA) (Frantz

and Carley, 2008). All identifying participant data was anonymized to protect the privacy of the participants.

Daily networks were created for the time period 1 September -31 December 2008. The nodes in the network are the individuals in each of the three groups. A weighted link connects nodes, where the weight corresponds to the number of emails exchanged from a source node to a target node. Figure 56 shows the network for 3 September, which is the Wednesday after Labor Day. The tightly clustered group in the right side of the figure is the ELDP group of officers.

The cadets in the regimental chain of command were not actually given BlackBerries until 18 September after they were in their duty position for one full month. The intent was to see if the introduction of the BlackBerry could be detected with SNCD methods. To make this experiment even more challenging, the faculty and ELDP groups are left in the network. Therefore, the change imposed on the network is the introduction of enhanced communication for approximately one third of the network.

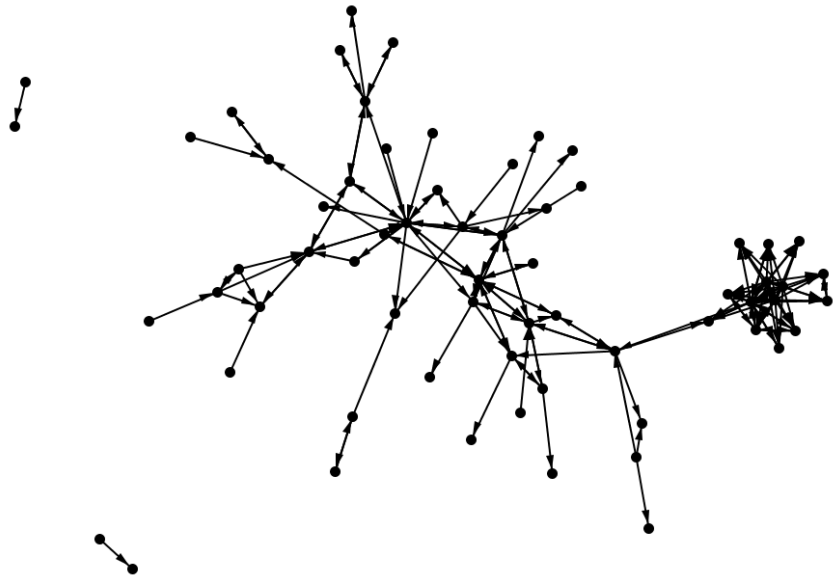


Figure 56. IkeNet 3, 3 September 2008.

The average of the betweenness of nodes in the network is plotted in Figure 57. I show this measure for the IkeNet data, so that the issue of periodicity might be observed. The volatility in the measure is partially due to the weekly patterns of email activity within the group. This periodicity introduces additional noise into the data, making it more difficult to detect change.

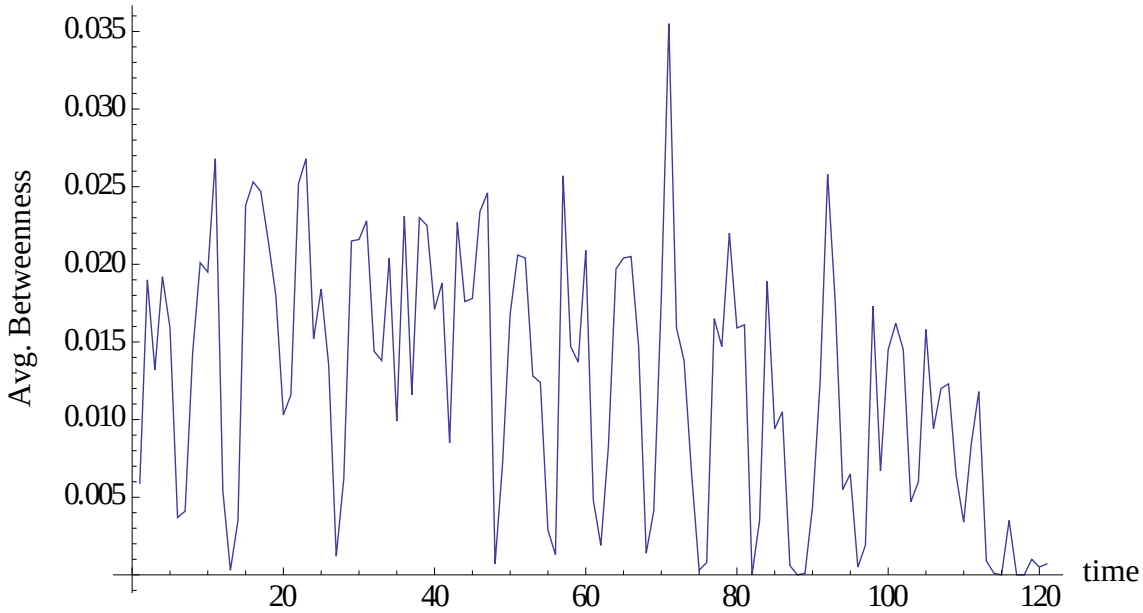


Figure 57. IkeNet3 Average Betweenness 1 Sep - 31 Dec 2008.

The longitudinal network data is monitored in real time. The first 10 networks corresponding to 1-10 September 2008 are used to establish the typical behavior of the network. Any change detected over time is really detecting networks that are a statistically significant departure from the first 10 networks in the data set. I expect periodicity in the network data, so I apply the Fourier Transform. The frequency plot of the data is shown in Figure 58. Dominant frequencies that are greater than two standard deviations from the mean frequency are shown in Figure 59.

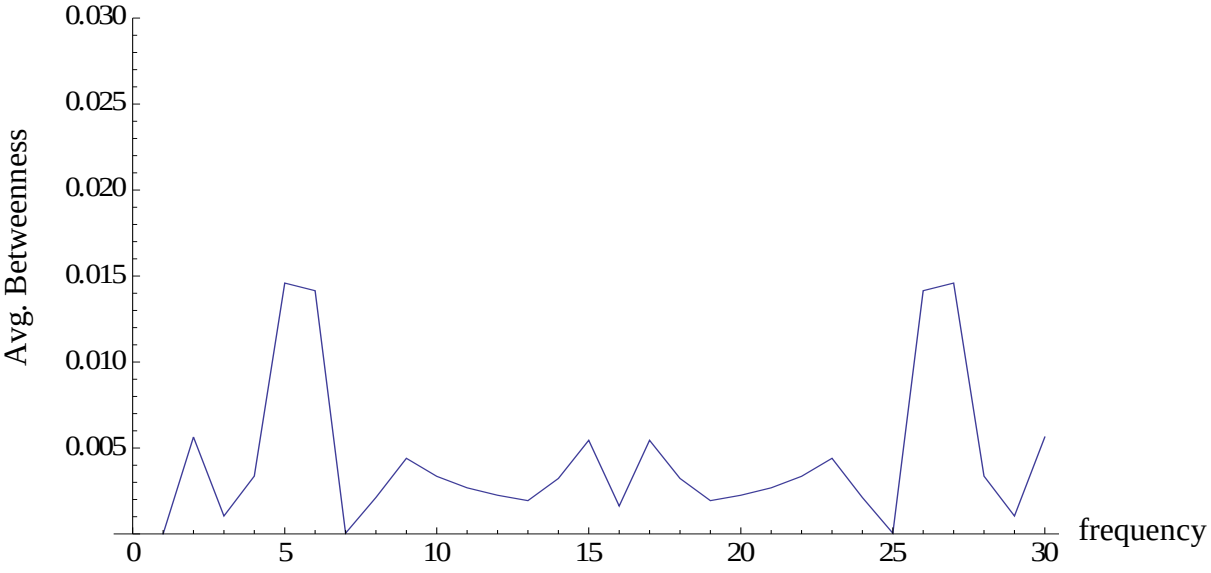


Figure 58. IkeNet 3, Fast Fourier Transform of Average Betweenness.

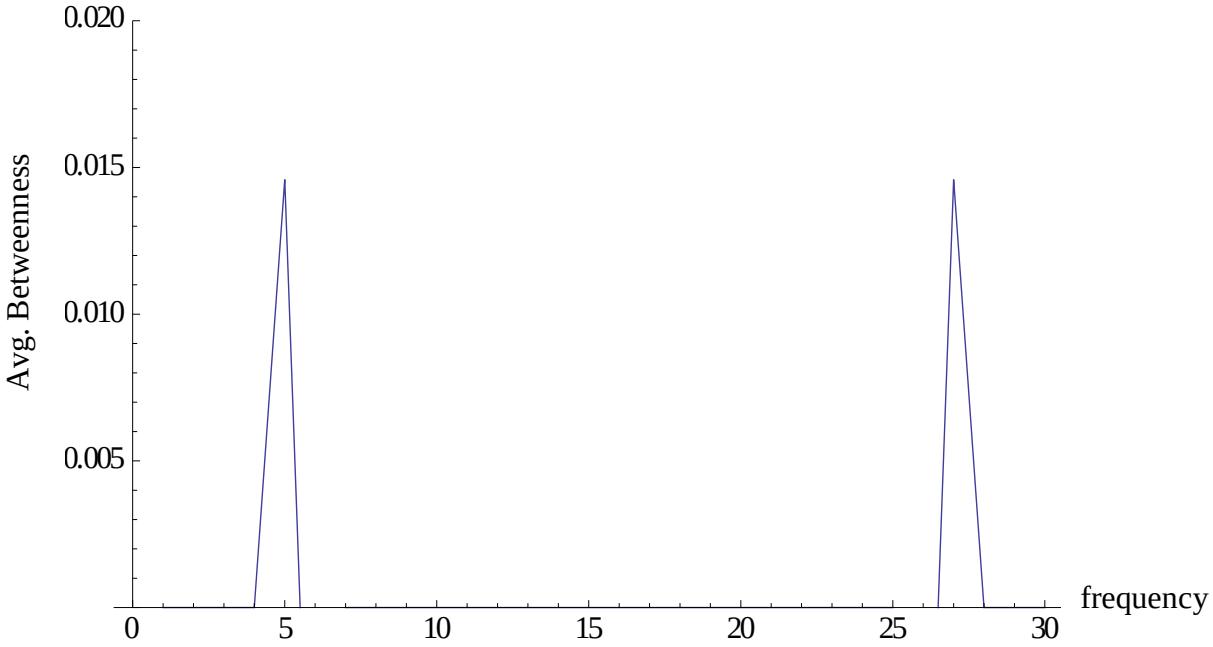


Figure 59. IkeNet 3, Dominant Frequencies of Average Betweenness.

The periodicity is determined by applying an inverse Fourier transform to the dominant frequencies. The resulting period plot is shown in Figure 60. The weekly periodicity can be identified by noticing the peaks and valleys that occur every seven days.

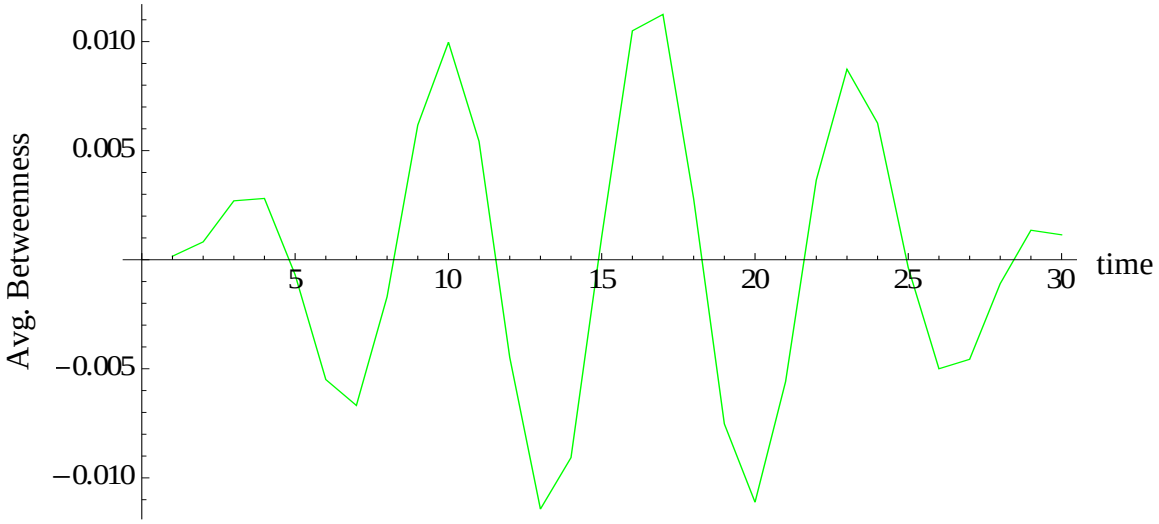


Figure 60. IkeNet 3, Period Plot of Average Betweenness.

The data is filtered by subtracting the periodicity from the average betweenness measure for each time period. The CUSUM procedure is then applied using a $k = 0.5$ and an $h = 3.5$, which corresponds to a false positive rate of 1%. A plot of the CUSUM is shown in Figure 61.

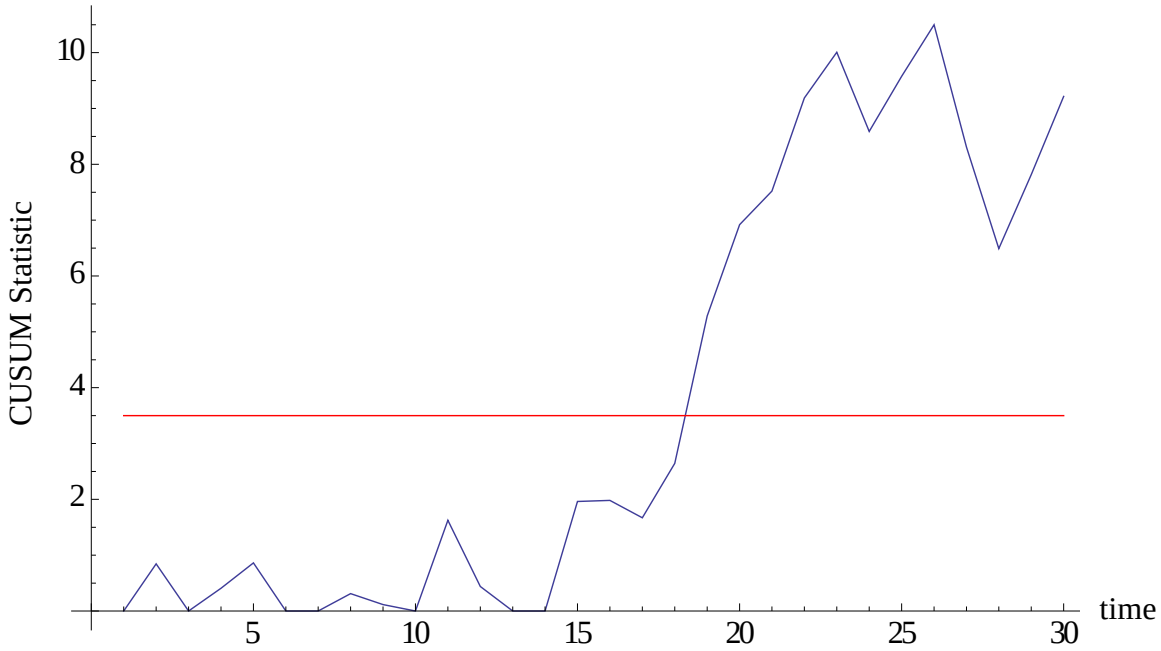


Figure 61. IkeNet 3, CUSUM Statistic on Filtered Average Betweenness.

A change is detected on 19 September. The estimated change point is 14 September. I therefore look for a potential cause for the change in email behavior in the IkeNet group sometime around 14 September. Part of the IkeNet experiment involved studying the effect that Blackberries had on the cadet chain of command. The cadet chain of command assumed their duties on 18 August 2008, therefore, I scheduled them to receive Blackberries on 18 September 2008. This was announced to the chain of command by their regimental commander at their weekly meeting on 14 September. I believe that this is the change. Initially the change occurred as they planned to receive the blackberries, and then continued after they had the device.

I must now look for a new equilibrium, beginning Sunday 21 September. The CUSUM control chart is restarted. The new typical network behavior is estimated from the first 10 days, beginning on 21 September.

I again detect weekly periodicity. Figure 62 shows the frequency plot of the data from the Fast Fourier transform. Figure 63 shows the resulting period plot. Again the weekly periodicity can be observed from the regular seven day peaks and valleys in the period plot.

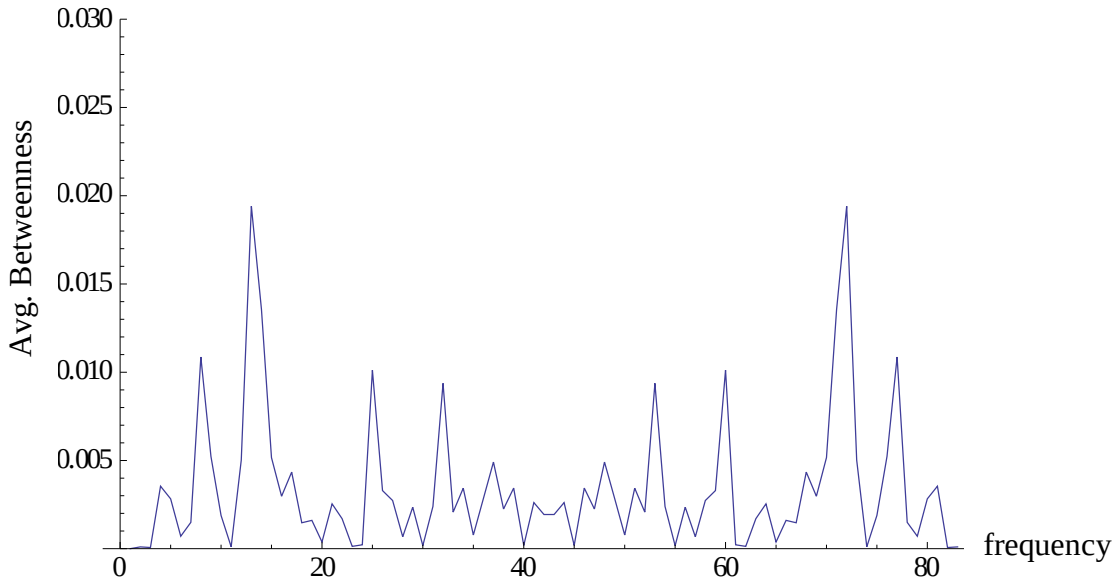


Figure 62. IkeNet 3, Fast Fourier Transform of Average Betweenness after BlackBerry Issue.

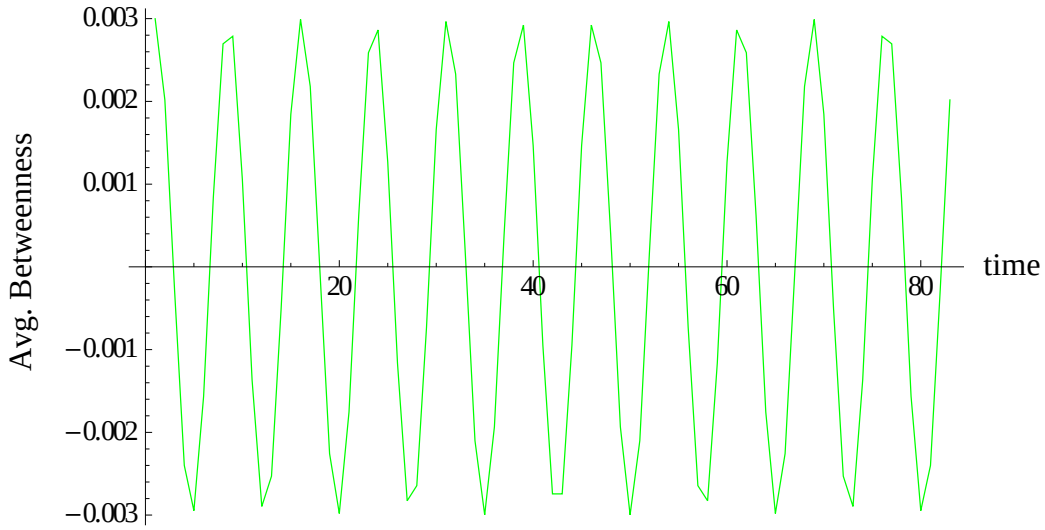


Figure 63. IkeNet 3, Period Plot of Average Betweenness after BlackBerry Issue.

The CUSUM procedure is applied to the filtered data. The CUSUM again uses $k = 0.5$ and $h = 3.5$. This time the procedure detects a decrease in the measure. Figure 64 shows the CUSUM statistic plotted over time.

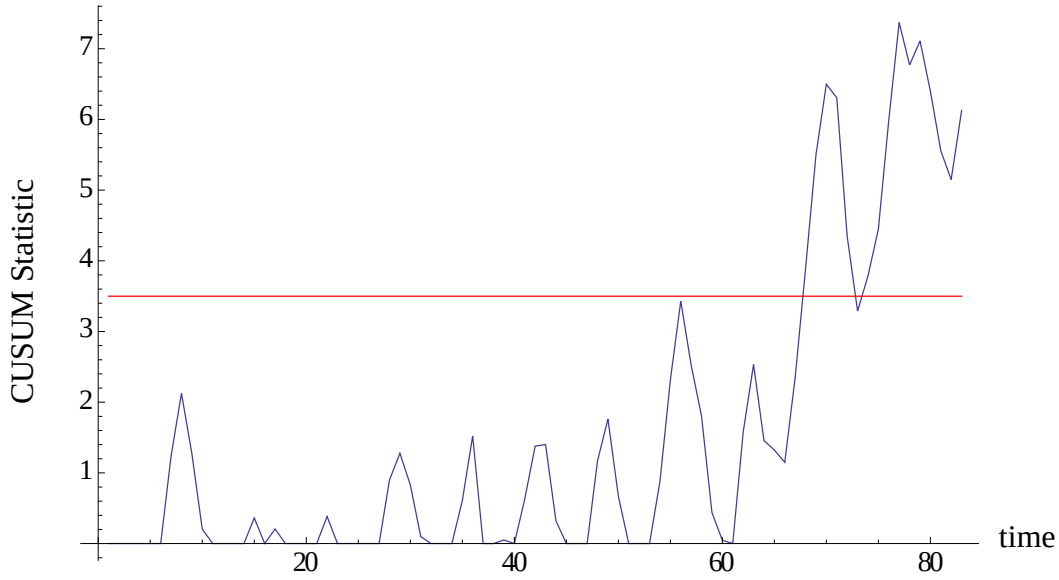


Figure 64. IkeNet 3, CUSUM Statistic of Average Betweenness after BlackBerry Issue.

The CUSUM procedure signals a change in the network on the Friday after Thanksgiving, 2008. The likely time the change actually occurred is 21 September, the Friday before Thanksgiving 2008. This is a reasonable change point in the network as both the ELDP and cadets finish major academic requirements and take a pause from their academics.

I set a new equilibrium for the first 10 days of December and re-run the CUSUM with $k = 0.5$ and $h = 3.5$. We detect a significant change in the network on 24 December. The likely time the change actually occurred was 18 December, which is the day before the last Final exam of the semester. By the afternoon of 19 December, all cadets and faculty were on Christmas leave. Figure 65 shows a plot of the CUSUM statistic over time.

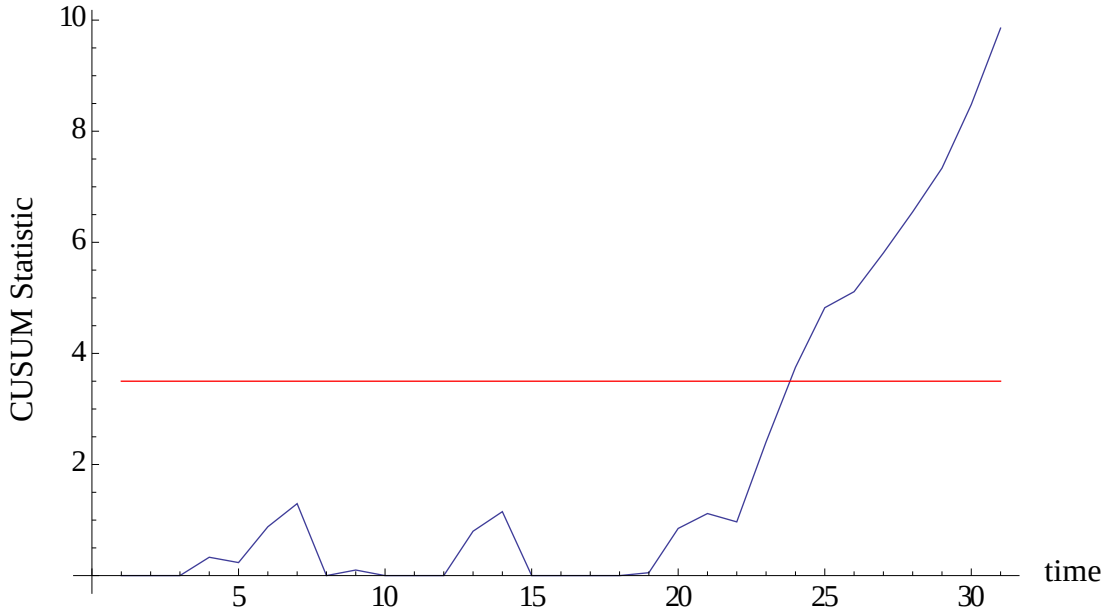


Figure 65. IkeNet 3, CUSUM Statistic of Average Betweenness after Thanksgiving.

This example shows the ability of SNCD to detect several changes in network communication over the course of an academic semester. The issuing of BlackBerries, Thanksgiving, and the conclusion of final exams represent the major significant events of the semester. SNCD was effective in identifying all three changes. In addition, the weekly periodicity inherent in the data was successfully filtered out, making the procedure perform more effectively.

Future IkeNet 3 data for the Spring 2009 semester will also include collecting friendship and trust networks. It will be interesting to apply SNCD to all of the networks collected on the group and compare the presence of any changes across the different networks. The data collected in the Spring will also hopefully allow investigators to look for evolutionary changes in friendship over time.

6.10 Cautionary Note on Findings

The empirical results described in this paper, such as the detection of change in the Al-Qaeda network should be viewed with caution. I present them here purely to illustrate the methodology. Limitations on the data make it difficult to determine the validity of the results; thus, we should simply view these results as showing the promise of this methodology. The Leavenworth data spans only four days and used self-reported survey data, therefore it is not likely that it captured all communication and interaction among officers. The fact that even in this data set we were able to systematically detect a key change suggests the value of the proposed approach. The Al-Qaeda data, was based on open source information. As such it is an incomplete representation of interaction in that terror network. We cannot be sure that we have the entire communication network, or even a true picture of the observed communication network. However, the fact that our technique detects a change corresponding with the 9/11 attacks is intriguing. This

work suggests that our approach may provide some ability to detect change even when there is incomplete information.

That being said, it is important that future work examine the errors associated with this technique, both the false positives and false negatives. Future work should also consider the sensitivity of this approach to missing information, and to the reason why the information is missing. For example, data sets collected post-hoc that focus on activity around an event, such as the Al-Qaeda data are prone to errors of missing nodes and as a result links prior to the event. Whereas, data sets collected based on opportunity, such as the Leavenworth data, are prone to missing links among the nodes.

6.11 Sensitivity to Risk of False Positive

Sensitivity to the risk of false positives is an important consideration in detecting change in longitudinal network data. False positives occur when a change detection procedure indicates that a change may have occurred, when in fact there is no change. There exists a trade-off between false positives and rapid detection. A statistical process control algorithm that is tuned to detect changes faster will also have an increased probability of false positive.

The balance between rapid detection and false positive is determined by the decision interval of the change detection procedure. In the example in Figure 66, a CUSUM statistic is plotted for notional data. The data is the same for both charts in the figure. A change is introduced in the data at time point 22. The decision interval of the left chart is set so that the change is actually detected at time point 25. If the decision interval was lowered so that the change might be detected earlier, at time point 24, then the chart would also signal a false positive at time point 7. Therefore, it is important to determine a desired risk for false positive, and then monitor longitudinal networks for change.

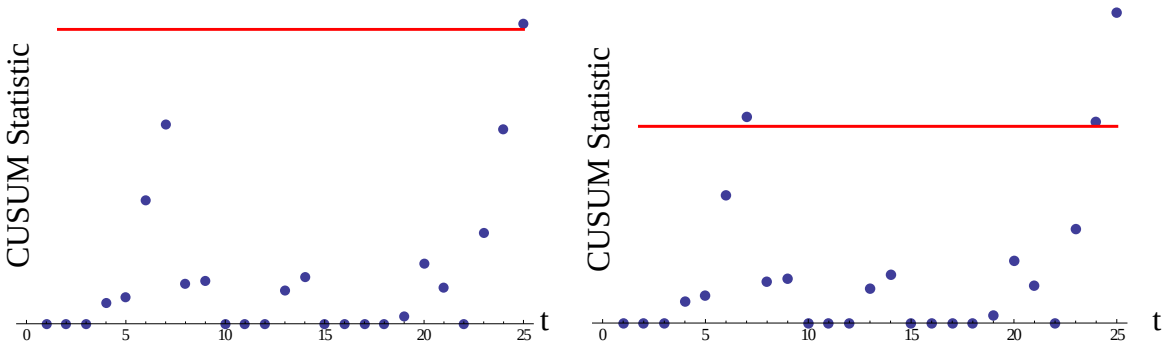


Figure 66. Trade-off Between False Positive and Rapid Detection.

I have used a risk of false alarm of 1% in the examples presented in this thesis to demonstrate the approach. A detailed discussion of risk in statistical process control is included in Appendix C. This is especially tricky for the CUSUM. The value of the CUSUM at any point in time depends on the value of the CUSUM at the previous point

in time. Therefore, estimating the values of the CUSUM involves nested conditional probability, which makes calculating the decision interval for a desired risk intractable. Through extensive Monte Carlo simulation I determine an analytic formula for the decision interval of the CUSUM which is over 99% accurate. The formula is given by,

$$h(\alpha,k)=\left(\frac{1}{5k\alpha^{0.1}}\right)\ln(k)-\left(0.53\ln(\alpha)+\left(\frac{\pi}{10}\right)\right)k^{-0.89},$$

where α is the risk of false positive, and k is the optimality constant of the CUSUM. Using this formula, the decision interval can be calculated for various values of α and k. Table 24 provides the decision interval for various values of α keeping $k = 0.5$.

Table 24. Decision Intervals for the CUSUM.

α	0.05	0.02	0.01	0.005	0.001
$h(\alpha, 0.5)$	1.99	2.85	3.50	4.15	5.65

Throughout this chapter, I have used a decision interval of 3.50, which corresponds to a false positive rate of 1%. Table 25 presents all of the example data sets that have a known change from this chapter. The known change point is included in the second column. The five remaining columns identify the time point that the first change is detected by the CUSUM procedure for that data set, with the given risk of false positive.

Table 25. Affect of Risk in Detecting Change in Real World Examples.

Data	Change	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
Fraternity	8	10	10	10	13	Never
Leav 07	3	5	5	5	Never	Never
Al-Qaeda	1997	1999	1999	2000	2000	Never
Winter C	May	Sept	Sept	Oct	Oct	Never
Winter A	May	Aug	Sept	Sept	Sept	Oct
Winter B	May	Sept	Sept	Sept	Oct	Never
IkeNet 2	25	26	26	27	27	27
IkeNet 3	14	15	18	19	19	20

It can be seen in Table 25 that when the risk of false positive is set very low, such as $\alpha = 0.001$, the change detection procedure can often miss real change. When the risk of false positive is lowered, the procedure is able to detect changes more rapidly. It is not apparent in the examples presented here that lowering the risk also introduces an increased probability of false positive. When $\alpha = 0.05$, there will be 20 observations on average between false positives. When $\alpha = 0.01$, there will be 100 observations on average between false positives.

The level of false positive that an analyst will be comfortable with depends on the application. If the analyst is creating daily networks, then he can expect a false positive more than once per month with $\alpha = 0.05$, or less than once per quarter with $\alpha = 0.01$. If the cost of missing a network change is high, then the analyst may accept greater risk in false positives in order to detect a network change more rapidly. In a terrorism application, a delay in detecting change may prevent the analyst from detecting a change until after the terrorists have carried out their attack. Accepting greater risk may improve the ability for the analyst to get inside the terrorists' decision cycle. In an organizational behavior application, there may be financial costs incurred as the analyst interviews members of the organization and searches for a potential cause of change in the network. In this situation, false positives may lead to increased unnecessary costs. Of course, missing a change in the organizational behavior example, or having a false positive in the terrorism example also have detrimental consequences. The analyst should carefully consider the trade-off between false positives and rapid detection when using SNCD.

7 Procedure for Small and High Variance Networks

Some networks may contain few nodes and high variance in network level measures. The high variance creates a random noise condition that can obscure the detection of change. This chapter outlines a procedure for handling network data with few nodes or high variance. The method is demonstrated on a unique data set collected at the U.S. Military Academy at West Point, NY.

The variance in network measures over time can be high as a result of several potential causes. The simplest explanation is that there are very few nodes included in the network. If a node is removed, it has the potential to impact $n - 1$ other links in the network out of a possible $n(n - 1)$. As the number of nodes increases, the ratio of impact $(n - 1)/n(n - 1)$ gets smaller, because the denominator of the expression grows faster than the numerator. With a small number of nodes, missing data or the erratic behavior of a single node can significantly bias the network.

The variance of a network can also be high as a result of typical changes in the day to day activities of the nodes in the network. For example, a holiday or an important deadline may affect the social network of the agents in an expected fashion. If these causes occur with regular frequency, the analyst can use the spectral analysis approach presented in Chapter 5 of this thesis. Otherwise, a different approach must be made.

If there are certain events which are likely to impact an organization and they can be known in advance, multiple linear regression can be used to model the behavior of key network measures. If the regression model can explain a sufficient amount of the variance in a given measure, then statistical process control can be applied to the residual error of this model. If the error becomes significantly high, the control chart implies that the regression model is no longer explaining the behavior of the key network measure being modeled. This in turn indicates that there may have been a change in the network.

7.1 IkeNet 1

The Eisenhower Leadership Development Program (ELDP)[4] is a one-year graduate program run as a joint effort by the United States Military Academy (USMA) and Columbia University. Each year, twenty-four Army officers (referred to in this study as Army 1 through 24) enter the program to earn a Master's degree in Social-Organizational Psychology with a concentration in Leadership and to prepare for service as mentors for West Point's cadet companies during the following two years. Social network data on email communication was collected for 24 weeks. Details regarding the data collection and network properties are described in McCulloh, et. al. (2007). The social network data collected through e-mail on this group has been referred to as the IkeNet data. The complete data includes 24 time periods collected on nine officers. Each

[4] The Eisenhower Leadership Development Program (ELDP) was originally called the Tactical Officer Education Program (ELDP). This data is referred to as the ELDP data in the original IkeNet Technical Report from the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The electronic data is available upon request from ARI.

time period is one week. The raw data consists of continuous e-mail data. To complete this clean data set, the following steps were taken: The first step of processing the raw data was to remove all emails sent by officers in ELDP to non-ELDP members. The primary concern of the study was to examine how email communication changed within the exclusive group of ELDP officers. This required that records of emails sent to non-ELDP members and email addresses of non-ELDP members in messages that were sent to mixed parties were deleted. Thus, all subsequent network pictures would only involve the email communication among the 24 officers. The network information can only be viewed as "near" complete as emails sent using Webmail were not collected because of limitations of the data collection software (McCulloh, et. al. 2007).

The continuous data were then separated it into 24 one week time periods. Weekly time periods were found to be the best resolution feasible for change detection. This data was collected before the spectral analysis method from Chapter 5 was developed. The aggregation level was therefore based on known email behavior exhibited by the group. If daily time segments were used, we would only detect Friday as a change point for most people. An average person intuitively maintains different communication patterns during the week when they are at work than they do on the weekend with friends and family. Monthly time segments can be very different as well. For an academic setting such as ELDP, the month of December includes Christmas Break and the month of March includes Spring Break. This significantly changes the communication behavior for the month. Weekly communication on the other hand aggregates normal daily fluctuations in e-mail activity, while providing a larger number of time periods to detect significant change.

Some of the officers stopped sending email at some point in the study and did not send email again. The principal investigator interviewed these officers and found that they had experienced technical problems during the study and had reformatted their hard drive, thereby erasing the collection patch. Other officers began to rely on webmail, which bypassed the collection patch. Therefore, the communication data collected was incomplete and not identically distributed. Officers, whose data collection was incomplete, were eliminated from further study. This reduced the number of officers in the data set from 24 to nine. Average network measures calculated on the reduced data set followed a normal distribution. A communication network for the reduced data set is shown in Figure 67 for the week of 29 October 2007.

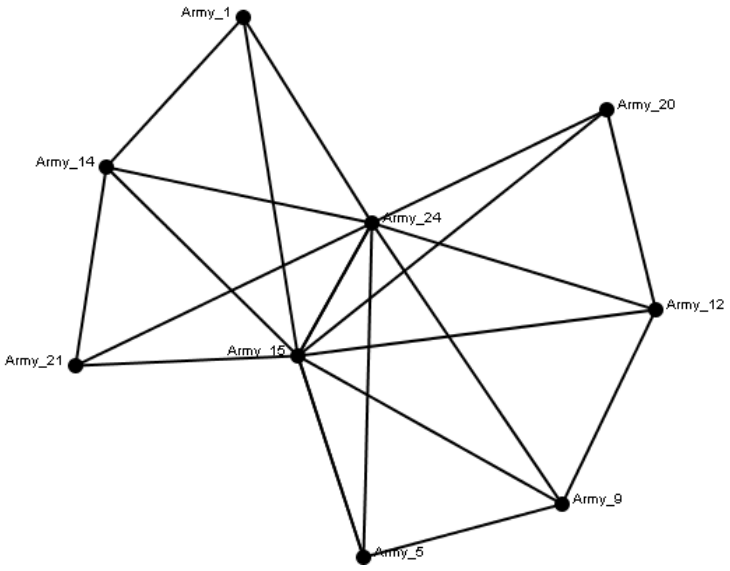


Figure 67. Email Network of ELDP Officers During Week of 29 October 2007.

Using this much smaller, but complete network, the average degree, betweenness, and closeness all appeared to be normally distributed. Determining baseline values, however, was still not possible because the network contained too much variance. There was no stable network measure behavior. In order to account for the variance caused by differing schedules week to week, I examined a copy of the ELDP planning calendar for the entire year. The calendar combined with interviews with officers allowed investigators to determine the number of significant events from a variety of categories that occurred each week. The significant events based on qualitative assessments by the officers were Academic Requirements, the Next Week's Academic Requirements, Administrative Events (such as a class trip or cancelled class), Group Projects, Social Gatherings, and Days Off.

Using MINITAB Statistical Software, analysis of variance (ANOVA) tests were run on predictors to determine if they were statistically significant factors in determining network measures for the first semester (12 weeks). Days Off was the most significant factor, due to Christmas break in the middle of the 24 week study, however once these weeks were removed from the study, Days Off was no longer a significant factor in any model. The best linear regression model obtained from first semester (12 weeks) data was for closeness based on the number of group projects, the number of social gatherings, and the number of emails sent each week. The ANOVA table from the regression is displayed in Table 26 and the regression equation is given by,

$$\text{Closeness} = 0.18 - 0.11(\text{Group Projects}) + 0.11(\text{Social Gatherings}) + 0.0074(\text{Number of Emails})$$

Table 26. ANOVA Table for Closeness Predictors.

Predictor	Coefficient	SE Coefficient	T	P	VIF
Constant	0.18	0.034	5.4	0	
Group Projects	-0.11	0.05	-2.1	0.05	1.3
Social	0.11	0.04	2.89	0.01	1.3
Number of Emails	0.0074	0.00084	8.77	0	1

This model has an adjusted R^2 value of 79.8%, accounting for a large majority of the variance in the network measure and a predictive R^2 value of 70.9%. Slightly surprising from this model is the effect of group projects on closeness. An increase in group project work was correlated with a decrease in communication. This might be due to the fact that as a group project comes due, the ELDP officers may communicate more with their immediate team of group members, and communicate more face-to-face, but overall they decrease communication outside of their working groups and through email in order to focus on the project. The positive effects of Social Gatherings and more emails sent over the week had the foreseen effect of improving group closeness.

The model created from the first semester was used to predict the average closeness value for the second semester. The CUSUM control chart was applied to the residual error between the prediction and the actual second semester data. This allowed me to conduct real-time monitoring of a social group for change.

Being able to predict the closeness of the ELDP communication network was essential in explaining much of the variance in the network. The control chart could then be used to determine when the network changed away from the model. In effect, when is the model no longer providing a good prediction? Using the closeness model developed from data obtained during the first semester of the ELDP graduate program, predicted values were calculated for each week of the second semester using the number of social gatherings and group projects from the ELDP calendar and the number of emails sent by observation. These were compared with the observed network measures. The residuals were verified as normally distributed to meet the prerequisites of the CUSUM Control Chart. The C^+ and C^- statistics were calculated for each week using a k value of 0.5 and a control limit of 3.5, which corresponds to a false positive risk of 1% (see Appendix C). A graph of the CUSUM statistic for the ELDP data is in Figure 68.

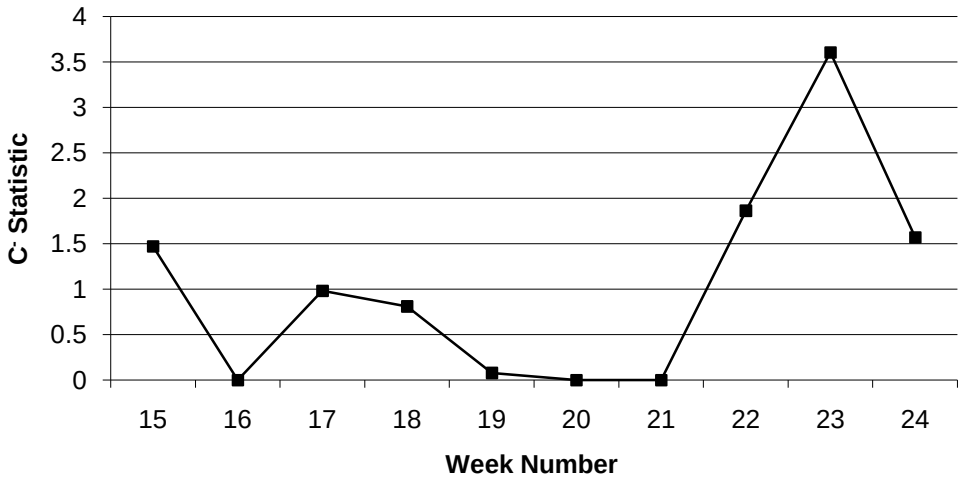


Figure 68. Plot of closeness CUSUM statistic for nine ELDP officers.

Figure 68 indicates that the control chart signals on Week 23 (see Table 26). Week 23 was the week that the ELDP officers took the comprehensive exam for their graduate program. It was the most significant academic event of the year. Tracing the C^- statistic back to the last time it was zero, the most likely change point was during Week 21. Upon first examination, Week 21 looks like it should be a typical academic week, with no unusual events or graded projects. However, based on interviews conducted with ELDP officers after the signal was detected, it was discovered that Week 21 was a critical preparation week prior to the comprehensive exam when the study questions for the exam were sent to the students. Thus, the CUSUM control chart signals on Week 23 as it represents a significant departure from the value predicted by the model.

Table 27. CUSUM Statistic Values for Closeness Network Measure.

Week	Closeness	Model	Z	C+	C-
15	0.3332	0.4712	-1.9714	0.0000	1.4714
16	0.5134	0.3798	1.9086	1.4086	0.0000
17	0.2760	0.3798	-1.4829	0.0000	0.9829
18	0.3332	0.3562	-0.3286	0.0000	0.8114
19	0.5406	0.5243	0.2329	0.0000	0.0786
20	0.6536	0.5745	1.1300	0.6300	0.0000
21	0.4977	0.3916	1.5157	1.6457	0.0000
22	0.1258	0.2913	-2.3643	0.0000	1.8643
23	0.2646	0.4215	-2.2414	0.0000	3.6057
24	0.5226	0.4152	1.5343	1.0343	1.5714

The CUSUM control chart implemented on the residuals of a communication model proved to be effective at detecting organizational change in the ELDP program. It is also interesting to note, that a decrease in communication can indicate that a major event is about to occur, as the officers rely less on email and more on face-to-face communication and study groups.

7.2 Discussion

This approach was demonstrated to be effective at modeling a social network measure of interest in a small social network data set, based on scheduled events. Change was then able to be detected as a departure from the typical behavior of the network. This was apparent after applying statistical process control to the residual error from a regression model, where the response variable was the network measure of interest and the predictor variables were scheduled events for the group.

For the example presented here, the average closeness measure was best explained by the scheduled events. This may not be true for all data sets. I recommend that several network measures be investigated for correlations between scheduled events or other known information on the group in question. Change detection should then be applied to the model that does the best job at explaining the response variable over time. In the example, that happened to be the average closeness. In other applications, it might be the average betweenness, diameter, or even density.

If the network is larger and free from high variance, change detection can be directly applied to the network measures. This is a much simpler approach for implementing change detection than the method presented in this chapter. If on the other hand, there are fewer than 20 nodes, there is high variance in the network measures, and factors contributing to network structure can be measured, this approach may be effective in detecting network change over time.

8 Robustness of Change Detection

A major concern in social network analysis is how random error can impact results and thus the conclusions reached by the analyst. By random error, I mean the random addition of links or the random removal of links. This is equivalent in practice to observation error, or possibly the deliberate attempt to mislead the analyst by the organization being monitored, such as in terrorist applications. This chapter will demonstrate that the proposed social network change detection methodology is relatively robust to error.

There are not many publications on network robustness in the literature. Borgotti, Carley, and Krackhardt (2005) investigated the impact of adding and removing edges from Erdos-Renyi random networks on the estimation of centrality measures. Frantz, McCulloh, and Carley (n.d.) investigated robustness for different network topologies. Costenbader and Valente (2003) looked at the robustness of centrality measures for inaccurate or incomplete network data. These are the only publications revealed in an extensive literature review. Even the recent papers identified here point to a lack of research in this area. The limited work on robustness, however, suggests that network analysis is relatively robust to error.

The real-world data from Chapter 6 is further investigated for its robustness to error. The links in all of the data sets were removed with probability *p*, which varied from 0.01 to 0.10. This introduced random error into the real-world data sets from Chapter 6. The social network change detection was again run on the modified data, using the same parameter settings as articulated in Chapter 6. A false-alarm risk level of 0.05 was used in all cases.

Four performance measures were investigated to evaluate the robustness of the proposed change detection approach. A *false alarm* (FA) in the robustness experiment is considered to occur when the procedure signals a potential change in the network at a time point earlier than the identified change in the Chapter 6 (no error) analysis. A *miss* in the robustness experiment occurs if the procedure does not detect a change in the network when the random error is introduced. A *different change point* (DCP) occurs when the estimate of when a change actually occurs is different when random error is introduced compared to the change point estimated in Chapter 6. A *late signal* (LS) occurs when the procedure experiences a delay in signaling the change in the network when random error is added. The random removal of links was replicated 100 times for each of the data sets investigated. The robustness performance measures are therefore the percentage of FA, miss, DCP, and LS that occurred in the 100 different replications of introducing random error into each of the real-world networks. Table 28 presents the results of the virtual robustness experiments.

Table 28. Robustness of Change Detection to Missing Links.

Data	Change	ε = 0.00	ε = 0.01	ε = 0.05	ε = 0.10
Fraternity	8	10	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 **DCP=15%** **LS=8%**
Leav 07	3	5	FA=0 Miss=0 DCP=0 LS=0	FA=0 **Miss=13%** DCP=0 LS=0	FA=0 **Miss=46%** **DCP=6%** LS=0
Al-Qaeda	1997	1999	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0
Winter A	May	Aug	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0
Winter B	May	Sept	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0
Winter C	May	Sept	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0
IkeNet 2	25	26	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0
IkeNet 3	14	15	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0	FA=0 Miss=0 DCP=0 LS=0

It can be seen in Table 28 that even when 10% of the links in the network are randomly removed, there is no difference in the change detection performance for most of the data sets. The Leavenworth 07 data set is the most severely affected, where a 10% random removal of links can cause the procedure to miss the change 46% of the time. In addition, there is a different estimate of the change point for both the Leavenworth 07 and Fraternity data when there is an error rate of 10%. The other six data sets are robust to the removal of links up to a rate of 10%.

It is not clear why the Leavenworth 07 and Fraternity data sets are affected by the random removal of links, when the others are not. The Leavenworth 07 data set has eight time periods, was collected by survey, and uses rating relations. The three wintering over data sets also consist of eight time periods, were collected by survey, and use rating relations, however, change detection does not appear to be affected in these data. The Leavenworth 07 data follows a free choice survey design, but the IkeNet data sets are also free choice. The Fraternity data is about as different from the Leavenworth 07 data as any other data set explored. The Fraternity data consists of almost twice as many time periods, uses rankings instead of ratings, and is a fixed choice design. Therefore, it does not appear that the number of time periods, method of collection, type of relations, type of survey design, or size of the network affect the robustness of the network to missing links.

The robustness of the data sets was further investigated by exploring the correlation between the node-level measure of an original network and the same network missing 10% of the links. Four node-level network measures were investigated; degree, betweenness, closeness, and eigenvector centrality. This was done for the eight data sets explored in Chapter 6 of this thesis: Fraternity, Leavenworth 07, Al-Qaeda, Wintering Over A, B, and C, and IkeNet 2, and 3. For all data sets the first time period was used. Since the missing links are removed at random, 100 instances of networks with missing links were generated and compared to the original data set. The results of the mean correlation and the standard error are reported in Table 29. Correlations that were statistically different than 0 are in bold.

The results of the correlations in Table 29 provide little insight into the robustness performance of the data sets. All correlations appear relatively high, with the lowest significant correlation being the closeness measure in the Wintering-Over C data set at 0.7074. The only correlation that does not appear statistically different than 0 is the closeness measure in the Fraternity data set. The Fraternity data set also has the lowest correlation in degree and a lower than average correlation in Eigenvector Centrality. The Leavenworth 07 data also has poor robustness; however, all the measures, degree, betweenness, closeness, and eigenvector centrality are more highly correlated than the average observed across all eight data sets. All of the wintering-over data sets exhibit the lowest correlation in the betweenness scores, which is used in the change detection procedure, however, they were found to be robust to missing links. Finally, a multiple linear and non-linear regression analysis was conducted to test if the number of nodes and links in the network affected observed correlations. There was no statistical significance for the number of nodes, links, or any interaction affect between the two factors.

Therefore, the robustness does not appear to be affected by the size of the network or the correlation between node-level measures in a network and the same network missing 10% of the links. The fact that the correlations are high, however, does suggest that missing data may not affect the estimates of the top central nodes in a network. Future work in this area should investigate the problem using simulation and response surface methods to explore the problem more thoroughly.

Table 29. Correlation of Measures Between Network and Network Missing 10% of Links.

Data Set	Degree	Betweenness	Closeness	Eigenvector
Fraternity	**0.9369**	**0.8692**	0.0912	**0.9079**
	(0.0249)	(0.0486)	(0.1554)	(0.0488)
Leavenworth 07	**0.9812**	**0.9476**	**0.9663**	**0.9765**
	(0.0034)	(0.0320)	(0.0448)	(0.0110)
Al-Qaeda	**0.9945**	**0.9780**	**0.9284**	**0.9899**
	(0.0005)	(0.0183)	(0.0344)	(0.0223)
Winter Over A	**0.9648**	**0.8320**	**0.8270**	**0.9820**
	(0.0099)	(0.0600)	(0.0448)	(0.0078)
Winter Over B	**0.9650**	**0.7510**	**0.8370**	**0.9764**
	(0.0126)	(0.0554)	(0.0362)	(0.0110)
Winter Over C	**0.9689**	**0.7695**	**0.7074**	**0.9874**
	(0.0107)	(0.0626)	(0.0712)	(0.0055)
IkeNet 2	**0.9692**	**0.9317**	**0.9675**	**0.9412**
	(0.0274)	(0.0740)	(0.0245)	(0.0739)
IkeNet 3	**0.9880**	**0.9567**	**0.8968**	**0.8898**
	(0.0036)	(0.0310)	(0.0533)	(0.1712)

A potential, non-obvious factor that may contribute to error probability in network data is statistical dependence. The correlation between certain network measures may provide some insight into the network structure and the dependence between network properties. In order to explore structural dependence, the degree, betweenness, closeness, and eigenvector centralities were calculated for all nodes, across all time periods, for the Fraternity, Leavenworth 07, Al-Qaeda, and the three Wintering Over data sets. The correlations between node-level measures were calculated. The correlations were averaged across the time periods and the standard errors of the correlations were also recorded. Tables 30-35 display the average correlations between network measures, with the standard error represented in parentheses. Statistically significant correlations are in bold.

Table 30. Newcomb Fraternity.

	Betweenness	Degree	Closeness	Eigenvector
Betweenness	1.00000 (0.000000)			
Degree	**0.439868** (0.192845)	1.00000 (0.000000)		
Closeness	-0.0166 (0.302627)	-0.42291 (0.246748)	1.00000 (0.000000)	
Eigenvector	0.377349 (0.159042)	**0.906408** (0.022901)	-0.3562 (0.286111)	1.00000 (0.000000)

Table 31. Leavenworth 07.

	Betweenness	Degree	Closeness	Eigenvector
Betweenness	1.00000 (0.000000)			
Degree	**0.80714** (0.01765)	1.00000 (0.000000)		
Closeness	**0.37296** (0.031345)	**0.572952** (0.067039)	1.00000 (0.000000)	
Eigenvector	**0.65571** (0.060462)	**0.820117** (0.04871)	**0.4535** (0.05176)	1.00000 (0.000000)

Table 32. Al-Qaeda.

	Betweenness	Degree	Closeness	Eigenvector
Betweenness	1.00000 (0.000000)			
Degree	**0.657298** (0.058526)	1.00000 (0.000000)		
Closeness	**0.225437** (0.011616)	**0.569288** (0.027522)	1.00000 (0.000000)	
Eigenvector	**0.141788** (0.041205)	**0.519142** (0.068754)	**0.147727** (0.050628)	1.00000 (0.000000)

Table 33. Wintering-Over A.

	Betweenness	Degree	Closeness	Eigenvector
Betweenness	1.00000 (0.000000)			
Degree	**0.528867** (0.18476)	1.00000 (0.000000)		
Closeness	**0.714927** (0.080143)	0.389236 (0.17361)	1.00000 (0.000000)	
Eigenvector	0.396394 (0.228421)	**0.949425** (0.00954)	0.275945 (0.189964)	1.00000 (0.000000)

Table 34. Wintering-Over B .

	Betweenness	Degree	Closeness	Eigenvector
Betweenness	1.00000 (0.000000)			
Degree	**0.444704** (0.143775)	1.00000 (0.000000)		
Closeness	0.510902 (0.214759)	**0.489819** (0.113464)	1.00000 (0.000000)	
Eigenvector	**0.412923** (0.168626)	**0.939555** (0.020432)	**0.366582** (0.126274)	1.00000 (0.000000)

Table 35. Wintering-Over C.

	Betweenness	Degree	Closeness	Eigenvector
Betweenness	1.00000 (0.000000)			
Degree	0.45036 (0.234517)	1.00000 (0.000000)		
Closeness	**0.725564** (0.087041)	**0.605238** (0.124091)	1.00000 (0.000000)	
Eigenvector	0.401038 (0.255925)	**0.96263** (0.011831)	**0.542198** (0.118581)	1.00000 (0.000000)

Several interesting insights can be gained from looking at the correlation between network measures in the six data sets investigated. There is a high, significant correlation between degree and eigenvector centrality in all data sets. There is no other pair of measures with a correlation that is significant in all data sets, to include the three wintering-over data sets which are all very similar. The correlations between degree and eigenvector centrality are lower in data sets with more nodes. Since the eigenvector centrality measures the influence of a node to the extent that its neighbors are central, this is not surprising. With less than 30 nodes, it is less likely to find a node with influential alters that is not highly connected itself. There is no pattern in this correlation that would indicate why one data set is more robust to missing links than another.

There is a generally positive correlation across all measures for all data sets except for the Fraternity data. The Fraternity data was the only data set that involved rankings. In addition, the rankings were dichotomized using an approach proposed in the literature by Krackhardt (1998). The dichotomization scheme may affect the network structure. Future work could explore the effects of dichotomization schemes on network analysis to include the correlation between measures, identification of key entities, identification of highly central actors and more.

One possible explanation for the Fraternity and the Leavenworth 07 data sets not being robust to missing links lies in the correlation between degree and betweenness. The Leavenworth 07 data possesses a very high correlation between degree and betweenness that is greater than any of the other data sets. The only significant correlations in the Fraternity data are the degree-betweenness and degree-eigenvector centrality, which is typical of all data sets. Since the robustness experiment was focused on detecting a change in the average betweenness of the network over-time, it may be plausible that a correlation between degree and the measure being investigated could affect the power of detection in the procedure. More research is required to reach definitive conclusions.

Social network change detection is a powerful, novel approach for detecting significant changes in organizational behavior over time. This approach has been demonstrated to be effective in both simulated and real-world data. In this chapter, I have addressed the robustness of the proposed approach to missing data and to statistical dependence in network measures. Social network change detection is found to be robust to missing data up to a level of 10% for most data sets investigated. In the case where the approach was not robust, such as the Fraternity and Leavenworth 07 data sets, there was a high correlation between the degree and the betweenness measure, which was monitored for significant change. This suggests that the robustness of social network change detection may be a function of the statistical dependence between network measures. This opens a new area of research into correlations in network measures, the robustness of network measures to missing data, and the robustness of change detection methods. In the near term, this chapter demonstrates that analysts can still gain valuable insights into dynamic networks using the methods proposed in this thesis.

9 Summary

9.1 Lessons Learned

Control charts are a critical quality-engineering tool that assists manufacturing firms in maintaining profitability (Montgomery, 1991; Ryan, 2000). The 10 examples presented in this paper demonstrate that social network change detection could enable analysts to detect important changes in a variety of different network measures over time. Furthermore, the most likely time that the change occurred can also be determined. This allows one to allocate minimal resources to tracking the general patterns of a network and then shift to full resources when changes are determined[5].

This paper describes an algorithm for change detection, and then demonstrates its ability to detect changes in networks. No doubt other change detection methods will emerge. My point, is that it is critical to be able to detect change in networks over time and to determine when those changes are not simply the random fluctuations of chance. The strengths of the proposed method are its statistical approach, ability to quantify the rate of false alarm, a wide range of social network metrics suitable for application, its ability to identify change points in organizational behavior, and its flexibility for various magnitudes of change. While the CUSUM may be effective in detecting change in non-normally distributed network measures, the false positive estimates may be biased. Good social network measures to use are those that scale well with the number of nodes and are averaged over all nodes in the network. Examples, demonstrated in this thesis are the average, maximum, and standard deviation of the betweenness, closeness, and eigenvector centralities. Average Degree and Density are not used, because they do not scale well with the number of nodes in the network. Individual nodes that are not present in all time periods may therefore bias the CUSUM statistic, causing an increased rate of false alarm, or reducing the power of the statistic. Other limitations of the algorithm cannot yet be determined as this is the first application of statistical process control methods to the problem of SNCD. Future research will provide much greater insight into the strengths and limitations of this approach to the problem. The remainder of this chapter will identify specific areas of caution when interpreting findings and identify areas for future research.

9.2 Limitations

The empirical results described in this paper, such as the detection of change in the Al-Qaeda network should be viewed with caution. I present them here purely to illustrate the methodology. Limitations on the data make it difficult to determine the validity of the results; thus, we should simply view these results as showing the promise

[5] Three social network change detection algorithms (Shewhart X-Bar, Cumulative Sum, and Exponentially Weighted Moving Average) are available in the "Over-Time Viewer" and the "Statistical Change Detection Report" in the software tool, Organizational Risk Analyzer (ORA) available through the Center for Computational Analysis of Social and Organizational Systems (CASOS), http://www.casos.cmu.edu.

of this methodology. The IkeNet data is a small sample capturing only email traffic and not all communication and interaction among officers. The fact that even in this small sample of behavior we were able to systematically detect a key change suggests the value of the proposed approach. The Al-Qaeda data, was based on open source information. As such it is an incomplete representation of interaction in that terror network. We cannot be sure that we have the entire communication network, or even a true picture of the observed communication network. However, the fact that our technique detects a change corresponding with the 9/11 attacks is intriguing. This work suggests that our approach may provide some ability to detect change even when there is incomplete information.

That being said, it is important that future work examine the errors associated with this technique, both the false positives and false negatives. Future work should also consider the sensitivity of this approach to missing information, and to the reason why the information is missing. For example, data sets collected post-hoc that focus on activity around an event, such as the Al-Qaeda data are prone to errors of missing nodes and as a result links prior to the event. Whereas, data sets collected based on opportunity, such as the IkeNet data, are prone to missing links among the nodes.

In an effort to define a manageable thesis, several limitations were placed on the scope of the research. This thesis did not attempt to completely define the probability structure of all network models. This thesis is focused on a fixed network that does not grow in size over time. Findings are limited to modeling and detecting changes, but not the causes of the change. The extents to which social network measures and data are used are only to demonstrate the mathematical soundness of the method.

The simulations and most of the real-world data sets all consisted of a fixed number of nodes that did not change over time. In some data sets the number of nodes can change, as in the Al-Qaeda data set. In one sense, the change in the number of nodes is a change itself. If the organization is growing or shrinking at a steady rate, however, one may be interested in monitoring the difference in a measure between time points. This may be equivalent to monitoring the rate of growth in the measure being monitored. In some cases, network measures may not be as sensitive to the addition or removal of nodes. Additional work is required to study the impact on change detection of adding or removing nodes at a steady rate.

This thesis did not specify any particular distribution for network measures. The estimates of false alarms are determined based on normality assumptions. As stated in Chapter 2, networks where the relationship between nodes requires a meaningful investment of time or other resources tend to have many network measures that appear normally distributed. In other networks, such as scale-free networks common for modeling the internet and certain biological networks, the false alarm rate may be adversely affected. Figure 67 shows the variance of data collected from a normal and right skewed distribution versus the number of observations sampled. The increased variance from the right skewed data will inflate the decision interval calculated on a few

initial observations, making it more difficult to detect change, or more susceptible to false alarm.

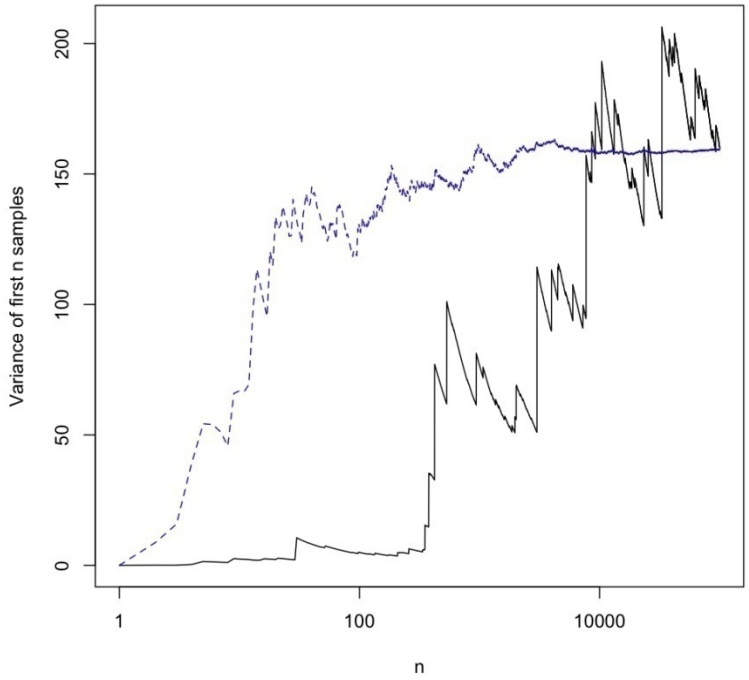


Figure 69. Bias Induced in Right Skewed Data

Some social scientists do not believe that groups can be adequately captured by quantitative analysis and statistical distributions (Morrow and Brown, 1994). I do not attempt to tackle this argument. Clearly, the work of this thesis contributes to quantitative methods in social science. I also do not claim that a detected change is definitive proof that the organization has in fact changed. This approach will only detect a statistically significant change in the observed network measure of an organization. This could be a false alarm, an expected event affecting the organization, among other causes. Change detection simply alerts an analyst or social scientist that a change may have occurred. It is incumbent on the analyst or social scientist to investigate the group using many different methods in the social sciences to determine if change has in fact occurred, the nature of that change, and the cause of change. The approach laid out in this thesis will narrow the scope of this task by quickly identifying potential change and estimating when the change occurred.

Network change detection was only demonstrated on 10 data sets. With a vast amount of existing data, change detection could be applied to many data sets with interesting and unique findings. The focus of this thesis is to develop new analytic methods that can be applied to any network over time. The selected data sets were chosen to clearly articulate these methods, and to create a concise and thorough thesis.

This thesis does not attempt to speak to the importance of change. Rather, this thesis articulates an approach to characterize network behavior based on a suite of user defined

network measures. Change detection seeks to identify a statistically significant change in the observed network measures over time. I do not claim that significant changes occur suddenly. They can emerge slowly over time, or rapidly. Of course, change detection will respond better to rapid changes, since slow changes are also slow to detect. Even in such a situation, change detection offers a more powerful detector of change than simple over time observation.

It is also important to point out that change detection can be deliberately obscured. Recognizing that change detection is only detecting a statistically significant change in a network measure, the monitored organization could take actions to hide or add links to make the observed measures appear consistent over time. Considering the complexity of networks, this is a more challenging task than it may seem. The nature of the network context significantly affects this problem. More work is required to investigate methods to conceal network change from the methods proposed in this thesis.

9.3 Future Directions

In order to rectify the above shortcomings, future research should focus on near-complete datasets with high resolution. Higher resolution involves taking many snapshots of the network. This may mean, simply an increase in frequency, e.g. changes by month, or it may mean a longer time horizon, e.g., more years. The right choice will depend on the problem where we want to detect network change. More data points will provide more opportunities to detect changes while they are still small, instead of allowing them to incubate and grow as was the case for the Al-Qaeda data. As a minimum two observed networks are required to estimate the normal behavior of a social group being monitored for change. In practice, five or more networks are preferred to reduce the variance in estimating the CUSUM parameters. Larger datasets will also provide near continuous network measures permitting the use of control charts for continuous data. Near complete data means that the data should cover the communication network, with little or no missing information for a large contiguous period. Here one might consider simply tracking a group in general, as opposed to focusing on tracking relative to a specific event. Data such as that on the US Congress or Supreme Court that is regularly output might provide a good source of data.

Research on the distributions is needed. Preliminary work on the distributions of network measures suggest that the assumption of normality does not hold for small networks, extremely sparse networks, and for certain metrics (Kim and Carley, working paper). Future work should consider these factors to determine the range of networks for which SNCD will work. Clearly, if the network measures are normally distributed, the CUSUM control chart can be used to monitor network change. If they are not, the false alarm probability will increase as demonstrated in Figure 69. Other topological properties may also affect change detection. Keeping in mind that change detection is focused on a defined set of measures, the distributional assumptions can be verified much like residual analysis in regression. When the measures appear to violate normality assumptions, it may be possible to develop transformations or develop more complex

change detection methods. Future work should address this issue. It is important to keep in mind that most real-world social networks satisfy the distributional assumptions for the change detection methods proposed in this thesis.

It may also be possible to extend change detection to node level measures. This would be done by simply monitoring the node level measure of an individual agent or agents over time using one of the algorithms proposed in Chapter 4. Again the distributional assumptions would need to be verified. Node level change detection may help further isolate change in an organization by monitoring the behavior of key individuals, without the noise introduced by less influential agents. More work in this area will prove beneficial.

Future research should also look at the sensitivity of the optimality constant, *k* and control limit values of the CUSUM Control Chart for network measure change detection. As stated earlier, these values are generally arbitrarily chosen and then optimized for the process. By using further Monte Carlo simulations, a researcher should determine which parameter value would be best in detecting certain types of changes such as sudden large changes or slow creeping shifts. Usage of control charts on comparing models and observations should also be studied to see what specific conclusions can be obtained.

Multi agent simulations provide valuable insight into the performance of control charts for social network change detection applications. Simulations allow an investigator to introduce various changes into a simulated organization and evaluate the average detection length for different algorithms. Simulations provide an efficient means of evaluating change detection on social networks. More importantly, however, is the ability to create more controlled experiments, by fixing certain variables, exploring others, and using many replications to estimate error. Simulation studies will continue to be extremely useful in exploring extensions of this methodology.

Social network change detection is important for identifying significant shifts in organizational behavior. This provides insight into policy decisions that drive the underlying change. It also shows the promise of enabling predictive analysis for social networks and providing early warning of potential problems. In the same way that manufacturing firms save millions of dollars each year by quickly responding to changes in their manufacturing process, social network change detection can allow senior leaders and military analysts to quickly respond to changes in the organizational behavior of the socially connected groups they observe. The combination of statistical process control and social network analysis is likely to produce significant insight into organizational behavior and social dynamics. Immediate applications to counter terrorism and organizational behavior are obvious. As a scientific community we can hope to see more research in this area as network statistics continue to improve.

10 References

Albert, R. (2008) Personal conversation at the 3rd US Military Academy Network Science Workshop, West Point, NY, 15 October 2008.

Albert, R., Jeong, H., and Barabasi, A.L. (2000). Error and attack tolerance of complex networks. *Nature*, 406, 378-382.

Alderson, D.L. (2008). Catching the "Network Science" Bug: Insight and Opportunity for the Operations Researcher. *Operations Research*, 56(5), 1047-1065

Anderson, C.J., Wasserman, S., Crouch, B. (1999). A *p** primer: logit models for social networks. *Social Networks* 21, 37–66.

Baller, D., Lospinoso, J., and Johnson, A.N. (2008). An Empirical Method for the Evaluation of Dynamic Network Simulation Methods. In *Proceedings, The 2008 World Congress in Computer Science Computer Engineering and Applied Computing*, Las Vegas, NV.

Banks, D.L., and Carley, K.M. (1996). Models for network evolution. *Journal of Mathematical Sociology*, 21, 173-196.

Barabasi, A.L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.

Barabasi, A.L. (2002). *Linked: The New Science of Networks.* Cambridge, MA: Perseus.

Barabasi, A.L., Jeong, H., Ravasz, E., Neda, Z., Schubert, A., and Visek, T. (2002). On the topology of the scientific collaboration networks. *Physica A*, 311, 590-614.

Barabasi, A.L. (2008). Presentation at the 3rd US Military Academy Network Science Workshop, West Point, NY, 15 October 2008.

Bernard, H.R. and Killworth, P.D. (1977) Informant accuracy in social network data II. *Human Communications Research*, 4, 3-18.

Bernard, H. R., Killworth, P. D., and Sailer, L. (1980). Informant accuracy n social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks, 2*, 191-218.

Bernard, H. R., Killworth, P. D., and Sailer, L. (1982). Informant accuracy in social network data V: An experimental attempt to predict actual communication from recall data. *Social Science Research, 11*, 30-66.

Bollobas, B. (1998). *Modern Graph Theory.* Springer-Verlag, New York.

Bollobas, B. and Riordan, O. (2003). Robustness and vulnerablility of scale-free random graphs. *Internet Math*, 1(2), 215-225.

Bonacich, P., Oliver, A., and Snijders, T.A.B. (1998). Controlling for size in centrality scores. *Social Networks*, 20(2): 135-141.

Borgotti, S.P., Carley, K.M., and Krackhardt, D. (2006). On the Robustness of Centrality Measures Under Conditions of Imperfect Data, *Social Networks*, 28:124-136.

Caldarelli, G. (2007). *Scale-Free Networks: complex webs in nature and technology.* Oxford University Press.

Carley, K.M. (1996). A Comparison of Artificial and Human Organizations, *Journal of Economic Behavior and Organization*, 31:175-191.

Carley, K.M. (1999). On the evolution of social and organizational networks. *Research in the Sociology of Organizations*, 16, 3-30.

Carley, K.M. (1991). A theory of group stability. *American Sociology Review*, 56(3):331–354.

Carley, Kathleen M. (1990). Group Stability: A Socio-Cognitive Approach. *Advances in Group Processes*, 7, 1-44.

Carley, K.M. (1995). Communication Technologies and Their Effect on Cultural Homogeneity, Consensus, and the Diffusion of New Ideas. *Sociological Perspectives*, 38(4): 547-571.

Carley, K. M. (2007). ORA: Organizational Risk Analyzer v.1.7.8. [Network Analysis Software]. Pittsburgh: Carnegie Mellon University.

Carley, K.M. (2006). *A Dynamic Network Approach to the Assessment of Terrorist Groups and The Impact of Alternative Courses of Action*. In "Visualising Network Information". Meeting Proceedings RTO-MP-IST-063. Neuilly-sur-Seine, France: RTO. Available from: http://www.vistg.net/documents/IST063_PreProceedings.pdf.

Carrington, P.J., Scott, J., and Wasserman, S. (2007). *Models and Methods in Social Network Analysis.* Cambridge University Press.

Coleman, T. F. and Moré, Jorge J. (1983). Estimation of sparse Jacobian matrices and graph coloring Problems. *SIAM Journal on Numerical Analysis*, 20 (1): 187–209.

Costenbader, E. and Valente, T. (2003). The stability of centrality measures when networks are sampled. *Social Networks* 25, 283-307.

Donninger, C. (1986). The distribution of centrality in social networks. *Social Networks*, 8, 191-203.

Doreian, P., and Stokman, F.N. (Eds.) (1997) *Evolution of Social Networks.* Amsterdam: Gordon and Breach.

Doreian, P. (1983). On the evolution of group and network structures. II. Structures within structure. *Social Networks*, 8, 33-64.

Doyle, J., Alderson, D., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R., and Willinger. W. (2005). The "robust yet fragile" nature of the Internet. In: Proceedings of the National Academies of Science, 102(41) 14497–14502.

English, J.R., Martin, T., Yaz, E. and Elsayed, E. (2001) *Change point detection and control using statistical process control and automatic process control*. Presentation at the IIE Annual Conference, 2001, Dallas, TX.

Erdos, P, and Renyi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6, 290-297.

Erdos, P, and Renyi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 17-61.

Erdos, P, and Renyi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12, 261-267.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions.* 3 Ed. New York: Wiley.

Feld, S. (1997). Structural embeddedness and stability of interpersonal relations. *Social Networks*, 19, 91-95.

Fisher, R.A., Thornton, H., and Mackenzie, W. (1922). The Accuracy of the Plating Method of Estimating the Density of Bacterial Populations, with Particular Reference to the Use of Thornton's Agar Medium with Soil Samples. *Annals of Applied Biology*, 9, 325–359.

Frank, O. (1991). Statistical analysis of change in networks, *Statistica Neerlandica* **45** (1991), 283–293.

Frank, O., Nowicki, K. (1993). Exploratory statistical analysis of networks. *Annals of Discrete Mathematics* 55, 349–366.

Frank, O., Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* 81, 832–842.

Freeman, L.C. (1996). Some antecedents of social network analysis. *Connections*, 19, 39-42.

Frantz, T. and Carley, K.M. (2008). CEMAP II: An Architecture and Specifications to Facilitate the Importing of Real-World Data into the CASOS Software Suite. *Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-08-130*

Frantz, T., McCulloh, I., and Carley, K.M. (n.d.) *Estimating the Reliability of Top-Actor Identification.* Unpublished manuscript.

Freeman, L. (1979). Centrality in social networks: I, conceptual clarification. *Social Networks* 1 (1979), 215-239.

Freeman, L.C. and Freeman, S.C. (1980). A semi-visible college: Structural effects of seven months of EIES participation by a social networks community. In: *Electronic Communication: Technology and Impacts.* Henderson, M. and MacNaughton, M. (Eds.). Washington, DC: American Association for the Advancement of Science, 77-85.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40** (1977), 35-41.

Goodreau, S.M. (2007). Advances in exponential random graph (*p**) models applied to a large social network. *Social Networks*, 29, 231-248.

Goodreau, S.M., Hunter, D.R., and Morris, M. (2005). Statistical Modeling of Social Networks: Practical Advances and Results. Center for Studies in Demography and Ecology, University of Washington, Working Paper No. 05-01.

Goh, K.I., Kahng, B., and Kim, D. (2001). Universal Behavior of Load Distribution in Scale-Free Networks. *Physical Review Letters,* 87 (27), 1-4.

Hamming, R.W. (1950). Error Detecting and Error Correcting Codes, *Bell System Technical Journal* **26**(2):147-160.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., Morris, M. (2006). Statnet: An R Package for the Statistical Analysis and Simulation of Social Networks. Manual. University of Washington, http://www.csde.washington.edu/statnet.

Handcock, M.S. (2003). Statistical models for social networks: degeneracy and inference. In: Breiger, R., Carley, K., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis. National Academies Press, Washington, DC, pp. 229–240.

Handcock, M.S. (2002). Statistical models for social networks:degeneracy and inference. In: Breiger, R., Carley, K., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis. National Academies Press, Washington, DC, pp 229-240.

Headquarters, Department of the Army (1992). *Field Manual 7-8*, *Infantry Rifle Platoon and Squad*. U.S. Army Infantry School, Ft. Benning, GA.

Holland, P. and Leinhardt, S. (1977). A dynamic model for social networks. *Journal of Mathematical Sociology*, 5, 5-20.

Holland, P.W., Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association* 76, 33–65.

Huisman, M., and Snijders, T.A.B. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological Methods and Research,* 32, 253-287.

Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M. (2008). ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24:3, 1-29.

Hunter, D. (2006). Curved Exponential Family Models for Social Networks. *Social Networks*, doi:10.1016/j.socnet.2006.08.005.

Hunter, D.R., Handcock, M.S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics* 15, 565–583.

Hunter, J.S. (1986). The Exponentially Weighted Moving Average. *Journal of Quality and Technology* 18, pp. 203-210.

Jehl, D. (1997). Islamic Militants Attack Tourists in Egypt. *The New York Times*, November 23, 1997. p. WK2.

Johnson, J.C., Boster, J.S., and Palinkas, L.A. (2003). Social roles and the evolution of networks in extreme and isolated environments. *Journal of Mathematical Sociology*, 27, 89-121.

Katz, L. and Proctor, C.H. (1959). The configuration of interpersonal relations in a group as a time-dependent stochastic process. *Psychometrika,* 24, 317-327.

Killworth, P.D. and Bernard, H.R. (1976) Informant accuracy in social network data. *Human Organization,* 35, 269-286.

Killworth, P. D. and Bernard, H. R. (1979). Informant accuracy in social network data III: A comparison of triadic structure in behavioral and cognitive data. *Social Networks,* 2, 19-46.

Kim, E. and Carley, K.M. (Working Paper). Confidence Intervals of Network Metrics.

Krackhardt, D. (1992). A Caveat on the Use of the Quadratic Assignment Procedure. *Journal of Quantitative Anthropology*, 3, 279-296.

Krackhardt, D. (2008). Center for Computational Analysis of Social and Organizational Systems Summer Institute, June, 2008. Carnegie Mellon University, Pittsburgh, PA.

Krackhardt, D. (1987a). Cognitive Social Structures. *Social Networks*, 9, 109-134.

Krackhardt, D. (1987b). QAP Partialling as a Test of Spuriousness. *Social Networks,* 9, 171-186.

Krackhardt, D. (1998) Simmelian ties: Super strong and sticky. In Power and Influence in Organizations (eds R. Kramer, M. Neale), pp, 21-38. Sage, Thousand Oaks, CA.

Leenders, R. (1995) Models for network dynamics: a Markovian framework. *Journal of Mathematical Sociology,* 20, 1-21.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. *KDD'05.*

Lospinoso, J. (2008). Utility Maximizing Networks. In: *Proceedings of the 2008 International Conference on Information and Knowledge Engineering*. Las Vegas, NV.

Lucas, J.M. and Saccucci, M.S. (1990). Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements. *Technometrics* 32, pp. 1-12

Martin, R. and Sunley, P. (2006). Path dependence and regional economic evolution. *Journal of Economic Geography*, 6, 395-437.

Marquand, Robert (2001). The tenets of terror. *Christian Science Monitor*,18 Oct 2001.

McCulloh, Ian. (2004). *Generalized Cumulative Sum Control Charts* (Master's Thesis, The Florida State University, 2004).

McCulloh, I., and Carley, K.M. (n.d.) *The Link Probability Model: An Alternative to the Exponential Random Graph Model for Longitudinal Data.* Unpublished manuscript.

McCulloh, Ian & Carley, Kathleen M . (2008). Social Network Change Detection. *Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-08-116*

McCulloh, I., Garcia, G., Tardieu, K., MacGibon, J., Dye, H., Moores, K., Graham, J. M., & Horn, D. B. (2007a). *IkeNet: Social network analysis of e-mail traffic in the Eisenhower Leadership Development Program.* (Technical Report, No. 1218). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McCulloh, I., Lospinoso, J., and Carley, K.M. (2007b). Social Network Probability Mechanics. *Proceedings of the World Scientific Engineering Academy and Society 12th International Conference on Applied Mathematics,* Cairo, Egypt, 29-31 December, 2007.

McCulloh, I., Ring, B., Frantz, T.L., and Carley, K.M. (2008). Unobtrusive Social Network Data from Email. In *Proceedings of the 26th Army Science Conference*. Orlando, FL: U.S. Army.

McCulloh, I., Webb, M., and Carley, K.M. (2007). Social Network Monitoring of Al-Qaeda. *Network Science*, 1, 25-30.

Montgomery, D.C. (1991). *Introduction to Statistical Quality Control*, 2nd Edition, John Wiley and Sons, New York.

Morrow, R. and Brown, D. (1994). *Critical Theory and Methodology*, pp. 199-225, Sage, California.

National Research Council (U.S.) Committee on Network Science for Future Army Applications (2005). *Network Science.* National Academies Press.

Naus, J. (1965). Clustering of Random Points in Two Dimensions. *Biometrika*, 52, 263-267.

Nelson, R. (1995). Recent evolutionary theorizing about economic change. *Journal of Economic Literature*, 33, 48-90.

Newcomb, T.N. (1961). *The Acquaintance Process*. Holt, Rinehart and Winston, New York .

Newman, M. (2005) Power laws, Pareto distributions and Zipf's law. *Contemporary Phys*ics, 46 323–351.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review* 45, 167–256.

Neyman, J. & Pearson, E.S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions Royal Society Series A.* **231**, 289-337.

Page, E.S. (1961). Cumulative Sum Control Charts. *Technometrics* **3**, 1-9.

Pastor-Satorras and Vesppignani, (2004). *Evolution and Structure of the Internet: A Statistical Physics Approach.* Cambridge University Press, Cambridge, UK.

Pattison, P.E., Robins, G.L. (2002). Neighbourhood-based models for social networks. *Sociological Methodology* 32, 301–337.

Pattison, P.E., Robins, G.L. (2004). Building models for social space: neighborhood based models for social networks and affiliation structures. *Mathematiques des Science Humaines* 168, 11–29.

Pattison, P.E., Wasserman, S. (1999). Logit models and logistic regressions for social networks. II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology* 52, 169–194.

Priebe, C.E., Conroy, J.M., Marchette, D.J., and Youngser, P. (2005) Scan Statistics on Enron Graphs. *Computational and Mathematical Organization Theory,* 11, 229-247.

Ring, B., McCulloh, I., and Henderson, S. (2008). Gathering and Studying Email Traffic to Understand Social Networks. In: *Proceedings of the 2008 International Conference on Information and Knowledge Engineering*. Las Vegas, NV.

Roberts, S.V. (1959) Control chart tests based on geometric moving averages. *Technometrics* **1,** 239-250.

Robins, G.L., Elliott, P., Pattison, P.E. (2001a). Network models for social selection processes. *Social Networks* 23, 1–30.

Robins, G.L., Pattison, P.E. (2005). Interdependencies and social processes: generalized dependence structures. In: Carrington, P., Scott, J., Wasserman, S. (Eds.), Models and Methods in Social Network Analysis. Cambridge University Press, New York, pp. 192–214.

Robins, G.L., Pattison, P.E., Kalish, Y., Lusher, D. (2007). An introduction to exponential random graph (*p**) models for social networks. *Social Networks*, 29, 173-191.

Robins, G.L., Pattison, P.E., Woolcock, J. (2004). Models for social networks with missing data. *Social Networks* 26, 257–283.

Robins, G. and Pattison, P. (2007) Interdependencies and Social Processes: Dependence Graphs and Generalized Dependence Structures. In: P. Carrington, J. Scott and S. Wasserman, Editors, *Models and Methods in Social Network Analysis*, Cambridge University Press, New York, 192-214.

Robins, G. and Pattison, P. (2001) Random graph models for temporal processes in social networks. *Journal of Mathematical Sociology*, 25, 5-41.

Romney, A.K. (1989). Quantitative models, science and cumulative knowledge. *Journal of Quantitative Anthropology*, 1, 153-223.

Rogers, Everett M. (2003). *Diffusion of Innovations*, 5th ed. New York, NY: Free Press.

Ryan, T. P. (2000). *Statistical Methods for Quality Improvement*. 2nd Ed, Wiley.

Saccucci, M.S. and Lucas, J.M. (1990). Average Run Lengths for Exponentially Weighted Moving Average Control Schemes Using the Markov Chain Approach. *Journal of Quality Technology* **22**, 154-159.

Sampson, S.F., (1969). Crisis in a cloister. Ph.D. Thesis. Cornell University, Ithaca.

Sanil, A., Banks, D., and Carley, K.M. (1995). Models for evolving fixed node networks: Model fitting and model testing. *Social Networks*, 17, 1995.

Schreiber, C. & Carley, K. (2004). *Construct - A Multi-agent Network Model for the Co-evolution of Agents and Socio-cultural Environments*. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report, CMU-ISRI-04-109.

Shewhart, W.A. (1927). *Quality Control*. Bell Systems Technical Journal.

Snijders, T.A.B., Steglich, C.E.G., Schweinberger, M. and Huisman, M. (2007). *Manual for SIENA version 3.1.* University of Groningen: ICS / Department of Sociology; University of Oxford: Department of Statistics

Snijders, T.A.B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* 3, 2.

Snijders, T.A.B. (2007). Models for longitudinal network data. In: P. Carrington, J. Scott and S. Wasserman, Editors, *Models and Methods in Social Network Analysis*, Cambridge University Press, New York, 148–161.

Snijders, T. A. B., Van Duijn, M.A.J.. (1997). Simulation for Statistical Inference in Dynamic Network Models. In Simulating Social Phenomena, (Ed. R. Conte, R. Hegselmann, and P. Tera) Berlin: Springer, pp. 493-512.

Snijders, T.A.B. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21, 149-172.

Snijders, T.A.B. (1990) Testing for change in a digraph at two time points. *Social Networks*, 12, 539-573.

Snijders, T.A.B. (2001) The statistical evaluation of social network dynamics. In: Sobel, M.E. and Becker, M.P. (Eds) *Sociological Methodology*, 361-395. Boston: Basil Blackwell.

Strauss, D., Ikeda, M. (1990). Pseudo-likelihood estimation for social networks. *Journal of the American Statistical Association* 85, 204–212.

Topper, C. and Carley, K.M. (1999). A Structural Perspective on the Emergence of network Organizations, *Journal of Mathematical Sociology*, 24(1):67-96.

Wald A. (1945). Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics* 16, 117-186.

Wald A. (1947). *Sequential Analysis*, Wiley, New York.

Wasserman, S. (1980). Analyzing social networks as stochastic processes. *Journal of American Statistical Association*, 75, 280-294.

Wasserman, S. (1977). *Stochastic Models for Directed Graphs*. Ph.D. dissertation, Harvard University, Department of Statistics, Cambridge, MA.

Wasserman, S. (1979). A stochastic model for directed graphs with transition rates determined by reciprocity. In *Sociological Methodology 1980* (Ed. Schuessler, K.F.) San Francisco: Jossey-Bass, 392-412.

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

Wasserman, S., Iacobucci, D. (1988). Sequential Social Network Data. *Psychometrika* 53:261-82.

Wasserman, S., Pattison, P.E. (1996). Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and *p**. *Psychometrika* 61, 401–425.

Wasserman, S., Robins, G.L. (2005). An introduction to random graphs, dependence graphs, and *p**. In: Carrington, P., Scott, J., Wasserman, S. (Eds.), Models and

Methods in Social Network Analysis. Cambridge University Press, NewYork, pp. 148–161.

Wasserman, S., Scott, J., and Carrington, P. (2007). Introduction. In: *Models and Methods in Social Network Analysis*. P. Carrington, J. Scott, S. Wasserman (Eds.) Cambridge Press (2007).

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. *Nature*, 393, 440-442.

Van de Bunt, G.G., Van Duijin, M.A.J., and Snijders, T.A.B. (1999). Friendship networks through time: An actor-oriented statistical network model. *Computational and Mathematical Organization Theory*, 5, 167-192.

APPENDIX A – Social Network Primer

In the 2005 National Research Council "Network Science" report, recommendation #1 stated,

> "The federal government should initiate a focused program of research and development to close the gap between currently available knowledge about networks and the knowledge required to characterize and sustain the complex global networks on which the well-being of the United States has come to depend." (p. 4)

Network Science differs from classical scientific methods in that it views the subject matter as being made up of many interacting entities that are called nodes. One application area of Network Science that has become extremely popular is Social Network Analysis (SNA). SNA looks at groups of people and their interactions. This type of analysis provides a methodology that does a very good job at explaining much of the complex behavior of these social groups. This work focuses specifically on detecting statistically significant changes over time in the observed social networks of several socially connected groups.

Social network analysis (SNA) examines relationships between social entities (i.e. people, groups, tasks, beliefs, knowledge, etc.). These entities are modeled with nodes and their relationships are modeled with links. Not all nodes are connected and some nodes may have multiple connections. This mathematical model is applicable in many content areas such as communications, information flow, and group or organizational affiliation (Tichy, et. al., 1979; Wasserman and Faust, 1994). SNA thus relies heavily on graph theory to make predictions about network structure.

To illustrate the importance of understanding social networks, I present an example of an informal network in a military organization. Consider the non-commissioned officer (NCO) support chain in an Army company of 150 soldiers. In this organization, the first sergeant (1SG) was new to the company. This was his second assignment as a 1SG. In his previous assignment, he brought some great ideas to his company and made significant improvements. The command sergeant major (CSM) thought that the 1SG would be able to make similar improvements in our example company. Unfortunately, the 1SG's ideas were not working in the same way that it did with the previous company. Good ideas can often fail when they are implemented poorly. Understanding the informal networks will provide insight into some important organizational dynamics.

Before we look at the informal network, we must understand the formal chain of support network in an Army company. The chain of support is illustrated in Figure 70.

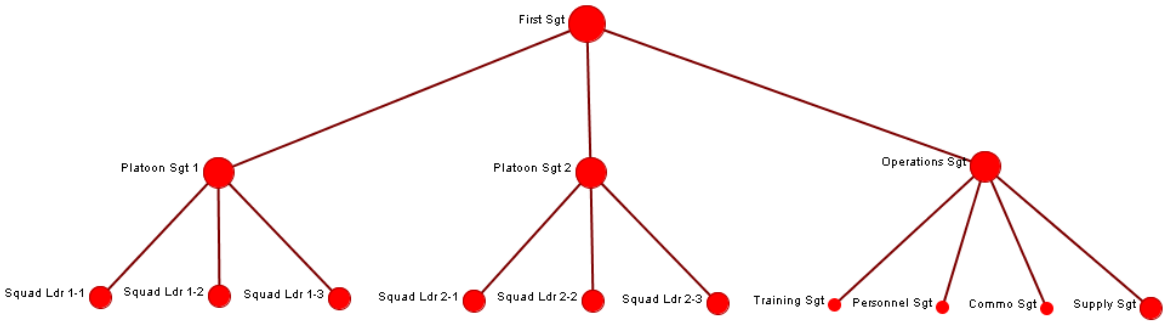


Figure 70. Formal NCO Chain of SupportFigure 1.

The 1SG is the senior NCO in the company. His pay grade is E-8 and he has usually served in leadership positions such as squad leader and platoon sergeant. There are two platoon sergeants (PSG) in the company as well as an operations sergeant. These NCOs are sergeants first class (SFC), or pay grade E7. The operations sergeant is sometimes referred to as the headquarters platoon sergeant. His soldiers are responsible for providing support to the platoons. The SFCs have usually served previously as a squad leader. Their direct supervisor is typically a lieutenant. While the chain of command consists of commissioned officers, the chain of support is a parallel network of NCOs that provide the officers with advice, experience, and logistic support. Each of the platoons has three squad leaders, who are staff sergeants (SSG), pay grade E-6. The company headquarters consists of one SSG who serves as the supply sergeant and three sergeants (SGT) of pay grade E-5. The three SGTs serve as the training sergeant, the personnel sergeant, and the communications sergeant. All the nodes represented in Figure 1 are sized according to their rank. Thus the 1SG is represented by a larger size node than the platoon sergeant, which is larger than the squad leader and so on.

Does the 1SG have the power necessary to make changes in the organization? According to the formal chain of support, he does. However, in our example, the 1SG is not effective in making change. His ideas are no different than they were in a previous company where he was effective. The difference lies in the informal network of the two companies.

Informal networks can be extracted in different ways. In the appendix, I present an unobtrusive method of estimating the informal network through collecting e-mail data on individuals in an organization. Command sensing sessions and surveys can also be effective in determining the informal networks in an organization.

The NCOs in this company were asked “What individuals in the company help you to get your job done.” The NCOs were not limited in the number of people they could list. This information represents the informal network. Individuals are not told who they have to go to for help or advice. People seek out this type of support from others for a variety of reasons. Sometimes informal relationships are determined by perceived competence, approachability, personality, common interests, or even racial/ethnic similarity. Figure 71 shows the informal network for our example company.

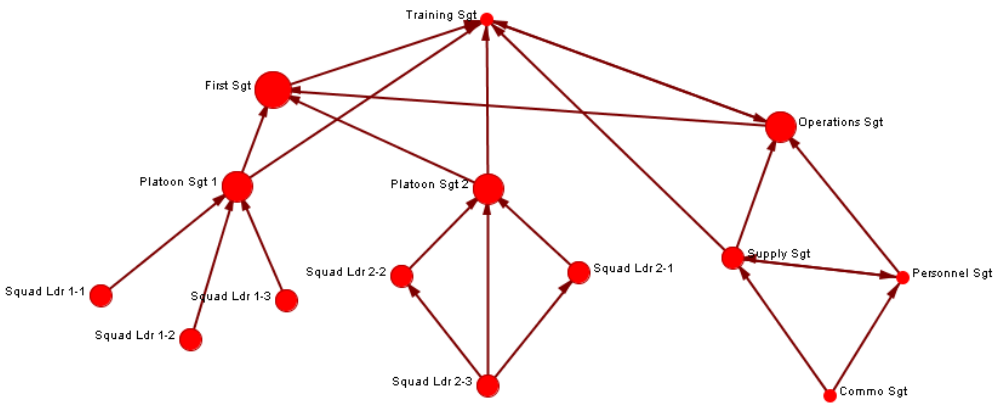


Figure 71. Informal NCO Network.

The NCOs in first platoon, on the left, tended to follow the NCO support chain for job assistance, as did second platoon in the middle. In second platoon, one of the squad leaders also sought assistance from the other two squad leaders in the platoon. Perhaps this squad leader was new and looked to his peers for mentorship. Members in the headquarters platoon had a similar dynamic. The node that stands out as unusual in the informal network is the training sergeant, who is looked to more than any other individual for assistance. Even the 1SG looks to the training sergeant for assistance.

There are several reasons for this informal organizational structure. The training sergeant had been serving in this capacity for almost a year and a half, which is a long time to serve in a duty position in the military. The training sergeant knew almost all of the NCOs that served in the battalion (higher) headquarters. He would often spend off-duty time with the battalion NCOs as well as with NCOs in his own company. In addition, he belonged to a couple of social groups which also had NCOs from different companies as members. Through this larger social network, the training sergeant was able to more effectively coordinate training resources through his friendship ties at the battalion level than other NCOs in other companies. The training sergeant was also able to mobilize his social network to find opportunities for squad leaders to send their soldiers to weapons ranges conducted by other units, use sections of training area reserved by other units, and coordinate similar training resources. The training sergeant enjoyed his position in the company. He liked the fact that senior NCOs would come to him for support. This position in the informal network gave the training sergeant more power.

When the 1SG first arrived at the company, he unknowingly hurt his effectiveness in the informal network. He wanted to assert himself as a leader and decided that he needed to make sure that soldiers maintained a high standard of military appearance and bearing. The training sergeant did not make a good impression on the 1SG. The 1SG felt that the training sergeant did not maintain a professional appearance and that he was cavalier and borderline disrespectful. This may not be entirely surprising, considering the power that the training sergeant held in the informal network. The 1SG made several corrections to the training sergeant and expressed his concerns about military discipline

to the operations sergeant. The training sergeant felt put-down and embarrassed by this first meeting. He did not openly discuss his feelings. As a result, the training sergeant was not eager to help make the new 1SG successful, or implement any of the 1SG's ideas. In addition, the training sergeant disagreed with many of the 1SG's ideas to make the company more efficient by requiring NCOs to brief him on training resources they were using for their training, and balancing those resources across the company. This of course, took away some of the power that the training sergeant enjoyed.

After a few months, it was time for the training sergeant to conduct a permanent change of station (PCS) and a new NCO assumed the duties of training sergeant and the 1SG's ideas slowly made the company better. However, these improvements might have been adopted more rapidly and more effectively had the 1SG been aware of the informal social network in the company. Perhaps, if the company leadership was able to monitor the social network, they would have prevented a relatively junior NCO from having so much power in the organization. Perhaps, the company leadership could have utilized the training sergeant's informal network through appropriate incentives. In any case, this example illustrates the importance of understanding the informal social network in an organization.

Social network analysis offers more than pictures: it provides an entirely new dimension of statistical analysis for organizational behavior. Traditional analysis focuses on individual attributes. Social networks focus on relationships between individuals. Traditional analysis assumes statistical independence, where social network analysis focuses on dependent observations. Traditional analysis seeks to identify correlation between significant factors and a response variable. Social network analysis seeks to identify organizational structure. The underlying mathematics behind traditional analysis is calculus, the language of change. The corresponding mathematics behind social network analysis is linear algebra and graph theory. These differences can be significant in terms of how someone looks at social dynamics.

Nodes are defined in terms of a set of n verticies, $V = v_1, v_2,\ldots,i,\ldots,j,\ldots, v_n$. The nodes are related to each other with a set of links L, where l_{ij} is a relationship between node i and j. A social network is often shown as an adjacency matrix, where the rows and columns correspond to the nodes and each cell a_{ij} can take on any numerical value corresponding to the link l_{ij}. In an unweighted network, cells are dichotomous and are represented as a 0 or a 1: the presence or absence of a link or relationship between nodes i and j. Networks where relationships between nodes are always mutual are called undirected networks, and their adjacency matrices will always be symmetric. Directed networks, on the other hand, can model both mutual and directional relationships. A value of 1 in cell a_{ij} represents a directed relation from node i to node j. In application, the diagonal of the adjacency matrix is rarely populated with anything but zeros, since interactions from an entity to itself are not generally interesting.

The potential complexity of interactions within even a small network, while discrete, grows exponentially with the number of entities. For this reason, algorithmic approaches to exploring distributions within constrained networks quickly become

computationally challenging. In a directed network, the number of possible relationships among nodes can be found by the expression, $n^2 - n$, where *n* represents the number of nodes in the network. The number of possible configurations of a network with a specified number of nodes (*n*) and links (*l*) can be thought of as the number of unique combinations of *l* nodes within the network given by,

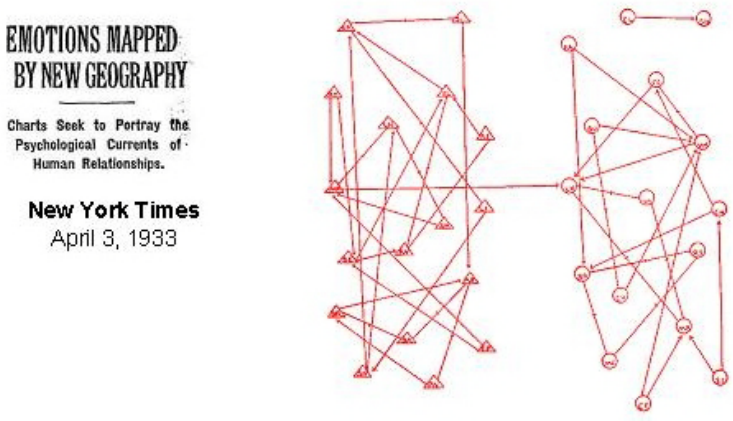
$$\binom{n^2 - n}{l} = \frac{(n^2 - n)!}{l! \cdot (n^2 - n - l)!}.$$

It follows that the total number of possible network configurations with *n* nodes can be represented by,

$$\sum_{l=1}^{n^2-n} \frac{(n^2 - n)!}{l! \cdot (n^2 - n - l)!}$$

A network of 30 nodes, for example, can be uniquely configured roughly 7.87 x 10^{261} different ways. With such large possible combinations of network structure, understanding how networks form and how they change over time is a complex problem.

In 1959, mathematicians Paul Erdős and Alfréd Rénia made revolutionary discoveries in the evolution of "random graphs." For our purposes a *graph* is synonymous with *network.* Erdős and Rénia use the term graph, referring to the field of mathematics called graph theory. This term was introduced by a chemist, Sylvester in 1878, as mathematicians were applying their ideas to chemistry. Social networks were independently introduced in the social sciences in 1933 by Moreno. Figure 72 is the first social network published in the New York Times. Erdős and Rénia's contribution to graph theory found great application in the social sciences, building on the work of Moreno and others, who use the term *network.*


EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

New York Times
April 3, 1933

Figure 72 The First Published Social Network, 1933.

In their eight papers Erdős and Rénia evaluate the properties of random graphs with n nodes and l links. For a random graph, G, containing no links, at each time step a randomly chosen link among the possible links is added to G. All of the possible links are equiprobable. A general model used to generate random graphs is as follows: "For a given p, $0 \leq p \leq 1$, each potential link of G is chosen with probability p, independent of other links. Such a random graph is denoted by $G_{n,p}$ where each link is determined by flipping a coin, which has probability p of coming up heads." In this model of random graphs each link has an equal probability of occurring or not occurring within the graph. This random graph model also assumes that all nodes in the graph are present at the beginning and the number of nodes in the network is fixed and remains the same throughout the network's life. Additionally, all nodes in this model are considered equal and are undistinguishable from each other.

Utilizing Erdos' theory of random graphs as well as the class of uniform distributions associated with these graphs, Holland and Leinheart (1981) developed a variety of statistical tests for the analysis of social networks. Using a uniform distribution these tests spread the total probability mass equally over all possible outcomes, therefore giving an equal probability to the existence of a link between any two nodes in the network. These statistical tests were used to develop a reference frame or constant benchmark to which observed data could be compared in order to determine how "structured a particular network was, or how far the network deviated from the benchmark." (Wasserman and Faust, 1994)

In 1969, Mark Granovetter proposed the strength of weak ties. In Granovetter's social world, our close friends are often friends with each other as well, leading to a society of small, fully connected circles of friends who are all connected by strong ties. These small circles of friends are connected through weak ties of acquaintances. In turn, these acquaintances have strong connections within their own circle of friends. The weak ties connecting circles of friends play an imperative role in numerous social activities from finding a job to spreading the latest fad (Granovetter, 1973; Aguirre, et. al., 1988). Close friends who have strong connections are often exposed to the same information; therefore, weak ties are activated to bridge out of our circle of friends and to the outside world.

Building off of Granovetter's model, Duncan Watts and Steven Strogatz (1998) developed the clustering coefficient, dividing the number of links of a node's first order connections by the number of links possible between these first order connections. This clustering coefficient illustrates the interconnectivity of a circle of friends, where a value close to 1 demonstrates all first order connections of a node are connected with each other. Conversely, a value close to 0 shows that a node's first order connections are only connected through that node.

Using the clustering coefficient, the Watts-Strogatz model of small world networks is the first to reconcile clustering with the characteristics of random graphs (Barabasi, 2003; Watts, 2004). According to the Watts-Strogatz model each node is directly connected to each one of its neighbors resulting in a high clustering coefficient.

By clustering alone, this model greatly increases the shortest path for a node to get to another node. However, by adding only a few random links between nodes of different clusters the average separation between nodes drastically decreases. This model while containing random links between nodes keeps the clustering coefficient relatively unchanged. While the Watts-Strogatz model originally did not add extra links to the graph, but randomly rewired some of the links to distant nodes, the addition of random links was proposed by Watts and M. Newman (1999).

According to Reka Albert and Albert-László Barabási (1999), the random graph theory of Erdős and Rénia was rarely found in the real world[6]. Albert and Barabási have found that many real world networks have some nodes that are connected to many nodes and others that are connected to few nodes. Their empirical tests showed that the distribution of the number of connections in many networks all followed a power-law distribution. These networks lack the characteristic scale in node connectivity present in random graphs, and therefore, are scale-free (Barabasi, 2003). As a result of the number of connections following a power distribution, hubs are created among nodes in the network. A hub is a highly connected node that contains most of the links in the network and creates short paths between any two nodes in the network.

Unfortunately, most of these network models do not have a defined probability space that proper statistical tests can be formed against. The statistical tests developed by Holland and Leinhardt (1981) apply for Erdős - Rénia random graphs, however, they do not apply to other network topologies (Scott, 2000; Albert and Barabasi, 2002; Borgotti, Carley, and Krackhardt, 2006; McCulloh et al, 2007). Without a defined probability space, detecting statistically significant change over time is impossible. Chapter 2 of this thesis explores the probability distributions of one network property, degree. The random graph of Erdős and Rénia is compared to the scale free graph of Albert and Barabasi. Chapter 3 provides a different frame work for viewing the probability space of a network. This alternate frame work is laid out as a basis and validation of multi-agent simulation for modeling networks. Chapters 4-7 then build off of this framework to propose change detection in social networks over time.

[6] I have found many social networks do not follow a power-law distribution and explore these claims in more detail in Chapter 2.

APPENDIX B - CONSTRUCT: Multi-Agent Simulation Model

Construct is a multi-agent simulation grounded in contructuralist theory (Carley, 1990, 1995). This multi-agent simulation is used to explore the performance and limitations of Social Network Change Detection (SNCD). The Link Probability Model (LPM), presented in Chapter 3 of this thesis, provides the stochastic engine for the multi-agent simulation. At each time step the link probabilities are determined by the nodes' perceived homophily, socio-demographics, and proximity. These social factors re-introduce the additional relational dependence missing in the raw LPM.

Construct is a dynamic-network multi-agent simulation model that can be used to examine the evolution of social, knowledge and activity networks in response to external interventions and the normal course of human interaction (Carley, 1990, 1991)[7]. Network evolution and the diffusion of information and beliefs through social networks can be examined using *Construct* (Carley, 1995; Hirshman & Carley, 2007b, Hirshman, Martin & Carley, 2008). *Construct* captures group dynamics under diverse cultural and technological configurations (Schreiber & Carley, 2004). Consequently it effectively models organizational change (Carley & Hill, 2001), socio-cognitive inconsistencies (Carley & Krackhardt, 1996), the impact of communication technologies (Carley, 1995; Carley 2002). To use *Construct* the researcher specifies both the agents replete with information processing capabilities (Hirshman, Carley & Kowalchuk, 2007a) and the networks in which they are embedded (Hirshman, Carley & Kowalchuk, 2007b).

Before, we explore the ability for network simulation to represent reality, we must first lay the foundational theory behind constructuralism as it applies to the multi-agent simulation *Construct.* Advances in both cognitive science and network theory have engendered the belief that it should be possible to develop analytical models of the relationships between individuals that would enable quantitative predictions of changes in interaction and that take into account both the self and the society, the individual and the group, the cognitive and the social. These advances have rekindled the dream, originally seen in social comparison theory (Festinger, 1954), cognitive dissonance theory (Festinger, 1957), and balance theory (Heider, 1958), that it is possible to build a mathematics of group change as a function of individual change, yet there is still a gap between the more cognitive and individual perspective in which changes in relationships between individuals result from independent dyadic encounters and the more social and structural perspective which changes in relationships between individuals result from gross changes to the group. Currently, a great deal of research is directed at bridging this gap. On the individual side the linking of symbolic interactionism and role theory can be viewed as a move to incorporate social or group factors into an otherwise predominantly cognitive.

Similarly, affect control theory is a move to incorporate the social, in terms of task constraints and social knowledge, into a cognitive and affective model of the individual's evaluation of; and hence determination of future action (Heise 1971, 1979, 1987; Smith-

[7] The Construct system itself is freely downloadable from the CASOS website, http://www.casos.cs.cmu.edu/projects/construct

Lovin 1987). The focus on the change in the individual or his or her relationships to an actual or a generalized other, treats the group or social world as present, but relatively fixed. This implicitly assumes that social or group behavior is somehow an aggregate of the results of independent encounters between pairs of individual. This last assumption is not exclusive to those who propose more cognitively rich models of behavior.

For example, we also see it in the work on status and dominance where hierarchies are viewed to result from independent dyadic encounters (Berger, Conner, and Fisek 1974; Rosa and Mazur 1979; Lamb 1986). On the up side, evidence is being amassed that group behavior cannot be accounted for by aggregating independent dyadic encounters (Chase 1974, 1980; Ridgeway and Diekema 1989) but is rather an emergent property of the simultaneous actions of all group members (Bales 1950; Homans 1950; Chase 1974, 1980; Fararo and Skvoretz, 1986). The mechanism by which such group behavior emerges remains elusive. As a step toward locating this mechanism, research in the structural and network traditions has been moving toward providing explanations, and hence predictions, of individual cognitive change in terms of the individual's social position.

This can be seen in Burt's model of action (1982) where perceived similarity and hence norms, attitudes, likelihood of adopting innovations, and so on is a function of social position. This is further supported by Krackardt's notion (1985, 1986, 1987) that the individual's social cognition (which he defines as the individual's perception of who interacts with whom) is a function of social position. These works reveal a more cognitive actor than that revealed by classic structuralist whose behavior is nonetheless socially situated. Yet, like the more cognitive individual models, these social models of individual change, still focus on the change in the individual while maintaining a relatively fixed social world. Thus, both the individual and the social perspectives treat the social world as fundamentally stable. Consequently, neither perspective provides a mechanism by which such individual changes can produce social change. Neither approach is sufficient to explain, let alone quantitatively predict, changes in the interaction patterns for all members of the society at once. Rather, the explanations of social change are highly contextual relying on situation specific factors, forces, and constraints such as goals, coercion, bureaucratization, change in group size, and membership rituals.

Every group has a population consisting of some number of individuals. In every group there is a set of information or facts that is potentially learnable by the members of the group. This set of information contains each piece of information that is known by at least one group member. The number of such facts will be denoted by K. At a particular point in time, say time period t. The individual, for any piece of information, such as k, either knows that fact or does not. This is denoted by $F(t) = 1$ if the fact is known by individual at time period t and 0 otherwise.

Every society has a culture, which can be thought of as the distribution of information across the population. At a particular point in time, say time period t, an individual i has a certain probability to interact with another other member of the society,

j. This is exactly where the LMP comes into consideration. Every society has a social structure, which can be thought of as the distribution of interaction probabilities across the population. The initial make-up of these probabilities and the transition of these probabilities at different time points are thus determined by several factors.

The first assumption of the *Construct* model posits that interaction leads to shared knowledge. It is generally demonstrable that individuals acquire information (and hence will come to share knowledge) during interactions. In order to represent this process a variety of simplifying assumptions are made. All pieces of information are entirely unstructured and undifferentiated. Thus, the individual may know conflicting information such as the sky is blue and the sky is green. Consequently, the overlap in what two individuals' know is just the sum of the pieces of information that they both know. When two individuals interact each communicates one fact to the other. Individuals always learn the piece of information that is communicated to them. Consequently, if individual *i* knows that the sky is blue and individual *j* knows that the sky is green and individual *j* communicates to individual *i* that the sky is green, the overlap in their knowledge increases. Hence they have more shared knowledge. All facts known by the individual are equally likely to be communicated.

According to constructuralism, both the individual cognitive world and the socio-cultural world are continuously constructed and reconstructed as individuals concurrently go through a cycle of action, adaptation, and motivation. During this process not only does the socio-cultural environment change, but social structure and culture co-evolve in synchrony. Carley (1991a) defined the following primary assumptions in describing constructuralism:

1. Individuals are continuously engaged in acquiring and communicating information
2. What individuals know influences their choices of interaction partners
3. An individual's behavior is a function of his or her current knowledge

In addition to these primary assumptions there were a series of implicit assumptions that upon explication serve to clarify and expand the primary assumptions. Following is an expanded list of assumptions, numbered to clarify their relation to the primary assumptions:

1a. Individuals, when interacting with other individuals, can communicate information
1b. Individuals, when interacting with other individuals, can acquire information
1c. Individuals can learn the newly acquired information thus augmenting their store of knowledge
2a. Individuals select interaction partners on the basis of relative similarity and availability
2b. individuals engage in interaction concurrently thus an individual's first choice of interaction partner may not be available.
3a. individuals have both an information processing capability and knowledge which jointly determine the individual's behavior

3b. individuals have the same information processing capabilities
3c. individuals differ in knowledge as each individual's knowledge depends on the individual's particular socio-cultural-historical background
3d. individuals can be divided into types or classes on the basis of extant knowledge differences.

These assumptions lead to a simulation template, which features a dynamic LPM as the stochastic engine. The LPM has convenient advantages in this capacity. The LPM avoids the issues of model degeneracy inherent in the ERGM. The probability of link occurrence is based on the historic presence of links and on social theory established in the literature, therefore, it does not use a Markov assumption or over specify a statistical model like other approaches. For these reasons, the LPM provides a reasonable stochastic engine for the *Construct* multi-agent simulation model. The multi-agent simulation simply adds additional relational dependence into a model that already performs well as we saw in Chapter 3, to make it more realistic and capable of evolution over time.

B.1 Importance of Simulation

The theoretical underpinnings of constructuralism as manifested in *Construct* lead us to a multi-agent simulation which utilizes a dynamic LPM as a stochastic engine for the development of knowledge diffusion and relationship building. What does this simulation provide the user?

The simulation provides an accurate, realistic simulation of social dynamics. We envision several ways in which this will be important to the military in particular and the wider academic audience in general.

Construct can be used as a valuable decision support tool for military commanders. The social dynamics of terrorist organizations, local culture, or friendly military forces can all be modeled with the simulation. A commander can war-game potential courses of action, and evaluate alternatives using *Construct.* It can be very difficult to reason through the many potential interactions, factors, and competing theories. This simulation provides a framework that is grounded in social theory, and validated against empirical evidence, that can be used to evaluate potential courses of action.

For example, a commander might consider detaining one or more suspected terrorists. By modeling the course of action in *Construct*, he can observe the impacts of removing the individual, on the organizations performance, situational awareness, and overall effectiveness. Given limited resources, the commander could even use the simulation to optimize the individuals to remove from the social group. The simulation provides the military analyst the ability to predict the future social dynamics of an organization. This is a powerful combat multiplier for today’s non-kinetic asymmetric war fighter.

The Army could also use *Construct* to evaluate the organizational structure of newly formed doctrinal units, such as the Future Combat System (FCS) operational units. The simulation can evaluate which personnel communicate more or less frequently. This can help inform efficient organization of soldiers from staff organizations to vehicle crews.

Focused research on social groups can follow better experimental design, and yield greater knowledge, if an array of research questions is first evaluated in simulation. Social dynamics are complex and it can be difficult to correctly reason through different scenarios. Simulation can provide insight that may shape the research questions to be more effective.

Finally, the normal behavior of an organization can be simulated many times. From the simulations, statistical distributions can be fit to various measures of group behavior. These statistical distributions can be used to evaluate statistical hypotheses or to detect statistically significant differences between observations of the group and normal behavior. This statistical framework, therefore, increases the relevant findings one can discover in socially dynamic organizations.

Using the social simulation program, *Construct* (Carley, 1990; Carley 1995; Schrieber and Carley, 2004), military units of varying size are simulated. A variety of changes will be introduced to the network at a known point. The Cumulative Sum (CUSUM) (Page, 1961), Exponentially Weighted Moving Average (EWMA) (Roberts, 1959), and Scan Statistic (Fisher and Mackenzie, 1922), statistical process control charts will be applied to several social network graph level measures taken on the network at each time step. The number of time steps between the actual change and the time that an SNCD method signals a change will be recorded as the Detection Length. The Average Detection Length (ADL) over multiple independently seeded runs is then a measure of the SNCD method's performance. The ADL will be compared for different changes and different SNCD parameters.

The basic military structure that will be simulated is an infantry training model. This is the most basic US military unit and is used for training soldiers and officers across the US Army Training and Doctrine Command (HQ, Dept of the Army, 1992). An organizational diagram is shown in Figure 73. Within this model, soldiers are organized into four man teams. Two teams and a squad leader form a 9 man squad. Three squads and a three person headquarters form a 30 man platoon. Three platoons and a 10 person command post form a company. Each soldier is trained in various skills that are distributed throughout the organization. Each team for example will have an automatic gunner, a grenadier and two riflemen. One member on a team will also be trained as a medic, another in demolitions, and two will be able to search enemy prisoners of war. Each soldier possesses individual skill in stealth, situational awareness, physical fitness, intelligence, military rank, and motivation. Homophily in these individual skills create stronger bonds between members of a unit which will increase their probability of communication. Organizational proximity will also affect communication, with individuals in the same sub-unit being more likely to communicate. The objective of the simulation will be to model communication within the military unit.

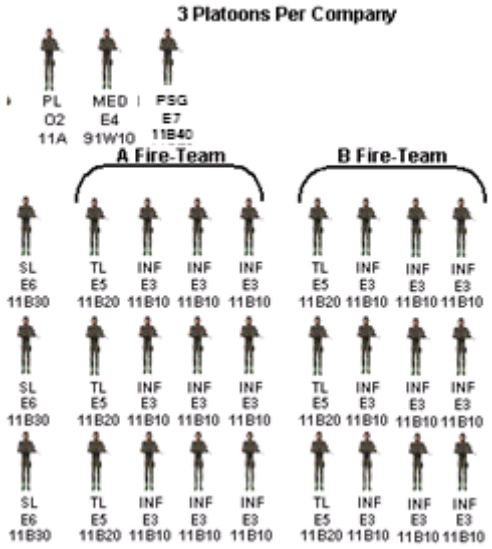


Figure 73. Platoon Organizational Diagram.

B.2 Docking

Other simulation studies have been performed on communication within military organizations (Kilduff, et.al., 2006; Rahimi, S., et.al.,). The Future Combat System (FCS) initiative has made extensive use of simulation to evaluate many system design considerations to include communication flow. The U.S.Army Research Laboratory's Human Research and Engineering Directorate (ARL-HRED) uses a simulation tool called C3TRACE (command, control, and communication: techniques for reliable assessment of concept execution). C3TRACE is a simulation environment that models organizations of varying sizes as they complete simulated tasks under various levels of workload. One study modeled communication within an infantry company very similar to the object of this paper's study. Some differences were introduced as part of the FCS revised personnel manning concept. The model equipped soldiers with differing communications devices to evaluate the impact on communication flow. The ARL-HRED focused on performance measures such as utilization, dropped messages, and decision quality. Their study did not look at the communication network however. The relationships between the modeled agents were not investigated. Change detection was not applied to the communication patterns in the simulated organization.

Another study investigated wireless platoon communication alternatives through simulation (Rahami, Mohamed, and Paredes, 2007). The focus of the study was on designing robust communication network topologies. While this study considered the relationships between simulated agents, they were essentially static. The model was a discrete event simulation created in Arena, which is better suited for process flow than it is for modeling relational network data. More importantly, the simulation did not explore dynamic changes over time.

The C3TRACE and Arena simulation models provide some insight into modeling a military organization, but they are not well suited to modeling dynamic network change over time. *Construct* was specifically designed to model relational network data and evolve it over time. In addition, *Construct* has the ability to vary the interactions of simulated agents based on their relative homophily, knowledge, and expertise. Dynamic social networks are easily exported to social network analysis software, where change detection can be evaluated. *Construct*'s ability to realistically simulate the social dynamics of an organization over time and provide network representations of communication makes this model uniquely well suited for SNCD exploration. An extensive search of the literature did not reveal any other relevant, similar models. Table 36 shows a comparison or docking of the three simulations.

Table 36. Docking: Construct, C3TRACE, Arena.

	Construct	C3TRACE	Arena
Simulated Organizations	Squad, Platoon, Company HQ	Squad, Platoon, part of Company HQ	Squad, Platoon
Size of Squad	9 men	9 men	9 men
Size of Platoon	30 men	49 men	undefined
Agent Details	Knowledge, expertise, beliefs, resources	Information quality, expertise	Knowledge
How agents interact	Uses an interaction sphere. Probability based on homophily or expertise.	Fixed in advance, based on doctrine	Interaction Strength is proportional to distance between agents
Output	Social network measures for each time step (Table 3).	Soldier utilization, soldier performance, and decision quality	Successful communication
Type of Simulation	Multi-Agent	Multi-Agent	Discrete Event
Virtual Experiment	Inject a change at a specified time point, to measure ARL	Change the agent interactions that will occur	No virtual experiment

While the three simulations are similar in terms of the organization being modeled, their objectives are different. The C3TRACE and Arena models are focused on measuring and improving unit performance. The purpose of this study is to realistically model the evolving social/communication network of an Army unit over time, so that methods of SNCD can be objectively compared to each other. In addition, *Construct* is much more sophisticated than the other methods at modeling how humans actually interact in an organization. This level of detail in modeling provides a much more rigorous test of the success of SNCD.

B.3 Verification and Validation

Verification and validation will focus on the communication dynamics within the organization. The input to the model is typical, frequent, verbal communication. The focus of the simulation is information sharing within a military organization. Issues of varying social capital for differing information is not considered in this model. Future research could investigate the effect of the importance of the information on communication dynamics. That type of investigation is beyond the scope of this thesis. Therefore, the model inputs are very simple.

There are several data sources that can be used to validate the process component of the simulation model. I have served as an instructor at the US Military Academy at West Point. During summer cadet training, I collected social network data on communication between soldiers ranging from team member to the company level. Data was collected for over 20 training missions from seven different company units. The communication within the simulation should be within the range of the real-world data. Simulated communication was also presented to four Army subject matter experts (Johnson, 2008; Gauthier, 2008; Smith, 2008; Trent, 2008) with recent combat experience to provide qualitative validation of model accuracy. The expert input was qualitative, where the soldiers were asked if a random sample of baseline (no change imposed) social networks appeared to reasonably describe communication patterns in an Army Infantry unit. Adjustments were made to weights placed on socio-demographic variables such as rank and job title to accurately reflect military communication. All four experts validated the final model.

The simulation output is validated by calculating several graph level social network measures for the baseline simulation and comparing those results to the data collected on cadet summer training. There was no statistically significant difference in the average betweenness, average closeness, or density. Unfortunately, more detailed information was not available. The output validation coupled with subject matter expert review provides reasonable evidence of the model's accuracy.

APPENDIX C - Analytical Derivation of Decision Interval

The Cumulative Sum (CUSUM) statistical process control chart is used to detect small changes in the mean of a random process. For quality control applications, it is desirable to detect any changes in the process mean as quickly as possible. For example, a manufacturing process may experience a change in mean as a result of tool wear, breakage, or adjustment, or any number of other unknown causes. The process is more likely to produce a product that does not meet quality specifications while the process mean is operating at its changed value. The product that does not meet quality specifications represents a financial loss to the company in terms of scrap product or re-working costs.

Methods that attempt to detect a change in a random process can sometimes signal that a change may have occurred, when in fact the process is still in-control. This is referred to as a false alarm. The probability of a false alarm occurring is sometimes referred to as Type I error. A false alarm in a manufacturing process can also represent financial loss for a company. If the company halts the process to search for a potential change that does not exist, the company is still paying for labor and overhead, while no product is being produced. Therefore, quality engineers must strike a balance between the probability of false alarm and the rapid detection of changes.

The determination of an appropriate balance between false alarm and rapid detection requires an expression that relates the probability of false alarm with control chart parameters. This was easily done with the Shewhart (1927) X-bar control chart, where an observation was compared against decision intervals set at $\mu \pm L\sigma$, where μ is the mean of the process, σ is the standard deviation, and L is the width parameter. The probability of false alarm, α, can be calculated from the expression, $\alpha = 2 * \int_{L}^{\infty} f(x)dx$, where $f(x)$ is the assumed symmetric probability density function of the process. The CUSUM control chart (Page 1954), on the other hand, was derived from the sequential probability ratio test (Wald, 1947), therefore the control chart statistic at each time point is conditioned on the previous time points. The CUSUM control chart statistic is given by, $C_t = \max\{0, Z_t - k + C_{t-1}\}$, where Z_t is the standardized observation at time t and k is an optimality constant. When the value of $C_t > h$, where h is the control chart's decision interval, the chart signals that a change in the process mean may have occurred. As a result of the max operator and the C_{t-1} expression, an analytical expression would somehow need to account for the nested conditional probability and the results are likely to be non-intuitive.

Several attempts have been made to provide quality engineers with insight into understanding the false alarm probability of the CUSUM. In situations where it is not necessary to know the precise probability of false alarm, but acceptable and rejectable quality levels have been established, an expression for the value k can be determined (Kemp, 1967). The optimal value of k is $k = (m_a + m_r)/2$, where $m \le m_a$ is an acceptable value from the random process, m, and $m \ge m_r$ is a rejectable value of m. It

has also been shown that the CUSUM is the most powerful test for detecting a change in the process mean of $2*k*\sigma$ (Moustakides, 2002). These results for the parameter *k* still do not provide a relationship between the parameter *k*, the decision interval, *h*, and the probability of false alarm. Expressions have been proposed that relate *k*, *h*, and a Brownian approximation to the expected number of observations until an in-control process signals a false alarm (Nadler and Robbins, 1971; Reynolds, 1975). This approximation was shown to overestimate the probability of false alarm (Reynolds, 1975). Thus, an accurate relationship between *k*, *h*, and the probability of false alarm has not yet been proposed.

In this chapter, Monte Carlo simulation is used to simulate the performance of the CUSUM on a random process consisting of independent and identically distributed observations for a range of false alarm probabilities between 0.001 and 0.05. A hybrid function is fit to the simulated values. The function provides a good fit to the simulated data with an R^2 value of 99.07%. Methods for using the newly proposed function for establishing CUSUM control chart parameters are discussed.

C.1 Method

Monte Carlo simulation was used to estimate the expected number of observations until an in-control CUSUM control chart signaled a false alarm. Results from the simulation were averaged over 100,000 independently seeded runs. Values of *k* ranged from 0.05 to 1.25 in increments of 0.05. Values for *h* ranged from 3.0 to 5.0. The specific values of *h* were adjusted for each setting of *k* to produce an expected number of observations until false alarm that fell within a range of 20 to 1000. This range was chosen for pragmatic reasons. The standard deviation of the number of observations until false alarm can be almost as large as the expected number of observations for an in-control process (Ewan and Kemp, 1960; Brook and Evans, 1972). Expected number of observations exceeding 1000 would, therefore, have such a deviation in the probability of false alarm as to be impractical. I submit that most practical applications using a probability of false alarm between 0.005 and 0.05.

The simulated data was plotted on a contour plot to observe trends in the data. The dependent variable was the decision interval. The optimality parameter, *k*, was an independent variable, and the expected number of observations until false alarm were contours. Initial observation suggested that the data exhibited the characteristics of exponential growth, where increasing *h* and *k* would lead to significant increases in the number of observations until false alarm. Three candidate functions were investigated: the exponential, power, and logarithmic functions, given by,

$$h(k) = \beta_1 e^{\beta_2 k}, \tag{1}$$

$$h(k) = \beta_1 k^{\beta_2}, \tag{2}$$

$$h(k) = \beta_1 Ln(k) + \beta_2. \tag{3}$$

Each function was fit to the data by the method of least square error. After analyzing the sum of square error and the R^2 values for the three functions, it was found

that the exponential and logarithmic functions estimated the decision interval better for the lower numbers of observation until false alarm, and the power function estimated the larger values better. In order to create a consistent function that provided accurate estimations for all values, a combination of the functions, or hybrid function was constructed to fit the full range of simulated data.

Several combinations of functions were investigated. These include linear combinations of two functions. The first hybrid function was the combination of the exponential and power function, given by,

$$h(k) = \left(\frac{1}{\beta_1 k}\right) e^{\beta_2 k} + \beta_3 k^{\beta_4} . \tag{4}$$

The alteration of the coefficient of the exponential function to $(1/\beta_1 k)$ allows the exponential portion of the equation to become insignificant as the optimality constant increases. This was applied because small changes in the *k* value cause larger changes in the decision interval as the expected number of observations until false alarm gets larger. This function was also fit to the data by the method of least squares.

A second hybrid function was created to combine the logarithmic and power functions. This function is also a linear combination of two functions, where the coefficient of the logarithmic function is replaced with the value $(1/\beta_1 k)$ and is given by,

$$h(k) = \left(\frac{1}{\beta_1 k}\right) \ln(k) + \beta_2 k^{\beta_3} \tag{5}$$

Both of these hybrid functions were designed to incorporate the strengths of both functions involved in the linear combination to create the best estimate of *h* based on the simulated data.

C.2 Results

The logarithmic-power hybrid combination provided the best fit to the simulated data. The function for the decision interval is therefore given by,

$$h(\lambda, k) = \left(\frac{\lambda^{0.1}}{5k}\right) \ln(k) + \left(0.53 \ln(\lambda) + \left(\frac{\pi}{10}\right)\right) k^{-0.89} , \tag{6}$$

where λ represents the expected number of observations until a false alarm occurs. Since all of the simulated data was used to fit the function, the performance of the function was evaluated using a 10-fold cross validation. The coefficient of determination, R^2, ranged from 98.79% to 99.81% with an average value of 99.07% across the 10 folds. The newly proposed function therefore provides a good approximation of the required decision

interval based on the optimality parameter and expected number of observations until a false alarm over a wide range of potential values.

The function can also be expressed in terms of the probability of false alarm. The probability of false alarm, α, is equivalent to the reciprocal of the expected number of observations until a false alarm, $\alpha = 1/\lambda$. Substituting α, in Equation 6 and simplifying provides an alternate expression for the decision interval given by,

$$h(\alpha, k) = \left(\frac{1}{5k\alpha^{0.1}}\right)\ln(k) - \left(0.53\ln(\alpha) + \left(\frac{\pi}{10}\right)\right)k^{-0.89}. \qquad (7)$$

This alternate expression provides an estimate for the decision interval based on a desired optimality parameter and the probability of false alarm. Both expressions are equivalent.

C.3 Discussion

A quality engineer can determine the parameters of the CUSUM control chart for a specific application, using the newly proposed expression in Equation 6 or Equation 7. First the engineer should investigate the costs associated with a change in the process. Does the product need to be scrapped, reworked, or sold at a lower price for certain quality characteristics? Based on these costs the engineer can determine the maximum acceptable and minimum rejectable process means for the process. The optimality parameter, k, should be set half way between these values and expressed in standardized units according to Ewan and Kemp (1960). The engineer must then decide on an acceptable risk level for false alarms. The engineer should choose a probability of false alarm, α, between 0.001 and 0.05 for most applications. The engineer could alternatively choose an expected number of observations until a false alarm, λ, if this is more intuitive for deciding upon a value. The choice of λ should be between 20 and 1000. Finally, the quality engineer can use the expressions proposed in this paper to determine an appropriate decision interval, h, for the CUSUM control chart.

Without these expressions, the quality engineer would either look up candidate values for h, k, and λ, as published in Van Dobben de Bruyn (1968), Nadler and Robbins (1971), Bagshaw and Johnson (1975), Vance (1986), Fellner (1990), Luceno (1999), or McCulloh (2004). Unfortunately, the published values of h, k, and λ, do not conform to the more ideal values associated with a specific process as determined in the method described above or according to the procedure laid out by Kemp (1962). Alternatively, a quality engineer could use an expression that approximates λ, under certain conditions as published in Reynolds (1975), or Luceno and Puig-Pey (2000). The expressions presented here more accurately estimate h for a range of k spanning 0.05 to 1.25, and a range of λ spanning 20 to 1000.

Perhaps the most useful applications of the newly proposed expressions are in software used to automate statistical process control. Many manufacturing processes include automated sampling, measurement, and control charting. The specific parameters for the process are still usually set by the quality engineer, however. With the analytical

expression for the decision interval, the parameterization can be automated as well. In addition, statistical process control is finding applications outside of manufacturing. The CUSUM has been used to identify changes in the organizational behavior of Al-Qaeda[8] (McCulloh et al, 2007), shifts in the membership commitment within on-line communities of practice (Galbreath, 2008), and changes within the semantic content of e-mail messages in the Enron corpus (McCulloh et al, 2008). For these new applications of statistical process control, determining appropriate control chart parameters is less clear. An analytic expression for the decision interval is necessary to broaden the potential application areas for statistical process control in general and for the CUSUM in particular.

[8] The Shewhart X-bar and Cumulative Sum statistical process control charts have been implemented in the software package Organizational Risk Analyzer (ORA) available from the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University, www.casos.cs.cmu.edu. ORA is used for social and dynamic network analysis.

APPENDIX D -- Longitudinal Network Data Collection

An important concern for the construction of a Link Probability Model (LPM) and monitoring a network for longitudinal change is a good source of data. Social network data can be time and resource intensive. Fortunately, email provides a rich source of longitudinal social network data that can be used for applications ranging from command and control, to military intelligence, to basic social science research. This Chapter reviews several methods available to extract email network data and compares them in terms of data quality and convenience of collection. In general, it is preferable to obtain email data directly from the central SMTP email server. In situations where this is not possible, alternative approaches presented here can be useful. These techniques for analyzing email data have been automated in the Organizational Risk Analyzer (ORA) software, which is freely available to DoD and academia.

Email has significantly changed how people communicate and interact. In many ways communication is easier and more reliable with email, however, there are many new challenges introduced. Over the past decade, many people have turned to email as the primary means to send information and to communicate (Ducheneaut, & Bellotti, 2001). It has enabled groups to work together, socialize and collaborate across any distances and outside of structured organizational boundaries. When organizational relationships do exist, email traffic among that group often mirrors this structure (Diesner, Frantz & Carley, 2005; Frantz & Carley, 2008; Tyler, Wilkinson, & Huberman, 2003). As a result, studying and analyzing communication patterns of email traffic can provide much insight into not only how an organization is structured, but also into how it actually operates (Carvalho, & Cohen, 2007). For example, a supervisor may typically send email to all his immediate subordinates and, likewise, those subordinates will respond. An increase in peer to peer collaboration may indicate that problems are being solved at a much lower level. Individual agents that connect disconnected groups might represent organizational vulnerabilities. Identifying these patterns from collected email data is extremely useful in identifying the underlying social network behavior of an organization.

I present two general methods for gathering and analyzing email data along with an analysis of each of these methods. During the course of this study, I gathered client-side email data over a seven month period to reveal the social network of a group of 24 mid-career Army officers. I also employed a centralized data collection procedure over a five month period directly from the central Simple Mail Transfer Protocol (SMTP) email server. The data collection schemes are compared in terms of data quality, ease of collection, and subject cooperation.

These email collection methods have been automated in a feature called CEMAP II contained in ORA (Carley, et al., 2008) --- a software package from the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University (Frantz & Carley, 2008b). The ORA software program is freely available to people in the DoD and at academic institutions at www.casos.cs.cmu.edu.

D.1 Background

Gathering email related data has shown to identify actual social and communal patterns among the email users (McCulloh et al, 2007; 2008). A collaborative group at Hewlett Packard Labs demonstrated that simply gathering the "TO" and "FROM" fields from a large collection of email messages can produce community structure when applied to a graph representation (Tyler, Wilkinson, & Huberman, 2008). This study focused on email data only at the organization's central mail server. In contrast, *Themail*, a visualization which shows an individual user's email exchange presents a visual network analysis of a user's email content simply by analyzing the archived mail on his or her personal computer (Viégas, Golder, & Donath, 2006). Users in this study were required to manually upload their entire Microsoft Outlook archive folder for analysis. Similar to this technique, Gloor and Zhao (2004) developed a software tool, TeCFlow, which gathers email data from a user's computer contained in various mailboxes and outlook archived files and stores that data into an SQL-database.

Communication via email can be divided into two types of relationships: the human-computer interaction; and the computer-computer interaction. People are usually most familiar with the human-computer interaction, where they sit at a computer, write an email, and push "send"; or they login to their email account and read messages contained in the "inbox". The computer-computer interaction, is actually an automated exchange between two computers, often with several other computers serving as intermediaries in the delivery process. A message sent from one computer is received by the target computer(s), in its electronic form, via a client software program that ultimately copies the email message from its host server. The email message is stored on a designated central server until the receiver "picks up" the message from the server. This process is the electronic version of picking up a package at the post office. The electronic email can be delivered to the post office repository for you to physically pick up, or directly to your personal mail box for you to pick up. Once a target computer picks up the message, the human-computer interaction allows the human to read, print, or store the electronic message via their email client software.

There are several different ways in which email can be delivered through the computer-computer interaction in the world-wide electronic email architecture. The message can be delivered to the equivalent of post office lobby-box, called an IMAP server. The email can be delivered to a personal mailbox, called a Post Office Protocol (POP) server. The email can also be routed through a Microsoft Exchange (MSEx) server. There are many technical differences between these email servers, but their purpose is the same. However, the principal difference between an IMAP and a POP email server is the storage feature of the server. An IMAP server will allow you (or your email client software) to persist, or store, your email physically on that server. A POP server only serves as a temporary holding station for a message that is removed once it has been retrieved by your email client software. An MSEx server is a Microsoft proprietary system that is widely used throughout the DoD. While it has some additional security features, it is more difficult to extract email network data from this system because of the propriety data format that Microsoft institutes. An IMAP server is designed to store the message even after the email has been initially retrieved. It should

be noted that the POP protocol calls for an email to be removed from the incoming mail box once it has been retrieved, however some software extensions do allow for a read-only access to the POP inbox, resulting in the message remaining in the inbox when retrieved and is therefore managed by the client software level. The popular Yahoo mail service implements this feature for paying customers.

Once a target computer receives an email, the human-computer interaction involves the computer displaying the message using client software. Email messages at the computer-computer interaction level are most often formatted in a world-wide standard format called MBOX. MBOX allows for different email client software programs to access the email from the server without confusion. The MBOX format specifies two sections of the email, the header section and the body section. The header section includes the From:, To:, CC:, BCC:, Subject:, and Date:, information. The body section contains the message text and any attachments to the email.

The MSEx server does not store messages in the MBOX format. Microsoft's proprietary standards create technical and licensing hurdles in accessing email data directly from the server in any manner other than using Microsoft software. Unfortunately, the MSEx format is widely used throughout DoD, making email data extraction more difficult. There are three approaches that we have discovered for extracting email content from an MSEx format. One approach is a custom client-side visual basic patch (McCulloh, et. al., 2007). Another client-side approach involves using .NetMap, which is a plug-in for Microsoft Excel 2007 that extracts email data from a proprietary *.pst file into an Excel format. The data can then be manipulated or saved to other file formats. The third approach involves parsing header data from a server log file. These approaches will be discussed in more detail in this paper. Analysis of dyad counts will be used to compare the performance of a client-side data collection with a centralized data collection.

D.2 Method

This study involves monitoring the email traffic of 24 mid-career Army officers in a one-year graduate program administered jointly by Columbia University and the U.S. Military Academy (USMA). Each of the officers participating were asked to sign a consent form in accordance with the institutional review board (IRB), approved by the USMA Human Subjects Research Review Board allowing their data to be collected for research purposes.

As part of this study, the participants permitted me to place a custom developed program (Appendix 1) that works in conjunction with their MSEx Outlook email accounts. This program allowed me to collect email data from the sent items folder found on participants' personal computers. The information included all of the header information associated with an email. I did not view or include the body of the email in the study. I was also able to collect similar email header information directly from the log files maintained by the Directorate Of Information Management (DOIM). The data collected from the custom program is referred to as the *Client-Side Method*, while data collected from the DOIM log files is referred to as the *Centralized Method.* I did not

investigate .NetMap as an approach as it has identical underlying email data-sourcing capabilites and functionality only with a different, al beit a more elegant, user interface. The email data collected from all methods was analyzed using a dynamic network analysis approach (Carley, 2003).

D.2.1 Client-Side Method

A client side Visual Basic for Applications (VBA) program was installed on the personal computers (PC) of all participants, in the session window of their Microsoft Outlook. This data collection approach was designed to overcome the difficulty in pulling information from a subject's sent mail folder in a proprietary Outlook Exchange system. This patch is easy to implement in Visual Basic and works harmoniously with Microsoft Outlook. The principal investigator could then compile the data from all participants into one master file and ensure anonymity of the names.

One of the chief advantages in managing a client-side patch is the low-level control in gathering data. A researcher does not have to obtain permissions from a network administrator to collect email data. They merely need the consent of the monitored individuals, who must login to their Outlook for the client-side patch to be installed. Furthermore, the program designers can pick and choose which data to import from the local client. If, for example, we wanted to include message content, then that could have been an option. We could have also just as easily gathered incoming email traffic, as opposed to only monitoring outgoing mail. This could provide further insight into areas such as whether a user classifies email as junk mail, whether they delete an incoming message, or even if they flag a particular message as important.

Managing the data collection from the individual participant required minimal effort. Once fully developed and installed, the Visual Basic patch is little to no overhead on the part of the user to manage. Furthermore, these participants felt more comfortable knowing that they have some degree of control in how the data is collected. While this could impede the data collection process, the subjects felt more comfortable knowing what was actually monitoring their email. Initially, most of the participants' email were sent to other students or people affiliated with their graduate program. Within two to three weeks, the participants began to email family members and friends. We suppose that this represents an increased level of trust. In the beginning, participants felt that their email needed to appear strictly business related. Gradually, as they incrementally sent personal email messages while they were "at work" without any negative consequences, they began to feel comfortable and appear to have returned into a normal cadence of email communication. Most of the participants knew how to remove the patch when their participation in the project ended. Several participants said they felt more comfortable knowing that the software sending the principal investigator information was on their computer, and that "Big Brother" was not pulling their information from somewhere else.

D.2.2 Centralized Method

As an alternative method, we developed a software application which analyzes email data gathered directly from a centralized email exchange server. This software gathered data over a five month period and extracted those email messages which were sent and received from the participants in this study. The server log files contain the email header information. This information was parsed into the same format as the client-side method.

With this method of data collection, the participants were not aware of the precise time that the collection process started. They did provide consent in accordance with the IRB, however, we were not required to inform them of the exact date when collection would begin. There was no significant observable change in the participants' pattern of communication. The centralized method was completely unobtrusive.

D.3 Dyad Analysis

It was not clear at the beginning of this investigation whether email communication within a homogenous group of people would appear random, if it would remain relatively consistent from week to week, or if there were identifiable factors that would affect changes in network structure. To investigate the structure of the network, we computed the dyad count. The dyad count, defined as the communication between two nodes (Wasserman, & Faust, 1994) distinguishes three different types of communication: asymmetric, mutual, and null. In an asymmetric dyad, one node talks to another, but does not receive a response. This type of communication could be an example of a group that has members who are sending out information. A mutual dyad signifies two nodes communicating with each other. This type of communication might occur in a group that collaborates equally, or one in which subordinates verify or clarify directives. Finally, a null dyad occurs when two nodes which are part of the network do not have any communication activity. In a dyad count, we conduct a census and tabulate the number of null, mutual, and asymmetric dyads. With 24 members in our study comprising a network, there exists 276 combinations of possible undirected pairs. Each of the 276 dyads could be either null, mutual, or asymmetric. The dyad counts are compared for data collected with the client-side method, centralized method, and with a calendar of significant events.

D.4 Results

There were significant differences in the client-side and centralized methods of data collection. The data from both methods was coded as a meta-network (Carley, 2002). Considering that the participants are a random sample of mid-career Army officers that all fulfill the same role of student in the organization, we might hypothesize that the email relationships formed in the network are random. Given that there are 24 nodes in the network, there exist 24 x 23 = 552 possible dyads. We can test the hypothesis:

H_o: Graph ~ Binomial(552, 0.5)
H_A: Graph ≠ Binomial(552, 0.5),

using the test statistic $z = (l - E(l)) / Sqrt(V(l))$, where l is the number of directed links in the graph. This reduces to $z = (l - 276) / 11.75$, where l is the sum of the mutual and asymmetric dyad counts. Under the null hypothesis, this number follows a standard normal distribution. The p-value was significant at the 0.05 level for most weeks, providing evidence to reject the hypothesis that email communication patterns are random binomial with a probability parameter of 0.5. A week with a corresponding p-value that was not significant at the 0.05 level can be identified in Table 37 by the 95% confidence interval on the Binomial parameter p that includes 0.5.

A confidence interval on the probability of communication can be constructed for each week according to the expression given by,

$$\hat{P} \pm z_{\alpha/2} \sqrt{\hat{P}\left(1-\hat{P}\right)/552}$$

where $\hat{P}$ is the maximum likelihood estimate of the unknown parameter p in the assumed binomial distribution and equal to l / 552. Table 37 shows the mutual, asymmetric, and null dyad counts recorded using the client-side and centralized methods. The right most column of Table 37 shows the 95% confidence interval on the random probability of communication. A confidence interval that spans 0.5 will correspond to a significant p-value in the random binomial hypothesis test above. For each week in Table 37, two values are shown for each of the dyad counts: Mutual, Asymmetric, and Null. The numbers in the top of each cell in Table 37 correspond to the client-side data collection method. The numbers in the bottom of each cell in Table 37 correspond to the centralized data collection method. The data presented in Table 37 corresponds to the time period beginning with the first week of the Spring semester and ending with the week before Spring break. The students took their comprehensive exam following Spring break and then began to transition to their military duties at West Point. Therefore, this data represents a reasonable time period for comparison of the client-side and centralized methods of data collection.

It can be seen in the Confidence column of Table 37 that there is a statistically significant difference in the probability of communication between the client-side and central data collection methods for all weeks, by observing that the 95% confidence intervals do not overlap. In all cases, the client-side method underestimates the probability of communication in the network. The general pattern of the probability parameter is correlated at a value of 0.69, which is low considering they are estimates on the same group of individuals during the same week. The client-side data collection method is therefore biased.

Table 37. Recorded directed links using client-side and central methods.

Week	Mutual	Asymmetric	Null	Confidence
13 Jan 2008	0 54	44 89	232 133	(0.06,0.10) (0.22,0.30)
20 Jan 2008	6 218	88 83	182 0	(0.14,0.20) (0.50,0.59)
27 Jan 2008	0 118	78 92	198 66	(0.11,0.17) (0.34,0.42)
3 Feb 2008	8 202	162 81	106 0	(0.27,0.35) (0.47,0.55)
10 Feb 2008	0 112	148 100	128 64	(0.23,0.31) (0.34,0.42)
17 Feb 2008	6 230	114 79	156 0	(0.18,0.25) (0.52,0.60)
24 Feb 2008	26 204	108 92	142 0	(0.21,0.28) (0.49,0.58)
2 Mar 2008	84 320	192 51	0 0	(0.46,0.54) (0.58,0.66)
9 Mar 2008	26 204	143 73	107 0	(0.27,0.34) (0.46,0.54)

* Client-side dyad counts are above central dyad counts.

The dyad count analysis can provide additional insight into the organizational dynamics of the participants by comparing their probability of interaction to significant events on their academic calendar. We restrict our investigation to data collected using the centralized method since it is complete. The centralized method captures all data sent or received through the central server. The maximum likelihood estimate of the parameter, *p*, in the binomial distribution of dyads is plotted over time and displayed in Figure 74.

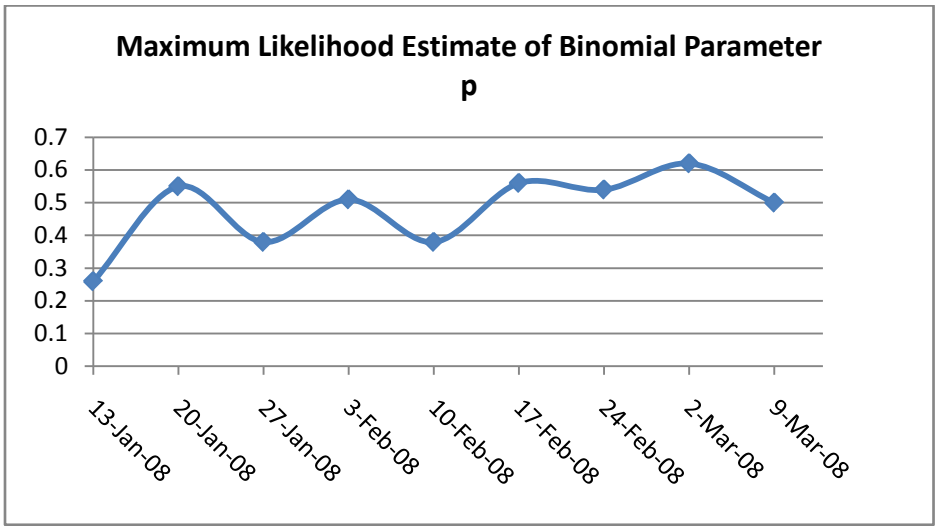


Figure 74. MLE of parameter p using centralized method.

The lowest MLE of *p* is shown in the first week of the semester, when the participants were just returning from Christmas leave. This was followed by an increase in communication as the group begins to plan for group academic assignments, carpooling, and other administrative issues. The low points in the MLE of *p* occur during the weeks of 27 January and 10 February when major group academic projects or presentations were due. This is consistent with the findings of McCulloh, et. al. (2007) who observed a similar decrease in email communication during times of group activity. They hypothesized that during times of increased face-to-face communication, people communicate verbally and have less time and need for email communication. Furthermore, during these times of increased subgroup activity, people have less time to write and respond to emails from individuals outside of their immediate subgroup. Following the group assignments due during the week of 10 February, the next major academic event was the comprehensive exam following Spring break.

A similar dyad analysis for the client-side method is shown in Figure 75. The characteristic dip in email communication corresponding to group activity is not clear. A careful review of the participants' academic calendar does not reveal any activities or events that would explain the behavior of the plot in Figure 75. This further suggests the importance of centralized email data collection.

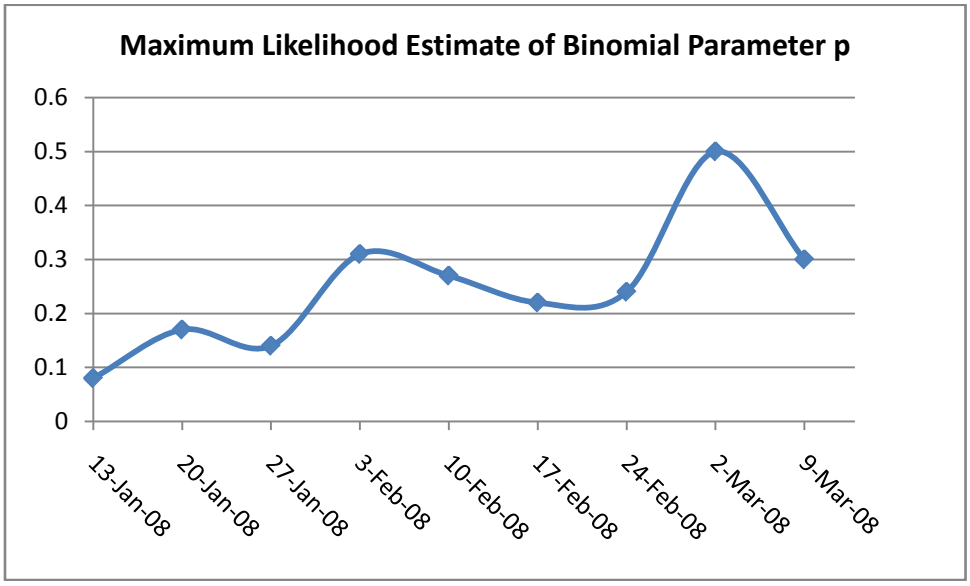


Figure 75. MLE of parameter p using client-side method.

The client-side method of data collection is not completely without merit. It can still be seen in Figure 75 that the first week has the lowest MLE of *p*. There is also a dip in the plot for the group assignment for the week of 27 January. The identification of the week of 10 February is missed however. This suggests that even client-side data can provide some insight into group behavior. This may be an appropriate method to use when complete centralized data is unavailable. Centralized data may be unavailable for reasons of security, privacy, damage, or other technical difficulties. In these situations, the client-side method may still provide valuable information on social network behavior.

D.5 Discussion

I found that the primary advantage to utilizing a server-side method to gather data is the improved data integrity. Every user with an email account must both send and receive data from that account's associated mail server. Therefore, to ensure that all data is gathered it must be collected at its source. All data contained within the centralized server is available for collection, such as from, to, cc, bcc, subject, time of receipt at the mail server, etc. Copying data directly from the server allows the social network analyst to accurately study all email communications within a study group for those utilizing their given email address.

Implementing a server based application also precludes the subjects involved in the study from corrupting and inserting bias into the data. With a client-side application, users had the ability to turn off, remove or disrupt the execution of the program used to monitor email. With a server-side collection technique, the clients are completely unaware or knowledgeable about when or what is collected. I found that while it takes more overhead to initiate the retrieval of email traffic from a mail server, there is surprisingly little overhead on the part of a server administrator to actually assist the research effort in gathering data. Since log files are typically stored in a common location on the server, the administrator need only make these files available. When operated across a network, he/she can easily copy these log files to a common location from where the server-based data collection program can import the data.

By presenting two methods for gathering and analyzing email data, we have shown both advantages and disadvantages for the social network analyst. These strengths and limitations must be considered by any social network analyst when studying email traffic. Even though gathering data at its source does provide better data integrity, such data collection means are not always feasible. In these cases, email data collected in a decentralized manner can still provide insightful analysis of the underlying social network.

I advise a practitioner to be highly sensitive to the privacy implications of this process, especially in the public and private sectors. People within the military typically do not maintain the expectation of email and internet privacy. This may not be true in other populations. Care must also be exercised with interpreting the results of these types of social networks. It is important that trained social network analysts provide proper interpretation of the organizational behavior, while respecting the privacy of individual identities. Revealing the position an individual maintains in the social network of an organization may lead to an overall decrease in trust and adversely affect the leadership climate within the organization. When used properly, however, social network analysis can provide a wealth of valuable information to the organization. Several commands within the Army have already implemented social network data collection from email. These methods have been automated in the software package ORA, which is maintained by CASOS at Carnegie Mellon University and can be freely downloaded by the military and academia.

Future research in this area will likely explore the impact of cellular phone communication and blackberries on social networks within the military. This line of research will further support the efficacy of Netcentric Operations within the Army. Focused research into the usage of cell phones, blackberries, e-mail, and face-to-face communication during major group activities will provide greater insight into social network data collection. Understanding the desired channels of communication for military leaders, may significantly contribute to shaping the communication technologies that the DoD invests in. This line of research may also provide data for real-time monitoring of organizational change. It will certainly be valuable in enhancing command and control systems used by the military.