UNITED STATES MILITARY ACADEMY

WEST POINT, NEW YORK

HONORS THESIS

ANALYSIS OF COMMUNICATION NETWORKS

by

Cadet Daniel Baller

April 2008

Thesis Advisor:	MAJ Anthony Johnson
Second Reader:	MAJ Ian McCulloh
Second Reader:	COL Steve Horton

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE April 2008	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE Analysis of Communication Networks **6. AUTHOR(S)** Cadet Daniel Baller	**5. FUNDING NUMBERS**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) United States Military Academy West Point, NY 10996	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) NA	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

Real world communication networks are based on an underlying edge probability structure. These networks are continuously changing and thus reliably estimating network measures becomes difficult. Therefore a new framework is proposed to determine the underlying probability distribution of specific communication network measures. When two individuals are socially connected, their communication may vary from day to day, yet their underlying relationship remains unchanged. In this case, estimates of network measures, such as density, degree centrality and others may be severely affected by the occurrence or absence of observed communication ties between individuals.
Two communication networks of groups of mid-career Army officers are modeled from empirical data using the network probability matrix (NPM) proposed by McCulloh and Lospinoso (2007). The NPM provides a framework to model a communication network by estimating the underlying edge probabilities between each dyadic pair in a network. This framework can model a specific social group regardless of their network topology: random, small-world, scale-free, cellular, etc. Monte Carlo simulation is used in conjunction with the NPM to generate 100,000 instances of each communication network. A statistical distribution is fit to the density measure. This probability distribution can then be used to detect statistically significant changes in density.
The linear dependence of each agent in the networks is also analyzed to explore adjacency matrix properties such as: rank, determinant and singular values via singular value decomposition. We hope to establish an interval for which rank deficient adjacency matrices occur.

14. SUBJECT TERMS			**15. NUMBER OF PAGES**
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT OF** UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

ANALYSIS OF COMMUNICATION NETWORKS

Cadet Daniel Baller
Cadet Captain, Military Intelligence
B.S., United States Military Academy, 2008

Submitted in partial fulfillment of the
requirements for the degree of
BACHELOR OF SCIENCE
in **MATHEMATICAL SCIENCES**
with Honors
from the
UNITED STATES MILITARY ACADEMY
April 2008

Author: Cadet Daniel Baller

Advisory Team: MAJ Anthony Johnson
Thesis Advisor

MAJ Ian McCulloh
Second Reader

COL Steve Horton
Second Reader

COL Michael Phillips
Chairman, Department of Mathematical Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Real world communication networks are based on an underlying edge probability structure. These networks are continuously changing and thus reliably estimating network measures becomes difficult. Therefore a new framework is proposed to determine the underlying probability distribution of specific communication network measures. When two individuals are socially connected, their communication may vary from day to day, yet their underlying relationship remains unchanged. In this case, estimates of network measures, such as density, degree centrality and others may be severely affected by the occurrence or absence of observed communication ties between individuals.

Two communication networks of groups of mid-career Army officers are modeled from empirical data using the network probability matrix (NPM) proposed by McCulloh and Lospinoso (2007). The NPM provides a framework to model a communication network by estimating the underlying edge probabilities between each dyadic pair in a network. This framework can model a specific social group regardless of their network topology: random, small-world, scale-free, cellular, etc. Monte Carlo simulation is used in conjunction with the NPM to generate 100,000 instances of each communication network. A statistical distribution is fit to the density measure. This probability distribution can then be used to detect statistically significant changes in density.

The linear dependence of each agent in the networks is also analyzed to explore adjacency matrix properties such as: rank, determinant and singular values via singular value decomposition. We hope to establish an interval for which rank deficient adjacency matrices occur.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION **3**
A. BACKGROUND 4
1. Erdös-Rnia Random graphs 4
2. Statistical tests 5
3. Strength of weak ties 5
4. Small World Networks 6
5. Scale Free Networks 6
6. Fitness Model 7
7. Winner Take All Model 7
8. Network Probability Model 8
9. Distribution of Network Level Measures 8
B. DATA 8
II. METHODS **11**
1. Construction of The NPM 11
2. Simulation Generation 11
3. Reliability and Consistency 12
4. Distribution Fitting 12
III. LINEAR DEPENDENCE OF THE AGENTS **13**
IV. RESULTS **15**
A. DISTRIBUTION OF THE DENSITY 15
B. ANALYSIS OF LINEAR DEPENDENCE 17
V. CONCLUSION **23**
APPENDIX A. ADDITIONAL STATISTICAL TESTS **25**
LIST OF REFERENCES **29**

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

1. Stepwise Plot of Density Data for Net07 and CDF of the Normal Distribution 16
2. Stepwise Plot of Density Data for Net05 and CDF of the Normal Distribution 17
3. Histogram of Density for Net07 18
4. Histogram of Density for Net05 19
5. List Plot of Sorted Determinants for Net07 20
6. List Plot of Sorted Ranks for Net07 20
7. List Plot of Sorted Ranks for Net05 21
8. Stepwise Plot of Rank Data for Net07 and CDF of the Normal Distribution . . 21
9. Stepwise Plot of Rank Data for Net05 and CDF of the Normal Distribution . . 22
10. Box Plot of Densities for Net07 25
11. Box Plot of Densities for Net05 26
12. Normal Q-Q Plot of Densities for Net07 27
13. Normal Q-Q Plot of Densities for Net05 28

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

IV.1. t-test of Average Hamming Distances for Net07 Data 16
IV.2. t-test of Average Hamming Distances for Net05 Data 17

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I am grateful to Steve Horton for his assistance in creating the mathematica simulations used in this research. Joshua Lospinosa's help with the simulations and t-tests was invaluable to the project. I also thank John Graham and Craig Scherieber for providing both Ft. Leavenworth datasets. Additionally, I would like to thank Todd Retchless for his assistance.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

This research analyzes the underlying distribution of the network measure of density and analyzes the linear dependence within the adjacency matrix of two real world communication networks. The empirical data sets are from warfighting simulations at FT Leavenworth, KS, in 2005 and 2007.

To preform this analysis the Network Probability Matrix (NPM) is used to approximate the underlying network probably structure that determines which agents communicate within a network. This NPM is used in Monte Carlo simulation to simulate 100,000 instances of each network. Using these simulated instances we have shown empirically and through the central limit theorem that the density of a communication network follows a normal distribution.

Using hamming average hamming distances the NPM is further validated and is shown to provide a more reliable and consistent approximation of the underlying network than the empirical data sets provide from time step to time step.

To analyze the linear dependence of the adjacency matrices the ranks and determinants were explored. It was determined that the ranks follow an underlying distribution and have a negative correlation with the density of the network.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Presently many structure based frameworks are used in the network science community for the simulation of networks. These frameworks are based on the presence of triads, dyads, cliques and other network structural components. However theses frameworks do not always consider all of the factors that contribute to the dyadic relationship between agents. In a network an agent may not be interested in the occurrence of a triad between 3 other agents or that certain agents in the network have dyadic ties. The agent is mainly concerned with its own dyadic relationships leading to an underlying dynamic equilibrium in the network., i.e. the underlying edge probability structure that contains a probability that each agent will communicate with every other agent in the network. For an example of a dynamic equilibrium, take a network of four agents. Within this network agent 1 has a 0.5 probability of communicating with agent 2, a 0.75 probability of communicating with agent 3, and a 0.0 probability of communicating with agent 4. These probabilities result from agent 1's relationship with each other agent in the network and remain constant regardless of who agent 1 is communicating with at any given time.

The underlying probability structure of a network remains independent of observations at any instance in time and is constant in the network. A single observation of a tie does not necessarily designate a relationship between two agents, since the communication could have been made in error. On the other hand a single observation of the lack of a tie does not designate that a relationship does not exist, since agents are not continuously communicating with every agent they have a relationship with at every instance in time. While a snapshot of the network at an instance in time does not indicate the dyadic relationships between agents, this snapshot is based on the underlying network probability that each agent will communicate with every other agent.

A new framework is proposed for the simulation of networks that is based on the underlying probability structure of the dynamic equilibrium. This framework is the network probability matrix (NPM) proposed by McCulloh and Lospinoso (2007)[1]. The network probability matrix estimates the edge probabilities for each dyadic pairs in the network. Probability estimation can vary from a proportion of communications in a series of observations or be estimated from more complex distributions depending on the amount and type of data present. This framework can be used to simulate a network regardless of its topology: random, small-world, scale free, cellular, etc.

The edge probability structure of the underlying dynamic equilibrium remains constant in the network while the network is at a stable state. However, the underlying probabilities may change as shocks to the network take place. A shock to the system can be caused by a variety of occurrences, such as two agents being assigned to work together on a project, the temporary absence of an agent from a network, or a change in leadership within an organization.

Once a shock takes place the underlying probabilities of communication within the network will then stabilize as the network returns to it dynamic equilibrium. Using Monte Carlo simulation with the NPM the underlying distributions of network measures can be determined while the network is in its dynamic equilibrium. These underlying distributions can be used in change detection and allow us to statically predict shocks to the network and determine when significant changes occur.

A. BACKGROUND

Social network analysis is a theoretical framework that examines the relationships between social entities (e.g. people, groups, organizations, beliefs, knowledge, etc.). These objects are known as nodes and their connections are referred to as edges. Not all nodes are connected; however, some nodes are connected with multiple relationships. This network framework is applicable in a plethora of content areas such as communications, information flow, and group or organizational affiliation [2]. Social network analysis relies heavily on graph theory to make predictions about network structure.

1. Erdös-Rnia Random graphs

In 1959 mathematicians Paul Erdös and Alfrd Rnia made revolutionary discoveries in the evolution of random graphs. In their eight papers Erdös and Rnia evaluate the properties of random graphs with n vertices and m edges. For a random graph G containing no edges, at each time step a randomly chosen edge among the $\binom{n}{2}$ possible edges is added to G. This graph contains N edges and each edge of the $\binom{\binom{n}{2}}{n} = C_{n,N}$ possible edges are equiprobable. Therefore, once an edge is chosen from the $\binom{n}{2}$ equiprobable edges the next edge is chosen among the remaining $\binom{n}{2} - 1$ edges and this process is continued so that if k edges are fixed, all remaining edges have equal probabilities of being chosen [3].

A general model used to generate random graphs is as follows: "For a given p, $0 \leq p \leq 1$, each potential edge of G is chosen with probability p, independent of other edges. Such a random graph is denoted by $G_{n,p}$ where each edge is determined by flipping a coin, which has probability p of coming up heads [4]." In this model of random graphs each edge 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[5].

2. Statistical tests

Utilizing Erdös' theory of random graphs as well as the class of uniform distributions associated with these graphs, Holland and Leinheart [6] 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 an edge 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 [7]."

3. Strength of weak ties

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 circle 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. 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 into the outside world [8].

4. Small World Networks

Building off of Granovetter's model Duncan Watts and Steven Strogatz [9] 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 nodes first order connections are only connected through that node.

The clustering coefficient of the Watts-Strogatz small world network model reconciles clustering with the characteristics of random graphs. 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 has a high average path length connecting two random nodes. 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[10]. 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.

5. Scale Free Networks

According to Albert-Lszl Barabsi the random graph theory of Erdös and Rnia was rarely found in the real world. Barabsi has found that many real world networks have some nodes that are connected to many nodes and others that are connected to few nodes. His 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 [11]. 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.

Barabsi's model of scale-free networks is constructed around two ideas–growth and preferential attachment. For each time step a new node is added to the network. This illustrates the principal that networks are assembled one node at a time[5]. Assuming that each new node connects to the existing nodes of the network with two links, the probability

that the new node will choose a given node is proportional to the number of links the chosen node has. Therefore, a node with more links has a higher probability of being connected to. This creates a ”rich get richer” scenario where nodes with many links continue to grow by collecting new links while newer nodes with lower degrees do not collect as many links[5].

Based on a scale-free network model where nodes make connections based completely on preferential attachment the probability that a new node will connect to a node with k links is given by $\frac{k}{\sum_i k_i}$ [11]. This causes the first nodes in the network to develop into hub nodes due to having the longest time to collect links. However it is not always the case that the first nodes in a network develop into the biggest hubs.

6. Fitness Model

In order to account for newer nodes overtaking older nodes as hubs, Barabsi constructed the fitness model. Fitness is a node’s ability to collect links relative to every other node in the network and is based on competition in complex systems [12]. In this new model a node’s attractiveness is not determined completely by its number of links, but preferential attachment is driven by the product of the number of links a node has and its fitness. In this model the probability a new node will connect to a node with k links and a fitness of η is $\frac{k\eta}{\sum_i k_i \eta_i}$ [12]. Nodes in this model acquire links following the power law distribution of the scale-free model, however, the dynamic exponent, β, which determines how fast a node acquires new links, is different for each node. This is proportional to a node’s fitness, therefore, a node that is twice as fit as another node will obtain nodes twice as fast because its dynamic exponent is twice as large. This ”fit-get-rich” model allows nodes to become hubs based on their attractiveness regardless of when they enter the network[12].

7. Winner Take All Model

Contrary to the scale-free network model Barabsi developed the ”winner take all model,” which strongly portrays monopolies. The ”winner-take-all-model” consists of a single hub and many tiny nodes. This network develops a star topology and nodes do not acquire links following a power law distribution.

8. Network Probability Model

Ian McCulloh and Joshua Lospinoso (2007)[1] proposed a new framework for random communication networks over time, based on empirical data collected on real world networks. This new framework, estimates distributions for the time between communication messages, then based on a given time interval the probability of an edge occurring in the network is calculated for every ordered pair of nodes. These probabilities can be constructed through multiple techniques. To derive the probabilities from empirical data collected over several time periods, a proportion of edge occurrences, e_{ij}, can be used to estimate probabilities for each cell in the adjacency matrix a_{ij}. These probabilities are displayed in a network probability matrix where each cell is the probability that node i communicates with node j. This frame work is capable of generating networks that are similar to scale free networks. Thus, this model can be used to construct any network topology: Erdös-Rnia random, Watts-Strogatz small world, Albert-Barabsi scale-free, star, cellular, ect. The McCulloh-Lospinoso model is estimated from empirical data and can be used to simulate realistic observations of relationships in specific organizations.

9. Distribution of Network Level Measures

Using the McCulloh-Lospinoso NPM framework Daniel Baller, Joshua Lospinoso, Ian McCulloh and Anthony Johnson (2008)[13] established a method to estimate the distribution of specific communication network measures. This method uses the NPM to simulate a communication network from empirical data. Based on these simulations the underlying distribution for density of a communication network of 68 field grade army officers has been determined to be normally distributed.

B. DATA

This research evaluates the density of two real world networks in order to find the underlying distribution of network density. The first data set was collected from a war fighting simulation in FT Leavenworth, KS in April 2007 by Craig Schreiber and Lieutenant Colonel John Graham. There were 99 participants in the experiment that were monitored over the course of four days while data was being collected. This 99 agent data set was then cut down to 68 agents. These 68 participants served as staff members in the headquarters of the brigade conducting the exercise. The data displays the interactions of agents in a

network which was collected by monitoring communications throughout the simulation.

The second data set is from a war fighting simulation in FT Leavenworth, KS in 2005, also collected by Craig Schreiber and Lieutenant Colonel John Graham. This data set contains 156 agents that were monitored over the course of nine iterations of the simulation. This data exhibits the communication agents in the network that was collected by monitoring communications throughout the simulation. For the duration of this paper the Ft. Leavenworth 2007 Data will be referred to as Net07 and the Ft. Leavenworth 2005 data sets will be referred to as Net05.

THIS PAGE INTENTIONALLY LEFT BLANK

II. METHODS

This research explores the distribution of the density measure in two simulated networks using the network probability matrix.

Below is an outline of the approach pursued in this study the following methodology was used for each data set:

1. Construction of The NPM

In order to simulate the network it is necessary for a network probability matrix (NPM), to be created. Once the datasets for Net07 were trimmed of the scripted agents, they were symmetrized across the main diagonal in the Organizational Risk Analyzer (ORA) to account for the lack of directionality of communication in the data. Symmetrizing the data also corrects for the informant error of agents not reporting other agents they have communicated with. Next, the datasets from Ft. Levenworth 2007 were dichotomized to remove the weighting set by the participants. Once the data is dichotomized a one represents communication between two agents and a zero represents the lack of communication between two agents. To construct the NPM all eight data sets were compiled into a single data set consisting of the total number of discrete time periods that each agent communicated with each other agent. This matrix was then divided by the number of discrete time periods to determine the underlying edge probabilities for the network in dynamic equilibrium.

The Net05 data sets were collected as unweighed data so they did not have to be dichotomized. It was also unnecessary to trim these data sets. The nine data sets from this network were symmetrized across the main diagonal in ORA to correct for informant of agents not reporting other agents they have communicated with. To construct the NPM all nine data sets were compiled into a single data set consisting of the total number of discrete time periods that each agent communicated with each other agent. This matrix was then divided by the number of discrete time periods to determine the underlying edge probabilities for the network in dynamic equilibrium.

2. Simulation Generation

The NPMs were then used as the edge probabilities for Monte Carlo simulations of these two networks. In these simulations a random number was generated for each

edge. If the random number is less than the edge probability then the edge is added to the graph. This algorithm was used to create 100,000 instances of the network. Once 100,000 instances of the network were completed the average density was taken from each simulation to create a dataset of 100,000 network densities for each network.

3. Reliability and Consistency

To analyze the reliability and consistency of our simulations hamming distances were utilized as a metric for the differences between two binary adjacency matrices. Using the NPM, 60,000 instances of each network were simulated. The average hamming distance from each empirical data set to every other empirical data set and from each simulated network to each empirical data set were then calculated. These average hamming distances were then analyzed using a t-test. The results of this test indicate whether the NPM predicts an instance of the empirical network with more or less error than the error introduced over time.

4. Distribution Fitting

The normal distribution was fit to the data of each network using Maximum Likelihood Estimation. An Anderson-Darling goodness of fit test and a comparison of the estimated cumulative distribution function to the data's empirical distribution function indicated a very good fit for the data. In addition, since the density is a linear function of the average node degree, the central limit theorem would suggest that the density is normally distributed for each network, given certain assumptions.

III. LINEAR DEPENDENCE OF THE AGENTS

Utilizing the NPMs created for the study of the density an analysis of linear dependence was conducted. To analyze the linear dependence of each agent the NPMs were used as the edge probabilities for Monte Carlo simulations of Net07 and Net05. In these simulations a random number was generated for each edge. If the random number is less than the edge probability then the edge is added to the graph. This algorithm was used to create 100,000 instances of each network. Once the 100,000 instances of each network were created the ranks and determinants of each network were calculated to create datasets of 100,000 ranks and determinants for each network.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS

A. DISTRIBUTION OF THE DENSITY

Using the t-test, it is illustrated that the networks simulated using the NPM have a smaller average hamming distance to the empirical data sets than each empirical data set is to each other for both the Net07 network and the Net05 network. This is evidence that the simulated networks give a more reliable and consistent approximation of the underlying distribution. The results of the paired t-test for both networks are shown below in Table 1 and Table 2 respectively. In each table column one is the average hamming distance from each empirical data set to every other empirical data set and column three is the average hamming distance from 60,000 networks simulated with the NPM to each of the empirical data sets. The p-value of each test is approximately zero indicating that there is a statistically significant difference between the empirical hamming distances and the simulated hamming distances. While the t-test is not statistically significant for data sets 6, 8, and 9 of Net05, since the p-values are over 0.05, in each case the mean hamming distance to the simulated networks is less to the mean hamming distance to the empirical networks, showing that these cases all follow the trend of the other data sets. Additionally, since it is shown that the simulated networks have, on average, less Hamming distance from each of the empirical data sets than the empirical data sets have from each other.

This test shows that if you select one of the empirical adjacency matrices there is more error in predicting it from the remaining empirical data sets then from predicting it with the NPM.

Once the reliability and consistency of the simulations created using the NPM were confirmed, the distribution of the density could be determined. Since density is a linear function of a sample average of a network statistic according to the formula

$$density = \frac{avg\ degree}{(n-1)}$$

and the sample sized, n, is greater than 30 for each network the central limit theorem can be used to determine that the underlying distribution of network density is the normal distribution, with $\mu = 0.0984374$ and $\sigma = 0.00396148$ for Net07 and $\mu = 0.0476886$ and $\sigma = 0.000972361$ for Net05. This is also shown in Figure 1 and Figure 2 respectively.

Each graph shows the stepwise plot of the 100,000 densities overlaid with the CDF

Data Set	Mean Hamming Distance to Empirical Networks	Standard Deviation of Hamming Distance to Empirical Networks	Mean Hamming Distance to Simulated Networks	Standard Deviation of Hamming Distance to Simulated Networks	Paired 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	38.247	320.164	12.793	4.450	0.00
4	377.857	36.100	330.674	12.773	3.489	0.00
5	375.286	38.159	328.377	12.796	3.675	0.00
6	349.857	207.944	306.078	12.785	3.245	0.00
7	373.8571	48.451	327.0728	12.826	2.731	0.01
8	362.4286	55.635	317.1509	12.778	2.302	0.02

Table IV.1 t-test of Average Hamming Distances for Net07 Data

of the normal distribution. The sum of squared error of this model for Net07 is 9.60609 and the sum of squared error of this model for Net05 is 1.41659. These small sums of squared error reinforce the models shown above in Figure 1 and Figure 2.

A histogram of the densities for Net07 and Net05 are shown in Figure 3 and Figure 4 respectively.

It is shown in Figure three and Figure 4 that the densities of Network 1 and Network 2 both fit a normal curve. This further reinforces that the densities for both Network 1 and Normal 2 follow a normal distribution. Additional Normality test can be seen in Appendix

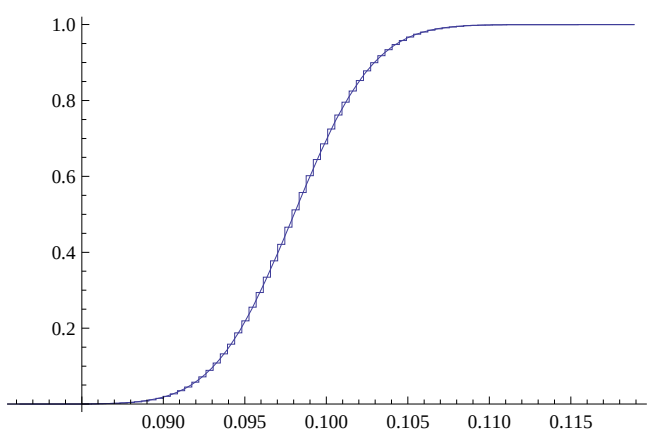

Figure 1. Stepwise Plot of Density Data for Net07 and CDF of the Normal Distribution

Data Set	Mean Hamming Distance to Empirical Networks	Standard Deviation of Hamming Distance to Empirical Networks	Mean Hamming Distance to Simulated Networks	Standard Deviation of Hamming Distance to Simulated Networks	Paired 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 IV.2 t-test of Average Hamming Distances for Net05 Data

A.

B. ANALYSIS OF LINEAR DEPENDENCE

Analyzing the determinants and ranks of the of Net07 data 17,650 of the simulated instances of the network had a nonzero determinant. In these instances each row of the adjacency matrix is linearly independent and the rank is 68 with no isolates in the network. The remaining 82,344 simulated instances have a determinant of zero and at least one row

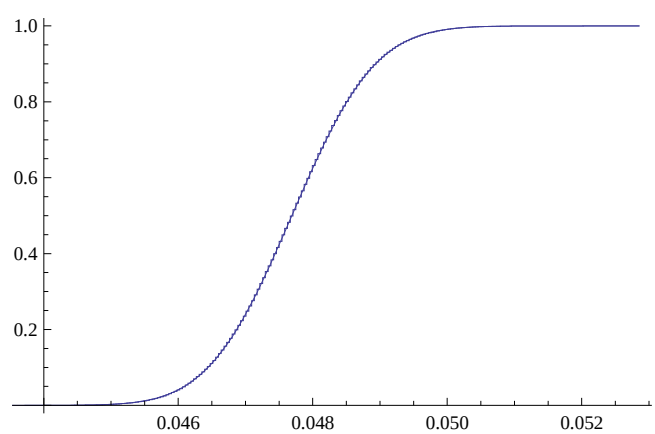

Figure 2. Stepwise Plot of Density Data for Net05 and CDF of the Normal Distribution

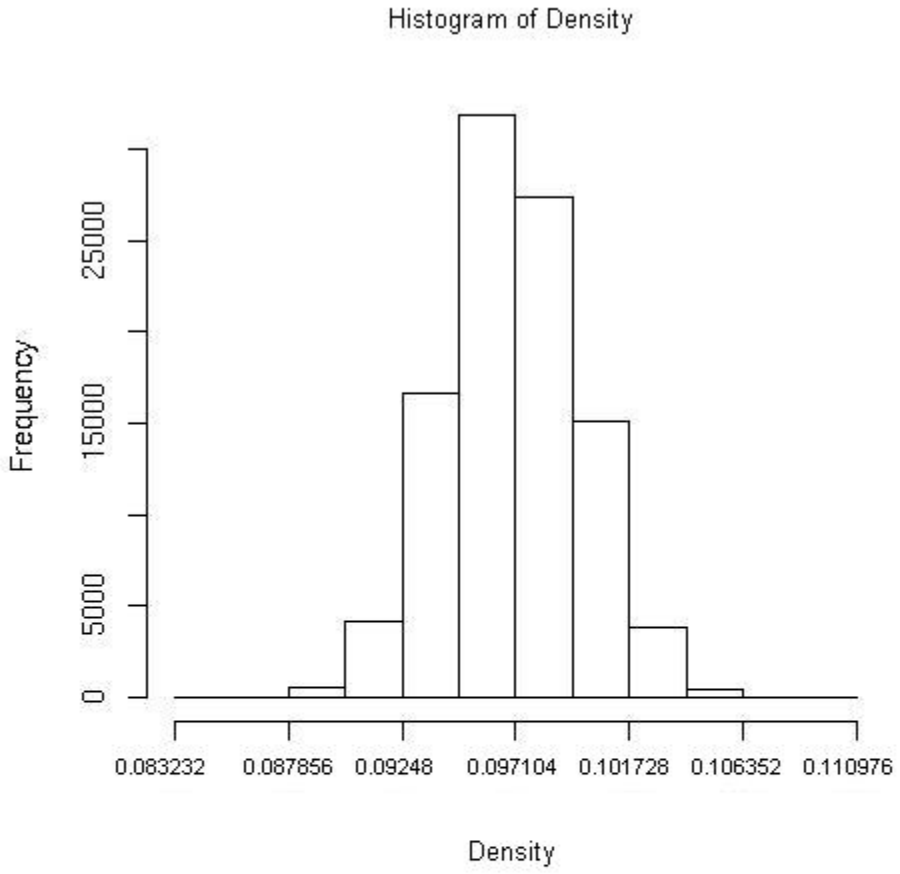

Figure 3. Histogram of Density for Net07

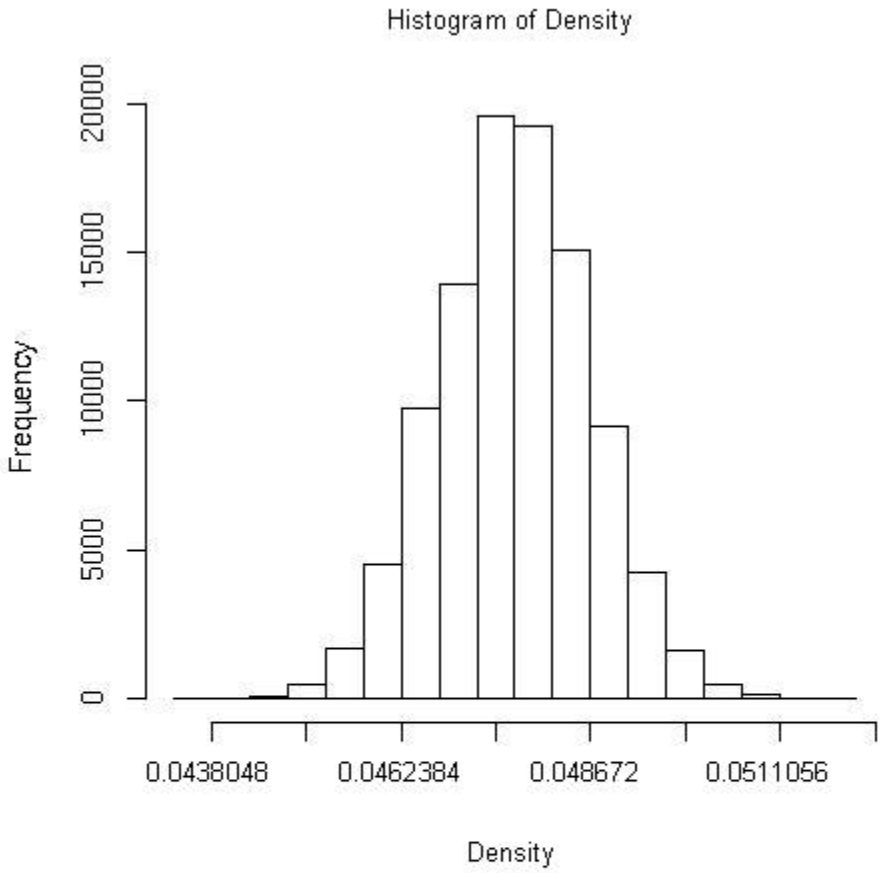

Figure 4. Histogram of Density for Net05

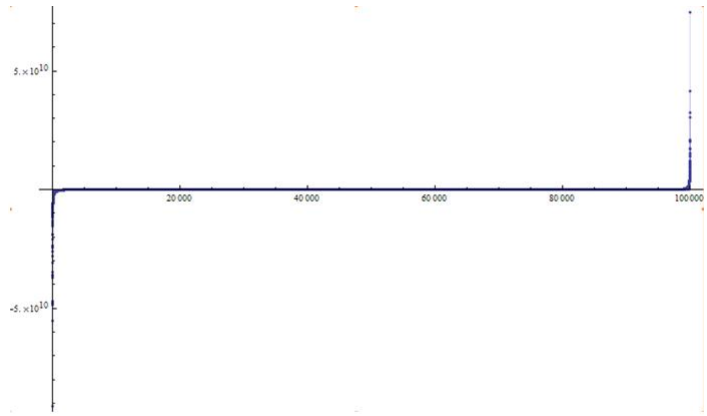

Figure 5. List Plot of Sorted Determinants for Net07

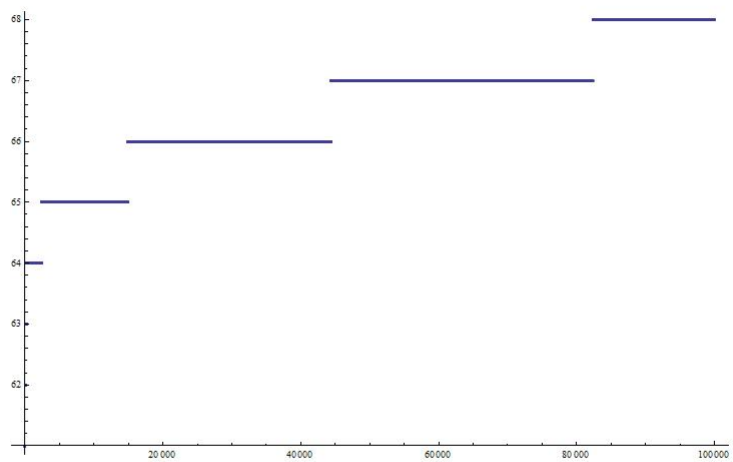

Figure 6. List Plot of Sorted Ranks for Net07

in the adjacency matrix that is linearly dependent. Additionally, these instances have a rank ranging from $n-1$ to $n-7$ where $n-rank$ is the number of isolates in the network. List plots of the sorted determinants and ranks for Net07 are shown in Figures 5 and 6 respectively.

Analyzing the determinants and ranks of the Net05 data all 100,000 of the simulated instances of the network had a determinant of zero and least one row in the adjacency matrix that is linearly dependent. Additionally, these instances have a rank ranging from n to $n-14$ where $n-rank$ is the number of isolates in the network. A list plot of the sorted ranks for Net05 is shown in Figures 7.

In this analysis it was noticed that the Net05 data having a density approximately half as large as the Net07 data has a maximum number of islets that is twice as large

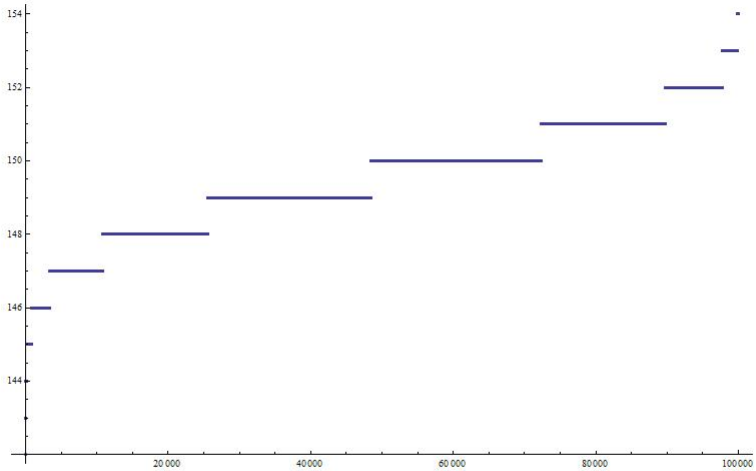

Figure 7. List Plot of Sorted Ranks for Net05

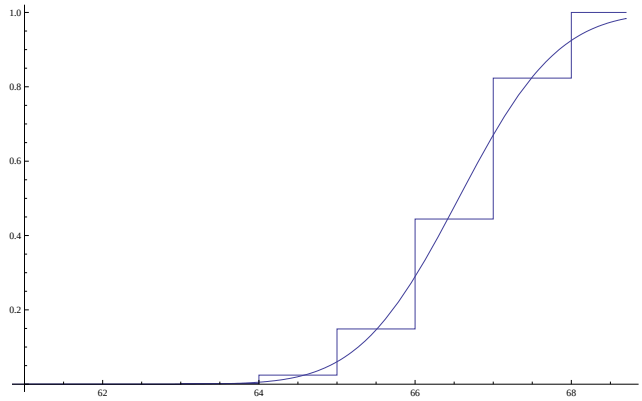

Figure 8. Stepwise Plot of Rank Data for Net07 and CDF of the Normal Distribution

indicating a negative correlation between the density and the number of isolates. The ranks for each network follow a normal distribution which further supports this finding. Figures 8 and 9 show the stepwise plots of Net07 and Net05 respectively shown against the CDF of the normal distribution.

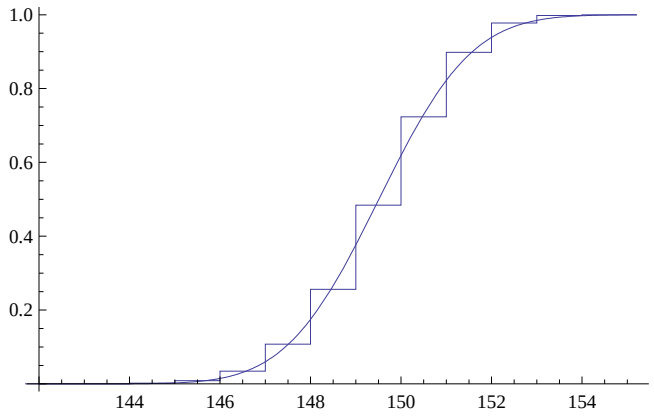

Figure 9. Stepwise Plot of Rank Data for Net05 and CDF of the Normal Distribution

V. CONCLUSION

This research validates the use of the NPM for simulating networks based on empirical data. based on the results of this research the NPM provides a reliable and consistent network simulation that is a strong framework for analysis.

It is also determined that $n - rank$ of the adjacency matrix will show the number of isolates within the network and that the maximum number of isolates in a network is dependent on the density.

This research can be extended in at least three aspects: assessing the underlying distribution for agent level statistical measures, assessing the underlying distribution for other network level statistical measures, and change detection. This allows changes and shocks in the network to be statistically predicted.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. ADDITIONAL STATISTICAL TESTS

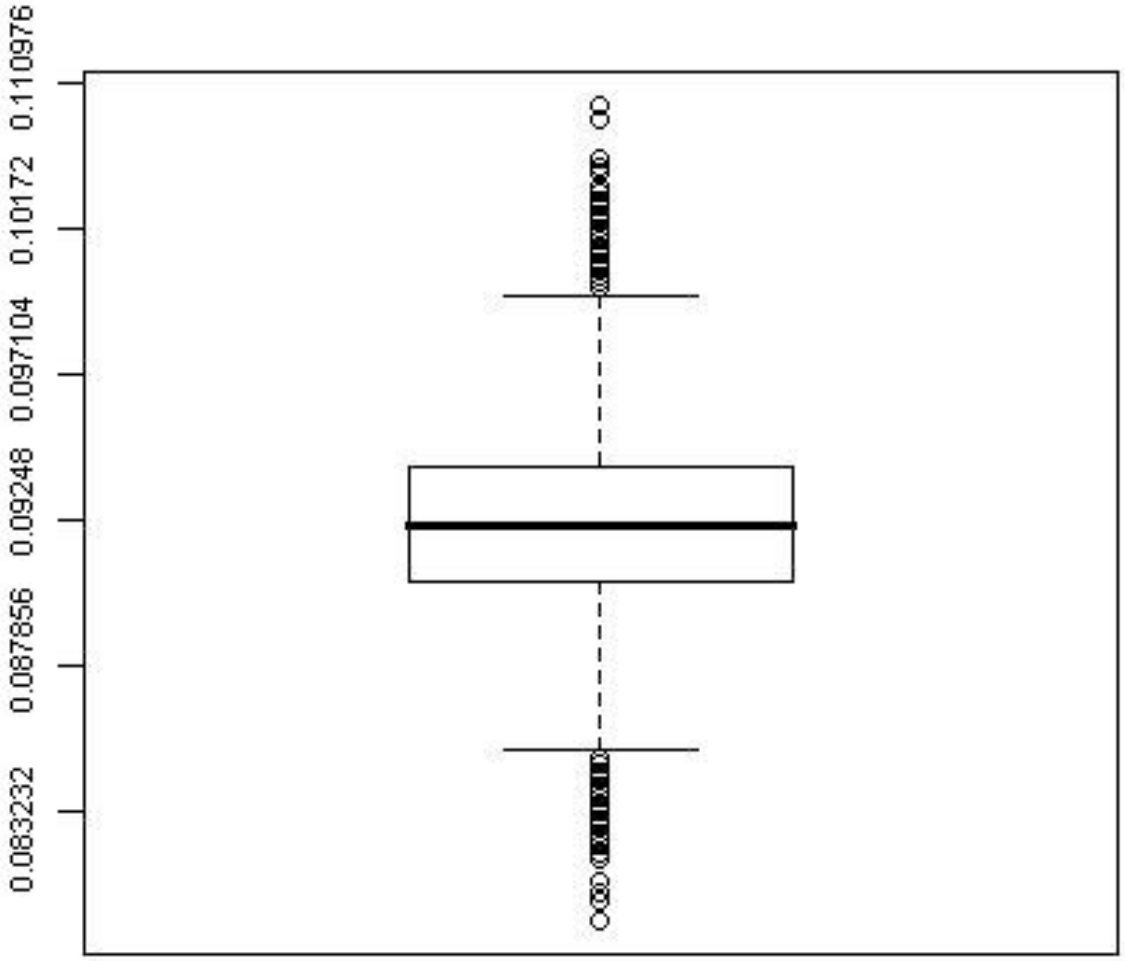

Figure 10. Box Plot of Densities for Net07

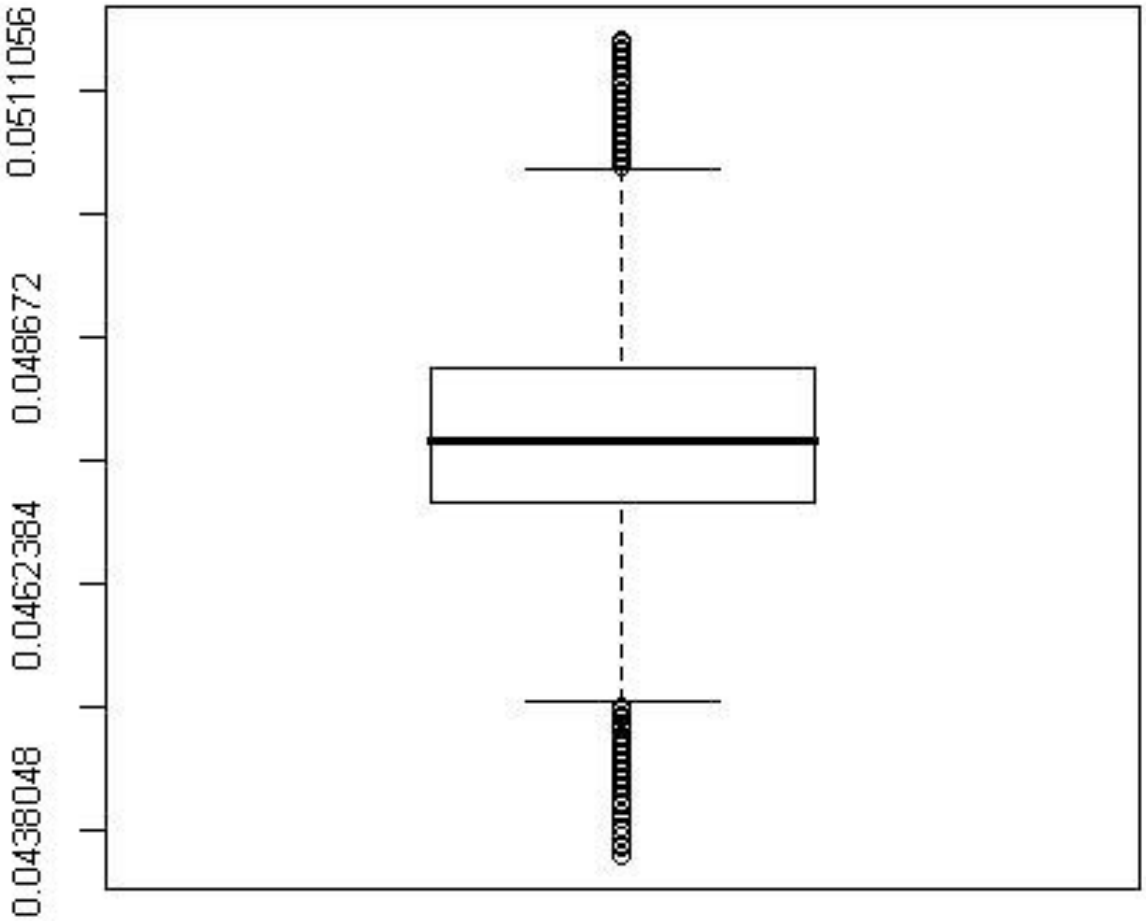

Figure 11. Box Plot of Densities for Net05

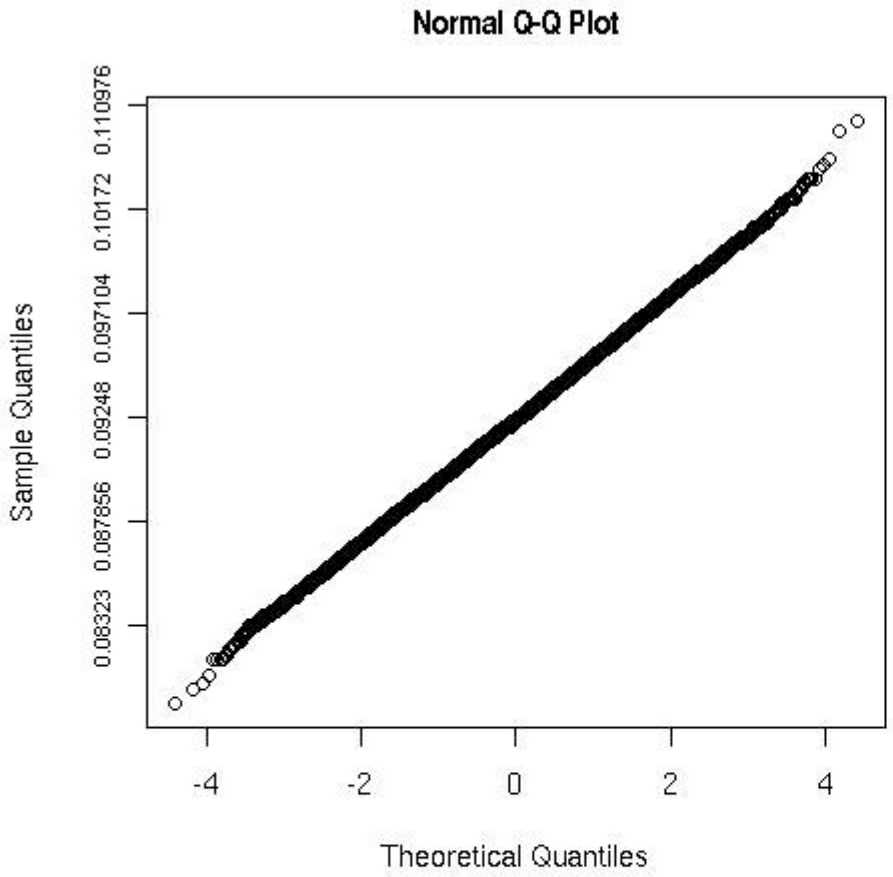

Figure 12. Normal Q-Q Plot of Densities for Net07

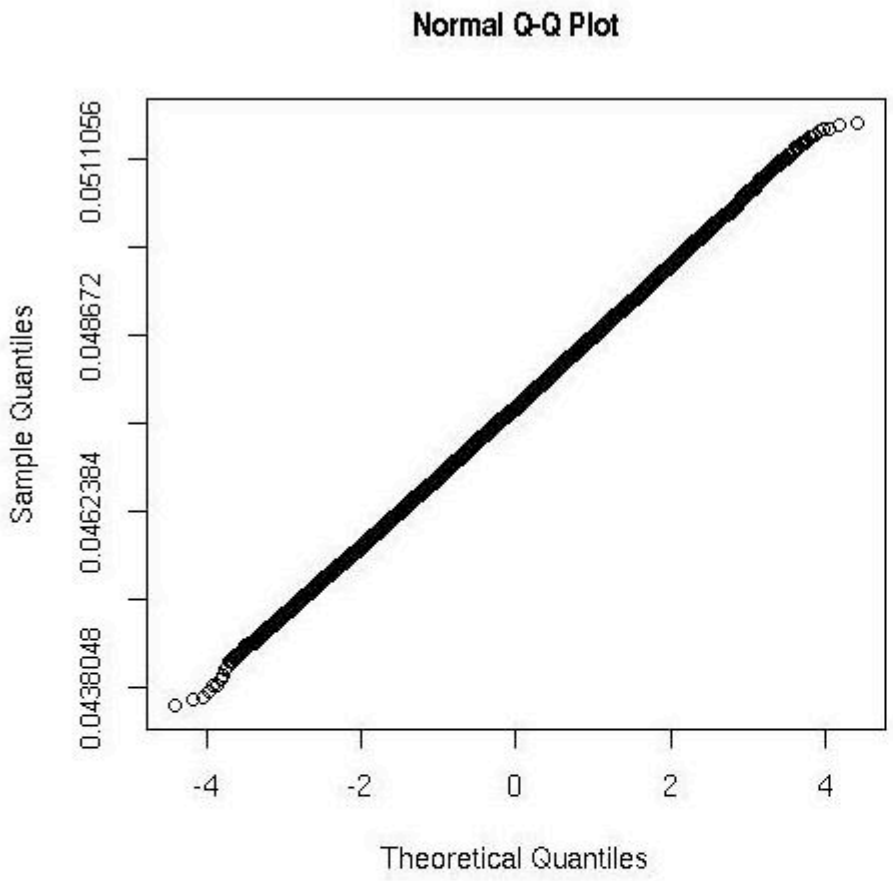

Figure 13. Normal Q-Q Plot of Densities for Net05

LIST OF REFERENCES

[1] Ian McCulloh and Joshua Lospinoso. Social networks probability mechanics. Applied Mathematics for Science and Engineering, Vol. 12:319–323, 2007.

[2] Noel M. Tichy, Michael L. Tushman, and Charles Fombrun. Social network analysis for organizations. The Academy of Management Review, Vol. 4(4):507–519, 1979.

[3] Paul Erdos and Alfred Rénia. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci., Vol. 5(370):17–61, 1960.

[4] Fan Chung and Ron Graham. Erdos on Graphs: His Legacy of Unsolved Problems. A K Peters, Wellesley, MA, 1998.

[5] A. Barabsi and Réka Albert. Emergence of scaling in random networks. Science, (286):509–512, 1999.

[6] P.W. Holland and S. Leinhardt. Transitivity in structural models of small groups. Comparative Group Studies, Vol. 2:107–124, 1971.

[7] Katherine Faust and Stanley Wasserman. Social Network Analysis: Methods and Applications. Cambbridge University Press, 1994.

[8] Mark S. Granovetter. The strength of weak ties. American Journal of Sociology, Vol. 78:1360–1380, 1973.

[9] J.D. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. Nature, Vol. 393:440–442, 1989.

[10] J.D. Watts and M. Newman. Renormalization group analysis of the small-world network model. Physics Letters, Vol. 263:341–346, 1999.

[11] A. Barabsi. Linked: How Everything is Connected to Everything Else and What it Means for Business, Science, and Everyday Life. Plume Books, New York, 2003.

[12] A. Barabsi and G. Bianconi. Competition and multisacaling inm evolving networks. Europhysics Letters, Vol. 54, 2001.

[13] Daniel Baller, Joshua Lospinoso, Ian McCulloh, and Anthony N. Johnson. Specific communication network measure distribution estimation. Network Science Submitted, 2008.