9-15-2016

A Statistical Approach to Characterize and Detect Degradation Within the Barabasi-Albert Network

Mohd-Fairul Mohd-Zaid

A STATISTICAL APPROACH TO CHARACTERIZE AND DETECT DEGRADATION WITHIN THE BARABÁSI-ALBERT NETWORK

DISSERTATION

Mohd Fairul Mohd-Zaid

AFIT-ENC-DS-16-S-003

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENC-DS-16-S-003

A STATISTICAL APPROACH TO CHARACTERIZE AND DETECT DEGRADATION WITHIN THE BARABÁSI-ALBERT NETWORK

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

Mohd Fairul Mohd-Zaid, BS, MS

September 2016

AFIT-ENC-DS-16-S-003

A STATISTICAL APPROACH TO CHARACTERIZE AND DETECT DEGRADATION WITHIN THE BARABÁSI-ALBERT NETWORK

Mohd Fairul Mohd-Zaid, BS, MS

Committee Membership:

Dr. Christine M. Schubert Kabban
Chair

Dr. Edward D. White
Member

Dr. Richard F. Deckro
Member

ADEDEJI B. BADIRU, PhD
Dean,
Graduate School of Engineering
and Management

Abstract

Social Network Analysis (SNA) is widely used by the intelligence community when analyzing the relationships between individuals within groups of interest. Hence, any tools that can be quantitatively shown to help improve the analyses are advantageous for the intelligence community. To date, there have been no methods developed to characterize a real world network as a Barabási-Albert network which is a type of network with properties contained in many real-world networks. In this research, two newly developed statistical tests using the degree distribution and the L-moments of the degree distribution are proposed with application to classifying networks and detecting degradation within a network. The feasibility of these tests is shown by using the degree distribution for network and sub-network characterization of a selected scale-free real world networks. Further, sensitivity to the level of network degradation, via edge or node deletion, is examined with recommendation made as to the detectable size of degradation achievable by the statistical tests. Finally, the degree distribution of simulated Barabási-Albert networks is investigated and results demonstrate that the theoretical distribution derived previously in the literature is not applicable to all network sizes. These results provide a foundation on which a statistically driven approach for network characterization can be built for network classification and monitoring.

To my daughter and my fiancée.

For showing me that you must first learn before you can understand, and only then can you truly love.

Acknowledgements

I would like to thank Dr. Schubert Kabban for guiding me through this challenging endeavor and for setting a high standard that I am proud to have met. Additionally, I would like to thank Dr. Deckro and Dr. White for their guidance and tutelage and for making sure that my ideas are presented in a well written manner.

I would also like to thank the 711^{th} Human Performance Wing, Battlespace Visualization Branch (711HPW/RHCV) for their financial support and for the mentorship provided by Dr. Blaha and Dr. Havig.

Lastly, I would like to thank a few of my classmates for the countless hours that they have helped me study for various rigorous courses, for lending their equipment to help me run my simulations, for helping brainstorm my research ideas, and for providing emotional support when I truly needed it.

Mohd Fairul Mohd-Zaid

Table of Contents

Page

List of Figures

List of Tables

Table Page

List of Abbreviations

Abbreviation		Page
SNA	Social Network Analysis	1
PCA	Principal Component Analysis	12
NP	non-polynomial time	14
TF	triad formation	16
PNDCG	prescribed node degree, connected graph	16
FSG	Frequent Subgraph Mining	17
SVM	Support Vector Machine	18
GMM	Gaussian Mixture Models	20
RBF	Radial Basis Function	20
GED	Graph Edit Distance	20
CDF	cumulative density function	24
PDF	probability density function	24
LS	least squares	27
SF	survivor function	27
MLE	maximum likelihood estimator	27
KS	Kolmogorov-Smirnov	28
PMF	probability mass function	29
MGF	moment generating function	31
PWM	probability weighted moments	33
PDQ	polynomial density-quantile	39
LRT	likelihood ratio test	45
UMP	Uniformly Most Powerful	48

Abbreviation		Page
UIT	Union-Intersection Test	50
iid	independent and identically distributed	73
MSE	mean squared error	110

A STATISTICAL APPROACH TO CHARACTERIZE AND DETECT DEGRADATION WITHIN THE BARABÁSI-ALBERT NETWORK

I. Introduction

Social Network Analysis (SNA) is heavily used by the intelligence community when analyzing the relationships between individuals or groups of interest [36]. Network information is often captured in the form of relational data that can be represented in the form of a graph. Thus, a main focus of SNA is to study both the relationships among entities (or nodes in graphical context), and the implications that these relationships may have on the collective and individual behavior resulting from the structure of the network. For the intelligence community, the level of analysis often involves network characterization, monitoring, and tracking; these are tasks which are normally performed visually, requiring some level of intuition by the intelligence analyst. Therefore, tools that can be quantitatively shown to improve SNA are advantageous for the intelligence community and others. For instance, a human study performed by Blaha and others [10] showed that the most appropriate technique for visualizing a network is dependent on various factors such as the network model and the task that is to be performed on the given network. Thus, if a network of interest can be characterized to the closest network model proxy, then the most appropriate visualization technique that gives the best insight into the structure of the network could be generated and the network could be studied in depth. For example, if a social network of interest is characterized as a Barabási-Albert network, then visualizing the network as the latter may provide more insights to the analyst in studying the given network. Additionally, for some applications, it is essential to monitor the

network once it can be characterized. Being able to detect network degradation with respect to its connectivity is necessary for monitoring the health of a network that might experience constant subtle perturbations. Using a computer network example, connections between servers may come down on occasions that are not due to an attack but rather due to operational factors. Therefore, having a test that can differentiate the two fairly quickly before any substantial damage is done to the network is very useful.

A property possessed by various real world networks, such as the World Wide Web, is the scale-free property [5]. This property was first described in networks by de Solla Price [90], and has been shown to govern many social, physical, and biological phenomena [6, 70, 16, 107]. The Barabási-Albert network exhibits the scale-free property and is highly desired for its simplicity in terms of its mechanics and parameters [24]. The Barabási-Albert model is driven by the concept of *preferential attachment* in which new entities are more likely to form connections with those that are already well connected within the network. Such preferential attachment can be attributed to some social networks where popularity drives the connection between entities. Despite both the scale-free property inherent in many real world networks and the attempts to statistically test for preferential attachment within a network [11], there have been no methods developed to characterize a real world network as a Barabási-Albert network [24]. Therefore, a method of classifying or characterizing a network as a Barabási-Albert network is novel because, although there is a link connecting the Barabási-Albert network to a scale-free network, the reverse connection has not been established. If such a connection can be made, it is possible to characterize a real world network as a Barabási-Albert network. One can then describe the real network's dynamics through the concept of preferential attachment. In addition, the real network can be easily simulated via the Barabási-

Albert network for characterization and monitoring.

Research Objectives

This research is driven by two primary objectives which are 1) to create a test of hypothesis in order to determine if a network or its sub-network can be represented as a Barabási-Albert network, and 2) to create a test of hypothesis in order to detect subtle degradation in a Barabási-Albert network attributable to nodal or edge deletion. For the first primary objective, properties of the degree distribution in a Barabási-Albert network are used to form a test of hypothesis. Specifically, the parameters associated with the *Pareto* distribution are utilized jointly in a test to classify network as Barabási-Albert. This approach is based on the knowledge that the Pareto distribution governs the degree distribution in the Barabási-Albert network. A test of hypothesis based on the two statistics of the degree distribution is then created and tested simultaneously after some inherent biases are corrected. For the second primary objective, L-moments from the empirical degree distribution of the Barabási-Albert network are used for creating a multivariate test of hypothesis. Analysis on the sensitivity of the multivariate test to changes within the network based on the proportion of edge and nodal deletion describes how quickly the test detects degradation in the network. Here, degradation refers to the changes in the structure of the network via its degree distribution and not the network performance which is the definition used in some research fields.

Finally, a secondary objective is also investigated which seeks to provide an accurate estimate of the parameters for the Pareto distribution and for which the hypothesis tests in the primary objectives are constructed. Such estimates have not been conclusively provided within the available literature, and for which the proper value is necessary in order to form the correct hypotheses for the tests developed. There-

fore, the parameters are estimated empirically through simulation of Barabási-Albert networks of various sizes.

The objectives in this research provide a foundation on which a statistically driven approach for network characterization can be built. Although the Barabási-Albert graph is the model of interest for this research, the statistical tests developed can be applied to other network models for network characterization and degradation detection.

The remainder of this dissertation is outlined as follows. A review on some graph theoretic concepts relating to graph degree, graph models, and graph similarity as well as statistical concepts on the degree distribution of the Barabási-Albert network, Power law and L-moments are first presented in Chapter II. Then, a novel method of characterizing a network as Barabási-Albert model through a Union-Intersection test of hypothesis using the Pareto distribution is presented in Chapter III. This is followed by a new method of detecting degradation within a Barabási-Albert network through a multivariate test of hypothesis based on the L-moments of the degree distribution presented in Chapter IV. Parameter estimation for the degree distribution of simulated Barabási-Albert networks are then presented in Chapter V in order to supplement the theoretical derivation available in the literature. Lastly, a summary of the results and a discussion on the impact of this research as well as possible areas where it could be expanded are presented in Chapter VI.

II. Background

2.1 Graphs and Networks Background

In this section, a literature review of some of the key areas pertaining to graphs and networks will be outlined. This includes a literature review on graph theory and SNA with a focus on network measures as well as network classification. If one can classify a real world network to a particular network algorithm with known parameters, then the knowledge gained from the algorithms can be used to describe the characteristics of the real world network. A human study conducted by Blaha and others [10] has shown that the most suitable network visualization techniques for various random networks (which will be defined later) are dependent on the particular network and the specific graph related question that is being asked. Therefore, suppose that an analyst is tasked with examining a real world network with a set of specific inquiries in mind. If said network can be classified as a specific Barabási-Albert network with defined parameters, then the analyst can use the best visualization method for that classification of network to visualize the real world network in order to give the best insights into the particular inquiries at hand. The analyst can then apply existing knowledge of the Barabási-Albert network to the real world network for analysis. The concept of creating a method of characterizing a real world network with a well defined network model is very useful.

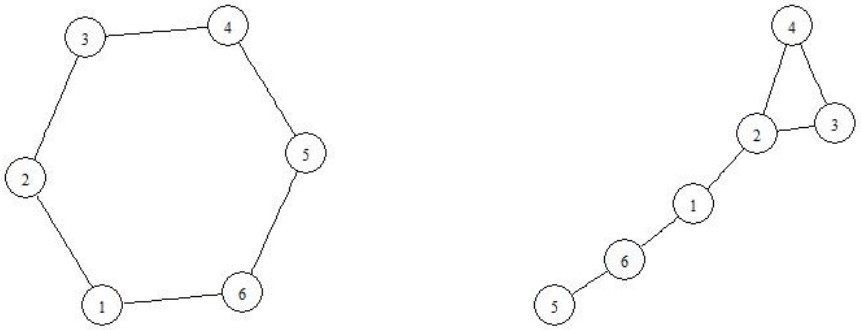
In order for such a task to be achieved, there is a need for a way to properly measure an empirical network so that it can be compared to the network model. A high fidelity measure can also prove useful in monitoring a network temporally and in keeping track of the evolution of said network. Wasserman and Faust [95] suggests that variability of nodal degree is one measure of both graph activity and centralization and further, that there is a relationship between density (the proportion

of all possible edges within a graph) and the mean number of nodes in a graph. In this section, graphs and network measures is defined and discussed. Note that a *graph* is defined to be the mathematical representation of a *network* although these two words will be used interchangeably.

2.1.1 Graph Theory Definitions.

The following notation is based upon network representation as graphs and is constructed from fundamentals in graph theory [97]. A graph G is defined as $G = (V, E)$ where $V = \{1, ..., N\}$ is a set of *nodes* or *vertices* of size N and $E \subset \{V \times V\}$ is a set of *links* or *edges*, $e_{i,j} = (i, j)$, connecting a pair of *nodes*. This research will concentrate only on *undirected simple graphs* which implies that $e_{i,j} = e_{j,i}$ and only one instance of $e_{i,j}$ is in E. A graph is called a *directed* graph if $e_{i,j} \neq e_{j,i}$ or a *multigraph* if there are more than one instance of $e_{i,j}$. Figure 1 gives an illustration of two undirected simple graphs, the Circle graph $G(V, E_1)$ and the Watts-Strogatz $G(V, E_2)$, both with a vertex set $V = \{1, 2, 3, 4, 5, 6\}$ and either edge set $E_1 = \{(1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 1)\}$ or edge set $E_2 = \{(1, 2), (1, 6), (2, 3), (2, 4), (3, 4), (5, 6)\}$, respectively. A *ring lattice* is defined as a graph with N nodes where each node is connected to K neighbors with $\frac{K}{2}$ neighbors on each side. The graph shown in Figure 2 is an example of a 4-regular ring lattice on 6 nodes. A graph can also be represented as a matrix, $\mathbf{A}$, called the adjacency matrix with elements

$$a_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise.} \end{cases}$$



$G(V, E_1)$ $G(V, E_2)$

Figure 1. Undirected simple graphs of size $n = 6$: left) Circle Graph $G(V, E_1)$ and right) Watts-Strogatz $G(V, E_2)$

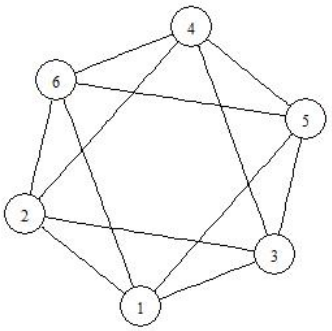


Figure 2. 4-regular lattice on 6 nodes

Shown below are the equivalent matrix representations of the previously defined graphs shown in Figure 1:

$$G(V, E_1) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} G(V, E_2) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Another way to represent a graph in matrix form is by using the incidence matrix, $\mathbf{I}(G)$, that lists which edges are incident with which nodes, with the nodes being indexed by the rows and the edges being index by the columns. Shown below are the incident matrices of the previous two graph examples as shown in Figure 1:

$$\mathbf{I}\left(G(V,E_1)\right) = \begin{pmatrix} & e_{1,2} & e_{2,3} & e_{3,4} & e_{4,5} & e_{5,6} & e_{6,1} \\ v_1 & 1 & 0 & 0 & 0 & 0 & 1 \\ v_2 & 1 & 1 & 0 & 0 & 0 & 0 \\ v_3 & 0 & 1 & 1 & 0 & 0 & 0 \\ v_4 & 0 & 0 & 1 & 1 & 0 & 0 \\ v_5 & 0 & 0 & 0 & 1 & 1 & 0 \\ v_6 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\mathbf{I}\left(G(V,E_2)\right) = \begin{pmatrix} & e_{1,2} & e_{1,6} & e_{2,3} & e_{2,4} & e_{3,4} & e_{5,6} \\ v_1 & 1 & 1 & 0 & 0 & 0 & 0 \\ v_2 & 1 & 0 & 1 & 1 & 0 & 0 \\ v_3 & 0 & 0 & 1 & 0 & 1 & 0 \\ v_4 & 0 & 0 & 0 & 1 & 1 & 0 \\ v_5 & 0 & 0 & 0 & 0 & 0 & 1 \\ v_6 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

2.1.2 Graph Measures.

Nodal *degree*, $d(n_i)$, is the number of direct edges connected from a given node to other nodes. Nodal *in-degree* is the number of edges that are directed from other nodes into a node. For directed graphs, the *in-degree* can be computed by summing the corresponding column of the adjacency matrix $\mathbf{A}$. Nodal *out-degree* is the number of edges that are directed from a node to other nodes. This can be computed by summing the corresponding row of the adjacency matrix $\mathbf{A}$ for a directed graph. The *in-degree* and *out-degree* are the same for undirected graphs. Table 1 lists the degrees

for the example graphs in Figure 1.

Table 1. Nodal degree for $G(V, E_1)$ and $G(V, E_2)$

Node	**Circle Graph** $G(V, E_1)$	**Watts-Strogatz** $G(V, E_2)$
1	2	2
2	2	3
3	2	2
4	2	2
5	2	1
6	2	2

A *nodal centrality index* is defined as the value assigned to a node to indicate its importance with respect to the whole graph. Define $C_A(n_i)$ as the nodal centrality index for node i denoted as n_i and $C_A(n^*)$ as the largest nodal centrality index of a network, then the general centralization index is

$$C_A = \frac{\sum_{i=1}^{N} (C_A(n^*) - C_A(n_i))}{\max \sum_{i=1}^{N} (C_A(n^*) - C_A(n_i))}$$

where $\max \sum_{i=1}^{n} (C_A(n^*) - C_A(n_i))$ is the maximum theoretical sum of degree difference for any network of size N [95, p.176]. The theoretical maximum is frequently incalculable and is often replaced using the variance of the difference, $Var(\mathbf{V})$, where $\mathbf{V} = \{v_1, \ldots, v_N\}$ and $v_i = C_A(n^*) - C_A(n_i)$ [95, p.177]. Group degree centralization applies degree centrality, or simply degree, to the general centralization formula which results in

$$C_D = \frac{\sum_{i=1}^{N} (C_D(n^*) - C_D(n_i))}{(N-1)(N-2)}.$$

A standardized degree centrality $C'_D(n_i) = \frac{d(n_i)}{N-1}$ measures the proportion of nodes

that are adjacent to node i. Degree variance

$$S_D^2 = \frac{\sum_{i=1}^{N} \left(C_D\left(n_i\right) - \bar{C}_D\right)^2}{N}$$

is also a measure of centralization and is dependent on N although a normalized variance is recommended by Snijders [89]. Wasserman and Faust state that, assuming that all edges are unweighted, the centrality of a node will decrease as it moves farther away from other nodes since there will be more edges linking that node to the other nodes [95, p.184].

Closeness is a measure that is related to the centrality measures, and it is defined as the inverse of the sum of pairwise distances between the nodes. Mathematically, closeness is expressed as

$$C_C\left(n_i\right) = \frac{1}{\left(\sum_{j=1}^{N} d\left(n_i, n_j\right)\right)}$$

and standardized closeness is given by

$$C'_C\left(n_i\right) = \left(N-1\right) C_C\left(n_i\right).$$

Group closeness as proposed by Freeman [25] is based on the standardized closeness given by

$$C_C = \frac{\sum_{i=1}^{N} \left(C'_C\left(n^*\right) - C'_C\left(n_i\right)\right)}{\left(N-2\right)\left(N-1\right)/\left(2N-3\right)} \tag{1}$$

where $C'_C\left(n^*\right)$ is the largest standardized closeness.

There are many properties of a graph that relate to closeness and centrality. Freeman showed that the maximum possible value for the numerator of Equation (1) is $\left(N-2\right)\left(N-1\right)/\left(2N-3\right)$. Group closeness and group degree centralization achieve unity when one node is connected to all other nodes. Bolland proposes a measure that

combines degree and closeness of nodes that is based on the number of paths that originate with each node [12]. Wasserman and Faust note that an "effective index [statistic] should reach its extremes in the cases of the circle graph (equal distances) and the star graph (one minimally distant node)" [95, p.187]. Circle graphs have a uniform distribution of centrality measures since all nodes have only degree two. In this case, none of the nodes are more central than the others which results in a group centrality of zero. On the other hand, star graphs have a single central node that is connected to the other nodes while those nodes are exclusively linked to the central node. This results in a very central graph having a group centrality measure of one.

The *clustering coefficient* for a given node measures the number of connections among the node's neighbors [96], and is related to the transitivity concept in the social network literature, where transitivity implies the idea of "a friend of a friend is a friend." It is defined as the proportion of local relationships among neighbors compared to the potential that all of the neighbors are connected. The mathematical formulation is defined as

$$C_{CL}\left(n_i\right) = \frac{|\{e_{i,j} \colon v_i, v_j \in K_{n_i}, e_{i,j} \in E\}|}{k_{n_i}(k_{n_i}-1)/2}$$

where $K_{n_i} = \{v_j \colon e_{i,j} \in E\}$ is the neighborhood of n_i and $k_{n_i} = |K_{n_i}|$ is the size of the neighborhood. The quantity in the denominator is the maximum number of edges possible in a graph of size k_{n_i}. The overall graph *clustering coefficient* can be computed by finding the mean of the nodal *clustering coefficients*. Another version of the *group clustering coefficient* is discussed by Wasserman and Faust [95] as a measure of triadic closure in a graph given by the ratio of existing triangular relationships, or triads, over the potential triads:

$$C_{CL} = \frac{6 \times (\#\text{ of triangles})}{\#\text{ of length two paths}}.$$

A length two path is simply an instance where one node is connected to two other nodes, but the two nodes are not connected to one another.

Diversity, as defined by Richards and Wormald [77], measures how uniformly connected the nodes are in a given graph. This measure is based on the square root of the ratio of edges that share no common end points, or *disjoint dipoles*, over the number of induced squares in the complete bipartite graph, $K_{\lfloor \frac{N}{2} \rfloor, \lceil \frac{N}{2} \rceil}$, and is given by

$$C_{DV} = \sqrt{\frac{\#\text{ of disjoint dipoles}}{\left(\frac{N}{4}\left(\frac{N}{2}-1\right)\right)^2}}.$$

For graphs that are not dense, the *diversity* measure is high. Therefore, this measure is somewhat similar to the *clustering coefficient* since a lowly clustered graph may indicate that there are few distinct communities, and that the graph is not diverse.

2.1.3 Other Network Measures.

An investigation on the correlation between twenty four network measures was conducted by Guzman and others [34], who categorized these measures into four groups using Principal Component Analysis (PCA). Their results demonstrate that group *degree centrality* is only highly correlated to one other measure, *pagerank*, and that together, these two measures form one of the four independent groups of correlated measures as reproduced in Table 2. Further, they show that *degree centrality* is significantly faster in terms of computation time when compared to pagerank. Therefore, not only is *degree centrality* shown to be correlated to the other more complicated measure, it was also more computationally efficient. However, the presence of associativity does not imply agreement, so even if some of the measures are correlated, they might not necessarily provide equivalent information. Additionally, a study by Mohd-Zaid and Schubert Kabban [64] demonstrates that some higher moments and

L-moments from the degree distribution of the Watts-Strogatz network is significantly more useful in classifying the network than just the mean degree. This suggests that using the characteristics of the whole degree distribution as opposed to a collection of multiple graph measures may provide more information in characterizing a network.

Table 2. Groups of highly correlated network measures as adapted from [34]

Group 1	**Group 2**
Clustering coefficient	Betweenness centrality
Soffer's clustering coefficient	Stress centrality
Squares clustering coefficient	Length-scaled betweenness
	Linearly-scaled betweenness
	k-betweenness
	Random walk betweenness
	Proximal source betweenness
Group 3	**Group 4**
Load centrality	Degree centrality
Proximal target betweenness	Pagerank
Uncorrelated Group	
Closeness centrality	Eigenvector centrality
Communicability centrality	Simple Diversity
General diversity	Communicability betweenness
Current flow betweenness	Approx. current flow betweenness
Closeness vitality	Average neighbor degree

As previously stated, nodal degree is a basic and simple graph measure, and there are other graph measures in the literature that attempt to characterize the complexity and entropy of a network. Mowshowitz and Dehmer present a taxonomy and overview of approaches to the measurement of graph complexity and probabilistic measures of graphs [67]. It was noted that there has also been considerable effort in applying various types of graph entropies in the field of network physics. Mowshowitz and Dehmer claim that there are two categories of probabilistic measures for graph complexity: intrinsic and extrinsic. Intrinsic measures use structural features of a graph to determine a probability distribution while extrinsic measures impose arbi-

trary probability distribution on graph elements [67]. For both intrinsic and extrinsic probabilistic measures, a numerical value is usually obtained by applying an entropy function to the distribution. For intrinsic measures, the measures include entropies on the symmetry of the graph, orbits, chromatic information, radial centric information, nodal degree, and weighted probablity schemes based on distances and degrees. For extrinsic measures, the measures include Körner entropy [50], parametric graph entropies, and non-parametric graph entropies. Körner entropy is not practical for measuring the complexity of large scale networks due to the graph coloring problem which is a non-polynomial time (NP)-hard problem. Parametric entropy measures rely on information functions to assign probablities to nodes of a graph whereas non-parametric measures are based on the eigenvalues of the characteristic polynomial of the graph. Although these measures exist, they will not be the focus of this dissertation as direct probabilistic measures computed on network characteristic rather than entropy based measures will be the focus in this research.

2.1.4 Graph Generating Algorithms.

The first commonly used random graph generating algorithm proposed by Erdös and Rényi [23] constructs a graph by connecting any pair of nodes via an edge with probability p, and by assuming each edge is independent from every other edge. This results in a graph of N nodes and m edges having an equal probability of $p^m(1-p)^{\binom{N}{2}-m}$ from all possible undirected simple graphs of N nodes and m edges. Despite being a truly random graph, a disadvantage of the Erdös-Rényi algorithm is that it is not scale-free [6] or small-world [96] which are properties that many real world social network such as the World Wide Web possess [5]. A scale-free network is defined as having a power law distribution for its nodal degrees. However, given its history, the Erdös-Rényi algorithm is widely used in the literature as a baseline

when making comparisons for network metrics and classifications.

A model based on two mechanisms that govern the scale-free power law distribution of real world networks are proposed by Barabási and Albert[6]. They define the two mechanisms to be: (i) networks expand continuously by the addition of new nodes, and (ii) new nodes attach preferentially to nodes that are already well connected. The model operates by first starting with an initial number of nodes N each having degree m. This is followed by an iterative process of adding one node with m edges where the edges are connected to an existing node i with degree k_i based on the preferential attachment probability

$$\pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2}$$

which is the probability that node i will be attached to the new node. Barabási and Albert state that

> [the capability that the model demonstrates in] reproducing the observed stationary scale-free distributions indicates that the development of large networks is governed by robust self-organizing phenomena that [is not specific to the domain be it social, biology, or the world wide web]. [6, p.509]

Watts and Strogatz propose a graph generator model that produces small-world properties [96]. Small-world networks possesses the property where the shortest path, L, between most pair of nodes in the networks are small and grows proportionately to the logarithm of the network size, N, such that $L \propto \log N$. The algorithm for this model functions by first starting with a ring lattice of size N. This is then followed by rewiring each edge in the lattice with probability β such that duplicates and self-loops are excluded. Many real world networks such as the neural network of the worm Caenorhabditis elegans, the power grid of the western United States, and the collaboration network of film actors are shown to possess small-world characteristic

of having small average shortest path [96]. However, one disadvantage of the Watts-Strogatz algorithm is that it produces a network that is not scale-free.

Therefore, an extension of the Barabási-Albert model was proposed by Holme and Kim [38] that induces clustering into the Barabási-Albert graph and, in doing so, combines both the scale-free and small-world properties into one single network model. The model works by the same iterative process as the Barabási-Albert model by connecting a new node, v, to a node, u, based on $\pi(k_u)$, but this is then followed by a triad formation (TF) step of adding an edge between node v to a neighbor of node u with probability p. An additional parameter, m_t, is added to the model which is the average number of TF trials per added vertex such that $m_t = (m-1)p$. The original Barabási-Albert model is obtained when $m_t = 0$. The additional TF step is added in order to model the previously mentioned transitivity phenomenon which has been observed in real world networks as shown by Newman [68].

Morris and others [66] created a prescribed node degree, connected graph (PNDCG) algorithm that allows the user to define the scale parameter as well as the clustering coefficient of the network. Comparisons of the average clustering coefficient with those from the Erdös-Rényi and Barabási-Albert generated networks show that their algorithm is able to generate networks with a wider distribution of average clustering coefficients.

Although the Barabási-Albert and the Watts-Strogatz algorithm are not fully representative of real world networks, they are the basis upon which other network models are built in terms of the properties that they possess, namely scale-free and small world. Therefore, it is important that both models first be investigated before all others. Mohd-Zaid and Schubert Kabban [64] have shown that the Watts-Strogatz network can be well characterized and classified with good accuracy using PCA on the moments and L-moments of some graph measures. Additionally, the model by Holme

and Kim [38] and Morris and others [66] are built upon the Barabási-Albert model and has the same underlying degree distribution which follows the power law, and thus far there are no method that is able to characterize a network as the Barabási-Albert. It is necessary that a well defined method first be established to characterize a network as Barabási-Albert before it can be extended to the other models. For these reasons, the focus of this research is aimed solely at the Barabási-Albert graph. Any extension to other types of networks is left for future works.

2.1.5 Graph Matching and Classification.

The research objective for this dissertation is to develop a test of hypothesis for comparing graphs with applications in network classification as well as monitoring of network degradation. Current state of the art methods for graph classification are now discussed which highlight gaps and help motivate the methodology for this research. Of particular interest is the concept of characterizing a given network based on its local structures. Graph classification is defined as any method of assigning a given graph to a library of graphs with known properties. Since the proposed method falls in the realm of graph classification, a survey on graph classification is presented to expose the reader to the various methods available in the literature.

A graph classification model using the maximum entropy extracted from local patterns was proposed by Moonesinghe and others [65]. This approach uses Frequent Subgraph Mining (FSG), which was intially proposed by Inokuchi and others [43], in order to generate frequent subgraph patterns that are then used to build the prediction model. FSG takes a graph and a minimum support threshold ϵ to generate all connected subgraphs that occur in at least $\epsilon\%$ of the graph. The subgraphs are then used to construct a binary feature vector which is then used to compute the maximum entropy of the graph in an iterative fashion until convergence in entropy occurs. The

approach has been compared to an AdaBoost and Support Vector Machine (SVM) classifier on the well known Chemical Compound [37], AIDS [103], Cancer [103], and Webspam datasets, and it performs comparatively with the other two methods, but not significantly better.

Jin and others [45] also proposed a graph classification method by creating classification rules based on pattern co-occurrence from the subgraphs. The method only performs pattern mining once by utilizing the Canonical Adjacency Matrix for pattern enumeration without repetition. As such, this method results in faster computation time than other methods that require multiple iterations. The method can be integrated into any subgraph mining algorithm by organizing patterns into groups of co-occurrence rules to form a rule set. Whenever a pattern is generated, the discrimination score of every rule is calculated with the pattern's inclusion and then the pattern is inserted into the rule that yields the greatest increase in discrimination score. The algorithm then finds a co-occurrence rule set that maximizes the number of graphs that can be classified correctly. The authors compared their method against LEAP[98]+SVM and gPLS (partial least squares) on the well known Protein [2] and PubChem chemical compound [1] datasets for truth classification. The results show that their technique performs comparably to the other techniques with magnitudes faster computation time.

Ketkar and others [46] presented an empirical comparison of the major approaches for graph classification, namely, SubdueCL which was first proposed by Gonzalez and others [32], FSG with SVM, a walk-based (direct product) kernel, FSG with AdaBoost, and DT-CLGBI which is a combination of FSG and decision trees as proposed by Nguyen and others [72]. SubdueCL is the pioneering algorithm for graph classification, and it operates by creating a decision list from subgraphs and performing an isomorphism test with a new graph for classification. FSG with SVM works by

using frequent subgraph mining to create a feature vector that is then used as inputs for SVM classification. The walk-based kernel is created by taking the direct product of two graphs as a similarity measure. FSG with AdaBoost works by using FSG to create a list of subgraphs and AdaBoost to create a list of positive and negative examples from the subgraphs that results in the upper bound of the gain that is associated with the supergraph. DT-CLGBI combines aspects of frequent subgraph mining and decision trees. The algorithms were compared using the Chemical Compound [37] dataset as well as artificial network data generated using an in-house data generating technique that assumes uniform distribution of vertex labels, edge labels, and degree for truth classification. The results show that the walk-based kernel performs poorly when the average degree is high and SubdueCL performs poorly when the graph is disconnected. Other methods perform similarly to one another.

Another technique for comparing graphs using subgraphs structures was introduced by Macindoe and Richards [61]. Their technique is performed by computing three summarizing features from the subgraphs, and then making a comparison using the Wasserstein metric [92] between the distributions of summarizing features to that of subgraphs from other graphs. The features used by the proposed method are the Leadership (also known as *Closeness Centrality*) [25], Bonding (also known as *Clustering Coefficient*) [95], and Diversity [77] measures as defined in Section 2.1.2. Macindoe and Owens then made comparisons between fifteen different network datasets of various types such as coauthorship, social, email, semantic, literature, economic, sports, neural, and citations. They used the proposed technique for network comparisons by performing clustering analysis on the set of measures from each graph. The results of their analysis suggest that graphs can be shown to be similar based on the full graph structure but dissimilar by their local structures.

Employing the statistical knowledge obtained from nodal attributes, Gibert and

others [28] proposed four fuzzy graph embedding methods that utilize known statistical techniques namely fuzzy k-means and Gaussian Mixture Models (GMM). These techniques were then applied to SVM using the linear as well as the Radial Basis Function (RBF) kernels for performing graph classification. The methods were compared against the k-Nearest Neighbor classifier as well as another Graph Edit Distance (GED) based embedding method that is applied to SVM on the Letter, Electronic Drawing, Digit, Fingerprint, and COIL databases [78] with the labels removed in order to illustrate the generalizability of their methods on unlabeled graphs. However, the results from their experiments show that their proposed methods performed no better than the two referenced methods with the exception of the result based on the COIL dataset. They claim that one advantage that their methods have over previously proposed methods is the computational efficiency provided by their embedding technique of transforming the graphs into vector forms.

By using already available network measures, Li and others [58] presented a graph classification technique that utilizes a feature vector of twenty graph measures which is then applied to SVM using the RBF kernel for classification. The approach was compared to other kernel based graph classification techniques such as Random Walk, Shortest Path, Cyclic Pattern, Subtree, and Graphlet and Subgraph kernels on the Chemical Compound [37], Protein [83], and Cancer datasets [19, 9, 8]. Their approach did not have an overall better accuracy although it was consistently faster in comparison to the kernel methods. A comparison between subgraph feature methods shows that subgraph features are not as good as the proposed approach of using the full graph measures. The authors also conducted a feature importance study for each dataset using SVM recursive feature elimination and found that the top features based on number of occurrence include: average clustering coefficient, number of nodes, number of eigenvalues, number of edges, energy (which is the squared sum of

the eigenvalues of the adjacency matrix), and average degree. A comparison with an augmented feature vector with ten additional features shows that the smaller feature set is sufficient to capture most of the important structural properties.

Ugander and others [91] proposed a coordinate system based on triadic structure within subgraphs of $k = \{3, 4\}$ for characterizing possible sub-networks within a social network. They use a Markov Chain to model the frequency space of triadic evolution for a size $k = \{3, 4\}$ subgraphs within a graph. Then, using graph homomorphism, they find inequalities governing subgraph frequencies which are then used as constraints for the linear program for finding the extremal bounds for the frequency space. They state that the bounds are not just properties of social graphs but are universal properties of all graphs. This is then used to identify regions that are theoretically inhabitable but not populated by the social graphs that they examine. They demonstrate their method on a Facebook® dataset by performing classification of sub-networks of various sizes into neighborhood, groups, and events through logistic regression by using the residuals from their method against the optimal parameter as features. This method was compared to the performance of using only global graph features such as size of k largest components, size of k-core (the maximal subgraph having degree k), number of components in k-core, number of composition in k-core, degeneracy, size of k-brace (subgraph formed by all edges of embeddedness less than k), and number of components in k-brace. The results show that their proposed method performed much better than the ones that uses only global measures.

Lagraa and others [53] proposed a new distance measure for comparing graphs using modular decomposition for obtaining prime graphs. This is then used to compare with other network's prime graphs using probe distance, which measures the number of edit operations needed to transform one graph into a second graph by label edits, and star distance [53]. Modular decomposition is first used to obtain prime graphs

which are graphs that have only trivial modules. They then used graph probing and star comparison to compute a distance between these graphs. They also prove that their distance is an upper bound of the star distance proposed by Zeng and others [105]. They note that the prime graph measure is only a pseudo-metric since it does not have a uniqueness property. The authors then used the AIDS, Protein, Chemical Compound, Electronic Drawing, Letter, Protein, and DNA/RNA datasets [78, 13] to perform comparisons and classifications. Comparisons were only performed on the AIDS and Proteins datasets. This was done by computing the pairwise distance of 10 graphs from each group. The distance and computation time were then compared against those obtained using regular GED and star distance. The results show that the prime distance is comparable to the star distance in terms of run time and does act as an upperbound for the star distance as claimed. The results from classification shows that prime distance is only comparable to star distance.

Despite all of the work presented in this subsection, as of this writing, there are no methods available in the literature that can classify a network as a particular network model such as the Barabási-Albert network. Instead, to date, methods have focused on comparing only specific features of particular graphs. As previously stated, the method of network characterization that is presented in this dissertation is useful for many reasons such as network visualization and network generation, to name a few. In addition, any knowledge that is available for the particular network model can now be applied to the network in question.

2.2 Statistical Background

In order to characterize a given network as Barabási-Albert network, various statistical tools that can utilize the properties of the network model are used. In this section, the definition of the power law distribution as it relates to the degree distribu-

tion of the Barabási-Albert graph will be described as will moments and L-moments which will be used to both derive the statistical tests to classify a network as a Barabási-Albert network and to detect degradation in a Barabási-Albert network.

2.2.1 Barabási-Albert Graph Degree Distribution.

The nodal degree of the Barabási-Albert scale-free graph can be derived by using the mean field theory as described by Barabási-Albert [6] and reproduced here. Let m_0 be the number of nodes initially included in the graph and m be the number of edges added at each iteration over t iterations, where

$$m \leq m_0. \tag{3}$$

The size of the graph at a particular iteration is then $n = m_0 + t$, the total number of edges $E = mt$, and consequently the total degree of the graph $\sum k_i = 2mt$. Given the preferential attachment in Equation (2), the rate of change for k_i over t iterations could be written as

$$\frac{\partial k_i}{\partial t} = m\pi\left(k_i\right) = \frac{k_i}{2t}. \tag{4}$$

Solving the differential equation for k_i we have

$$\begin{aligned} \frac{\partial k_i}{\partial t} &= \frac{k_i}{2t} \\ \frac{1}{k_i}\partial k_i &= \frac{1}{2t}\partial t \\ \ln k_i &= \frac{1}{2}\ln t + c \\ k_i &= t^{1/2}e^c. \end{aligned}$$

Since the initial degree of a newly added node is m, then the initial condition for the differential equation is $k_i\left(t_i\right) = m$ where t_i is the iteration at which node i was

created. Therefore, $m = t_i^{1/2} e^c$. Substituting in for e^c, the degree for node i is

$$k_i = \left(\frac{t}{t_i}\right)^{1/2} m. \tag{5}$$

Using Equation (5), the cumulative density function (CDF) for k_i may be expressed as

$$F_{k_i}(x) = P_{k_i}(k_i \leq x) = 1 - P_{t_i}\left(t_i \leq t\frac{m^2}{x^2}\right).$$

Since each node i is added at equal time intervals, then the probability density function (PDF) for the time at which node i is added, t_i, is $f(t_i) = \frac{1}{n}$, which results in k_i having the PDF

$$\begin{aligned} f_{k_i}(x) &= \frac{\partial}{\partial k} F_{k_i}(x) \\ &= \frac{\partial}{\partial k}\left[1 - \int_0^{t\frac{m^2}{x^2}} \frac{1}{n} dt_i\right] \\ &= \frac{\partial}{\partial x}\left[1 - \frac{tm^2}{nx^2}\right] \\ &= 2m^2 x^{-3}\frac{t}{n} \end{aligned}$$

where $t = n - m_0$ and $x \in [m, \infty)$.

Thus for finite n, the degree distribution is

$$f_{k_i}(x) = 2\left(m\sqrt{\frac{(n-m_0)}{n}}\right)^2 x^{-3}$$

which implies that the degree distribution of the Barabási-Albert graph follows a $Pareto\left(m\sqrt{\frac{(n-m_0)}{n}}, 2\right)$ distribution [7]. However, as the size of the graph increases, $n \to \infty$, the distribution of the degree converges to a $Pareto(m, 2)$ distribution.

Now consider a starting condition using a complete graph with $m_0 = m$ to guarantee a connected graph for any iteration. Here, the total degree at the end of any

given iteration is $\sum_j k_j = 2mt + m(m-1)$, thus Equation (4) becomes

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t + m - 1}$$

which results in

$$k_i = \left(\frac{2t + m - 1}{2t_i + m - 1}\right)^{1/2} m.$$

However, although each node i is added at equal time intervals, there are only $t + 1$ intervals since all nodes at time t_0 are created simultaneously, thus $f(t_i) = \frac{1}{t+1}$ resulting in k_i having the PDF

$$\begin{aligned} f_{k_i}(x) &= \frac{\partial}{\partial x}\left[1 - \frac{(2t + m - 1)m^2}{2(t+1)x^2} - \frac{(m-1)}{2(t+1)}\right] \\ &= 2\left(\sqrt{\frac{2t + m - 1}{2t + 2}}m\right)^2 x^{-3}, \end{aligned}$$

where $t = n - m_0$, and since $m_0 = m$,

$$f_{k_i}(x) = 2\left(\sqrt{\frac{2n - m - 1}{2n - 2m + 2}}m\right)^2 x^{-3}$$

which is a $Pareto\left(\sqrt{\frac{2n-m-1}{2n-2m+2}}m, 2\right)$ distribution that converges to $Pareto(m, 2)$ for $m = 3$ or as $n \to \infty$. Thus, the degree distribution of the Barabási-Albert graph converges to a $Pareto(m, 2)$ for very large networks with either an empty or connected starting condition.

The Pareto distribution has the form of the power law distribution, and was named after the early works by Pareto and Busino [73]. It also has a discrete form, the *Zipf's law* [108]. The continuous power law degree distribution, however, is only an approximation of the truly discrete degree distribution. The discrete version of the power law distribution, the Zipf's distribution, can be obtained using the power series

distribution. The form of the Zipf's and Yule-Simon distribution is demonstrated later in Section 2.2.3, both of which have been suggested as a better proxy for the degree distribution of the Barabási-Albert network [70].

2.2.2 Power Law Estimation.

The degree distribution of the Barabási-Albert graph is heavily skewed right [6, 4]. However, when the density is binned and plotted using a skewed logarithmic scale on the vertical and horizontal axes, the distribution follows an approximately straight line as shown in the degree distribution of a simulated Barabási-Albert graph in Figure 3. This line may be estimated by observing the relationship between the expression for

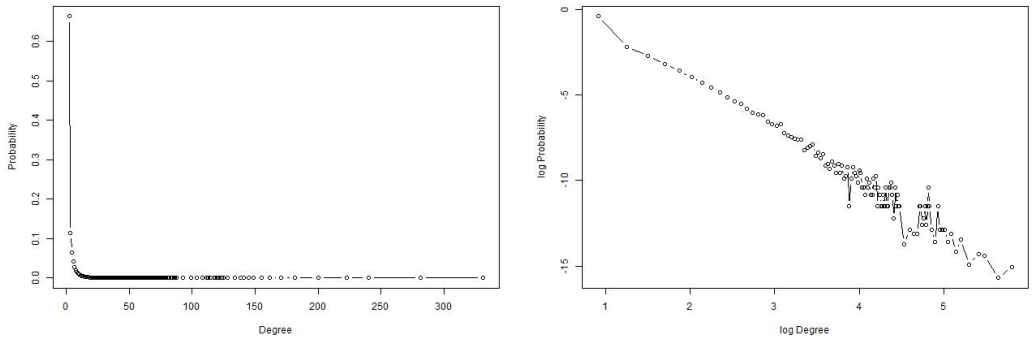


Figure 3. Plot of a Barabási-Albert $(n = 100000, m = 2)$ **degree distribution and its doubly log plot.**

the power law distribution and its log. A probability distribution is said to follow a *power law* if its PDF is of the form

$$f(x) = Cx^{-\beta}. \tag{6}$$

The log of this pdf is

$$\ln f(x) = \ln C - \beta \ln x$$

which is an expression for a straight line. The scalar β is the *exponent* of the power law whereas the constant, $\ln C$, is often set to satisfy the conditions of a probability distribution (i.e. Equation (6) integrates to unity). Once the linear form is estimated via data through least squares (LS), so is the power law distribution.

To provide a better approximation of the power law distribution, Newman [70] suggests the CDF of the data be plotted instead of the PDF. He argues that using the CDF bypasses the need for binning the data in order to create the PDF; a process which is more subjective. To maintain a similar downward trend, the complement of the CDF, which is often called the survivor function (SF), is plotted instead of the CDF itself. The complement of the CDF of the power law distribution is given by

$$S(x) = 1 - F(x) = C \int_x^\infty t^{-\beta} dt = \frac{C}{\beta - 1} x^{-\beta - 1}.$$

This SF also follows a power law, but with a shallower slope since the exponent is now $-(\beta + 1)$. Figure 4 demonstrates a plot of the SF for the data of the degree distribution in Figure 3. As can be seen, the straight line function in Figure 4 is more

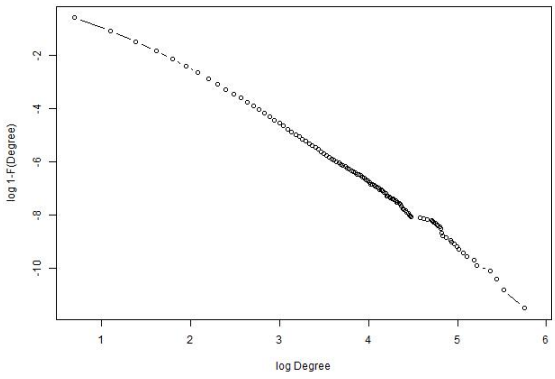


Figure 4. The survivor function of the Barabási-Albert ($n = 100000, m = 2$) **degree distribution.**

well behaved with less perturbation on the tail in comparison to its PDF in Figure 3.

It is easily shown that the maximum likelihood estimator (MLE) of β for the

$Pareto(m, \beta)$ distribution can be derived by considering the loglikelihood function

$$\mathcal{L}(\beta|\underset{\sim}{\mathbf{x}}, m) = n\beta \ln m + n \ln \beta - (\beta + 1) \sum \ln(x_i)$$

in which $x \in [m, \infty)$. Thus $\hat{m}_{MLE} = x_{(1)}$, and

$$\hat{\beta}_{MLE} = n \left[\sum \ln \frac{x_i}{x_{(1)}} \right]^{-1} = n \left[\ln \frac{\prod x_i}{x_{(1)}^n} \right]^{-1}.$$

However, a recently developed method of estimating the parameters of the Pareto distribution was proposed by Clauset and others [16] that is a combination of a MLE for β and nonparametric estimate for m. In their method, the estimate for β is the modified MLE

$$\hat{\beta}_{MLEnp} \simeq n \left[\sum \ln \frac{x_i}{x_{(1)} - \frac{1}{2}} \right]^{-1}.$$

For m, the value that minimizes the distance between the hypothesized degree distribution and the empirical degree distribution based on the Kolmogorov-Smirnov (KS) statistic [52] is chosen, denoted $\hat{m}_{MLEnp}$. Based on simulated data, they also suggested that overestimation of m is preferred to underestimation because the latter causes a larger deviation of $\hat{\beta}$ from the true β.

2.2.3 Discrete Power Law.

The power law as applied to the degree distribution assumes that the degree is continuous. However, a power law distribution can be derived assuming the degree is discrete. For the discrete case, let the probability $p(x)$ for an integer $x \geq 1$ be

$$p(x) = Cx^{-\beta}. \tag{7}$$

To assure that $p(x)$ is a probability distribution, a constant C is needed such that

$$1 = \sum_{x=1}^{\infty} p(x) = C \sum_{x=1}^{\infty} x^{-\beta} = C\zeta(\beta)$$

where $\zeta(\beta)$ is the Riemann ζ-function. Thus, the probability mass function (PMF) is

$$p(x) = \frac{x^{-\beta}}{\zeta(\beta)}$$

and if the power law has a lower bound m, then

$$p(x) = \frac{x^{-\beta}}{\zeta(\beta, m)} \tag{8}$$

where $\zeta(\beta, m) = \sum_{x=m}^{\infty} x^{-\beta}$ and $x \in \{m, m+1, \ldots\}$. The distribution in Equation (8) is referred to as the Zipf law after the work by Zipf [108].

Despite the straightforward development, Newman [70] suggests that the distribution derived from Equation (7) is not a good generalization of the power law for the discrete case. Another form of the power law for the discrete case is the Yule-Simon distribution [101, 87] which has been preferred over Zipf's law because the summation of the beta-function which makes up the distribution can be expressed in closed form unlike the Zipf's law that requires a form of the Riemann function. Yule [101] and Simon [87] proposed that the power law for the discrete case has the form

$$p(x) = C\frac{\Gamma(x)\Gamma(\beta)}{\Gamma(x+\beta)} = CB(x, \beta)$$

where $B(x, \beta)$ is the Legendre beta-function. Specifically, the normalizing constant

C for the Yule-Simon distribution is obtained by

$$1 = \sum_{x=1}^{\infty} p(x) = C \sum_{x=1}^{\infty} B(x, \beta) = C \frac{1}{\beta - 1},$$

thus the PMF of the Yule-Simon distribution is expressed as

$$p(x) = (\beta - 1) B(x, \beta)$$

for $x \in \{1, 2, \ldots\}$. However, given a condition where $x \geq m$, the truncated Yule-Simon distribution is given by

$$\begin{aligned} p_{X \geq m}(x) &= \frac{(\beta - 1) B(x, \beta)}{\sum_{x=m}^{\infty} (\beta - 1) B(x, \beta)} \\ &= \frac{B(x, \beta)}{\sum_{x=m}^{\infty} \int_0^1 t^{x-1} (1-t)^{\beta - 1} dt} \\ p_{X \geq m}(x) &= \frac{B(x, \beta)}{B(m, \beta - 1)} \end{aligned}$$

for $x \in \{m, m+1, \ldots\}$. Note that the truncated distribution for this application will be considered since the edge parameter m for the Barabási-Albert graph imposes a lower bound on the support for the degree distribution.

There are many forms of expressing the power law phenomenon depending on the nature of the data. However, the knowledge of having the power law within the degree distribution of the Barabási-Albert is useful since it allows us to test and act on the known distribution of degree for network of interest. Despite those variations, the continuous form is often used for network applications due to the relatively large size of networks in general.

2.2.4 Moments and L-Moments.

In probability and statistics, a moment is a quantitative measure that describes a characteristic of a random variable. An extensive set of moments may give a more descriptive summary for some random variables over the traditional approach of reporting the mean and variance since the latter only describes two moments of a distribution related to the center mass and scale of the distribution. It should be noted that while the existence of a moment generating function (MGF) for a random variable implies that there exists an infinite set of moments, the converse is not true. The characterization of a set of moments is not enough to uniquely define a random variable because there may exists another random variable having the same set of moments. However, uniqueness of moments is guaranteed if the random variables have bounded support or if the MGF exists in the neighborhood of zero [14]. The ℓ^{th} moment of a continuous PDF, $f(x)$, is defined as

$$\mu_\ell = E\left[X^\ell\right] = \int_{-\infty}^{\infty} x^\ell f(x)\, dx$$

and for a discrete PMF, $P(X = x)$, as

$$\mu_\ell = E\left[X^\ell\right] = \sum_{x \in X} x^\ell P(X = x)$$

where the mean μ is defined as the first moment of a distribution. The ℓ^{th} central moment is defined as

$$\acute{\mu}_\ell = E\left[\left(X - E\left[X\right]\right)^\ell\right].$$

It is often useful to scale the upper moments ($3^{rd}, 4^{th}$, and so forth) by a function of the variance so that comparison can be made between different distributions regardless

of the variance. An ℓ^{th} standardized moment is defined as

$$\gamma_\ell = \frac{E\left[(X - E[X])^\ell\right]}{\sigma^\ell} \tag{9}$$

where $\sigma^2 = \acute{\mu}_2$ defines the variance of the distribution. The 3^{rd} and 4^{th} standardized moments measure the symmetry and peakedness, respectively. However, not all moments exist for every distribution and no moments exist for some distributions, notably the Pareto distribution when $\beta \leq 1$. One set of metrics that solves the issue of nonexistent moments for some distributions is the L-moment. L-moments have a theoretical advantage of being able to characterize a wider range of distributions since the set of L-moments for a random variable exists if and only if the random variable has a finite mean [40]. However, this does not solve the problem for distributions where the mean does not exist such as the Pareto distribution when $\beta \leq 1$. Instead, other techniques can be used for those distributions such as the trimmed L-moments as defined by Elamir and Seheult [21]. Trimmed L-moments, however, are not the focus of this dissertation.

L-moments were first proposed by Hosking [40] as a conglomerate result that was derived from a collection of previous results [30, 84, 85, 20, 15, 49, 62, 33]. L-moments are linear combinations of order statistics that describe the location and shape of the probability distribution analogous to classical moments. The r^{th} L-moment is defined as

$$\lambda_r = \frac{1}{r}\sum_{i=0}^{r-1}(-1)^i \binom{r-1}{i} E[X_{r-i:r}] \tag{10}$$

where $X_{j:n}$ denotes the j^{th} order statistic (j^{th} smallest sample value) in an independent sample of size n. Note that $\lambda_1 = E[X_{1:1}] = E[X] = \mu$. The r^{th} L-moment ratio is defined as

$$\tau_r = \frac{\lambda_r}{\lambda_2}; r = 3, 4, ...$$

and is akin to the standardized conventional moment as defined in Equation (9) but has an open bound of $(-1, 1)$. The 1^{st} and 2^{nd} L-moments are referred to as L-mean and L-scale, respectively, whereas the 3^{rd} and 4^{th} L-moment ratios are referred to as L-skewness and L-kurtosis, respectively.

Hosking [40] states that a set of L-moments is unique to a particular distribution as long as the mean of the distribution exists. Further, Hosking stated that if the L-moments do exist, then the first two L-moments, λ_1 and λ_2, as well as the third and fourth L-moment ratios, τ_3 and τ_4, are enough to summarize the main features of a probability distribution. Additionally, the set of L-moments is considered more robust to outliers than conventional moments [40]. For example, a distribution with one very outlying point will cause the variance to increase quite notably but does not affect the L-scale to the same extent.

Estimating L-moments can be achieved by considering the probability weighted moments (PWM) which are used mainly to estimate the parameters of distributions that can be expressed in inverse form. Hosking [40] has shown that the first four L-moments can be expressed as

$$\lambda_1 = \xi_0 \tag{11}$$

$$\lambda_2 = 2\xi_1 - \xi_0 \tag{12}$$

$$\lambda_3 = 6\xi_2 - 6\xi_1 + \xi_0 \tag{13}$$

$$\lambda_4 = 20\xi_3 - 30\xi_2 + 12\xi_1 - \xi_0 \tag{14}$$

where ξ_r are PWMs defined by Greenwood and others [33] as

$$\xi_r = \int_0^1 x(G)G^r \, dx$$

such that G is a nonexceedance probability. Nonexceedance probability is the prob-

ability that an event X is smaller than or equal to the reference value X_r. Unbiased estimators for ξ_r were defined by Landwehr and others [54] as

$$\widehat{\xi}_r = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)(i-2)\cdots(i-r)}{(n-1)(n-2)\cdots(n-r)} x_{(i)}.$$

Estimates of the L-moments can be obtained by applying the estimated PWMs to Equations (11) to (14) resulting in the L-moment estimate, l_r, given by:

$$l_r = \binom{n}{r}^{-1} \sum_{1\le i_1<i_2<\ldots<i_r\le n} \frac{1}{r}\sum_{j=0}^{r-1}(-1)^j \binom{r-1}{j} x_{i_{r-j}:n},\ r = 1, 2, \ldots, n. \qquad (15)$$

Direct estimators of the first four L-moments were derived by Wang [94] that circumvent the need for using PWMs. These estimators are defined, respectively, as

$$\begin{aligned}
\widehat{\lambda}_1 &= \tbinom{n}{1}^{-1} \textstyle\sum_{i=1}^{n} x_{(i)} \\
\widehat{\lambda}_2 &= \tfrac{1}{2}\tbinom{n}{2}^{-1} \textstyle\sum_{i=1}^{n} \left(\tbinom{i-1}{1} - \tbinom{n-i}{1}\right) x_{(i)} \\
\widehat{\lambda}_3 &= \tfrac{1}{3}\tbinom{n}{3}^{-1} \textstyle\sum_{i=1}^{n} \left(\tbinom{i-1}{2} - 2\tbinom{i-1}{1}\tbinom{n-i}{1} + \tbinom{n-i}{2}\right) x_{(i)} \\
\widehat{\lambda}_4 &= \tfrac{1}{4}\tbinom{n}{4}^{-1} \textstyle\sum_{i=1}^{n} \left(\tbinom{i-1}{3} - 3\tbinom{i-1}{2}\tbinom{n-i}{1} + 3\tbinom{i-1}{1}\tbinom{n-i}{2} - \tbinom{n-i}{3}\right) x_{(i)}.
\end{aligned}$$

Recall that L-moments are unbounded, but L-moment ratios fall in the interval $(-1, 1)$. Additionally, the second L-moment is strictly positive, and the fourth L-moment ratio, L-kurtosis, is shown to have a tighter bound of $\frac{1}{4}(\tau_3^2 - 1) \le \tau_4 < 1$ [40].

Even though the definitions presented are for continuous random variables, L-moments can also be used with discrete distributions since Equation (10) remains valid and Equation (15) still provides unbiased estimators for λ_r; however, expressions for L-moments of common discrete distributions tend to be complicated [40]. Hosking also stated that since discrete random variable can be approximated by a continuous

random variable, certain results for continuous variable L-moments are also valid for discrete random variables [39].

An extension to multivariate L-moments analogous to comoments, coined L-comoment matrices, was proposed by Serfling and Xiao [82]. Comoments are similar to the variance-covariance matrix but instead of estimating the second central moment, they instead estimate the higher moments. Let $(X^{(1)}, X^{(2)})$ have CDF F with marginal distributions F_1 and F_2 and L-moment sets $\{\lambda_\ell^{(1)}\}$ and $\{\lambda_\ell^{(2)}\}$, respectively, then the ℓ^{th} *L-comoment* of $X^{(1)}$ with respect to $X^{(2)}$ is defined by Serfling and Xiao [82] as

$$\lambda_{\ell[12]} = \lambda_{\ell[21]} = Cov(X^{(1)}, P^*_{\ell-1}(F_2(X^{(2)})))$$

where

$$P^*_\ell(u) = \sum_{j=0}^{\ell} (-1)^{\ell-j} \binom{\ell}{j} \binom{\ell+j}{j} u^j$$

and $P^*_0(u) \equiv 1$. The orthogonal polynomials $P^*_\ell(u)$ for $0 \leq u \leq 1, \ell = 0, 1, 2, \ldots$, comprise the *shifted Legendre* system [82]. Now suppose that $\{(X_i^{(1)}, X_i^{(2)}), i = 1, \ldots, n\}$ is ordered by $X_i^{(2)}$, then the variate $X_i^{(1)}$ that is paired with $X_{r:n}^{(1)}$ is called the *concomitant* $X_{[r:n]}^{(12)}$. Serfling and Xiao [82] have shown that

$$\begin{aligned} E(X_{[r:n]}^{(12)}) &= nE\left(X_1^{(1)} | X_1^{(2)} = X_{[r:n]}^{(12)}\right) \\ &= n\binom{n-1}{r-1} E\left(X^{(1)}[F_2(X^{(2)})]^{r-1}[1 - F_2(X^{(2)})]^{n-r}\right) \end{aligned}$$

so the ℓ^{th} L-comoment of $X^{(1)}$ with respect to $X^{(2)}$ can be expressed as

$$\lambda_{\ell[12]} = \ell^{-1} \sum_{j=0}^{\ell-1} (-1)^j \binom{\ell-1}{j} E\left(X_{[\ell-j:\ell]}^{(12)}\right).$$

Given a random vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, then the ℓ^{th} L-comoment for each pair $(X^{(i)}, X^{(i)})$ for $1 \leq i, j \leq d$ is the $(i, j)^{th}$ entry of the ℓ^{th} *multivariate L-moment*

matrix, $\mathbf{\Lambda} = (\lambda_{\ell[ij]})_{d\times d}$. $\mathbf{\Lambda}_1$ is simply the vector mean, whereas $\mathbf{\Lambda}_2, \mathbf{\Lambda}_3$, and $\mathbf{\Lambda}_4$ are termed *L-covariance, L-coskewness, and L-cokurtosis*, respectively. Although the concept of L-comoments are presented here for completeness, they are not the focus in this research.

Distribution of Moments and L-Moments.

Since the estimates of the moments and L-moments are functions of random variables, then they are also random variables with underlying distributions. The distribution for the sample mean of certain distributions, such as the χ^2, Exponential, and Gamma distributions, can be easily derived using their MGF, whereas others can be obtained through transformation of variables [14]. With the exception of the Normal distribution and the sample moments for the distributions to be described in this subsection, no extensive work is available in the literature in characterizing the exact distributions of the sample mean, variance, skewness, and kurtosis. Table 3 lists the expected mean, variance, skewness, and kurtosis of the χ^2, Exponential, Gamma, and Normal distributions. The distribution of sample skewness and kurtosis of the Normal distribution is a simulated approximation where the values are compared to the percentile values of the proposed distributions and thus were shown to be the closest fit. The closed form function for the distribution of the sample variance for the Gamma distribution was derived by Royen [79].

The expected values of L-scale, L-skewness, and L-kurtosis for a noninclusive selection of distributions are listed in Table 4 as given by Hosking [39, 40]. As of this writing, no publication on characterizing the distributions of the L-moments have been found in the literature. However, Elamir and Seheult [22] have derived the expressions for the exact variances of the first four sample L-moments as well as their covariances.

Table 3. Distribution of various moments and central moments for four well known distributions

Distribution	Mean	Variance	Skewness	Kurtosis
Chi-square(p)	p $\hat{\mu}_1 \sim \Gamma(\frac{np}{2}, \frac{2}{n})$	$2p$	$2\sqrt{\frac{2}{r}}$	$\frac{12}{r} + 3$
Exponential(λ)	λ $\hat{\mu}_1 \sim \Gamma(n, \frac{\beta}{n})$	λ^2	2	9
Gamma(α, β)	$\alpha\beta$ $\hat{\mu}_1 \sim \Gamma(n\alpha, \frac{\beta}{n})$	$\alpha\beta^2$ $\hat{\sigma^2} \sim$ derived by Royen [79]	$\frac{2}{\sqrt{\alpha}}$	$\frac{6}{\alpha} + 3$
Normal(μ, σ^2)	μ $\hat{\mu}_1 \sim N(\mu, \frac{\sigma^2}{n})$	σ $\frac{(n-1)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-1}$	0 $\hat{\gamma}_3 \sim t$ [74, 75]	3 $\hat{\gamma}_4 \sim$ Pearson-IV[74, 75]

Table 4. L-scale, L-skewness and L-kurtosis of well known distributions as derived by Hosking [39]

Distribution	L-Scale	L-Skewness	L-Kurtosis
Exponential(λ)	$\frac{1}{2\lambda}$	$\frac{1}{3}$	$\frac{1}{6}$
Gamma(α, β)	$\frac{1}{\sqrt{\pi}} \frac{\beta\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}$	$6I_{1/3}(\alpha, 2\alpha) - 3$	Available in [39]
Normal(μ, σ^2)	$\frac{\sigma}{\sqrt{\pi}}$	0	$30\frac{1}{\pi}\tan^{-1}\sqrt{2} - 9$
Pareto(α, β)	$\frac{\alpha}{\beta(1-1/\alpha)(2-1/\alpha)}$	$\frac{1+1/\alpha}{3-1/\alpha}$	$\frac{(1+1/\alpha)(2+1/\alpha)}{(3-1/\alpha)(4-1/\alpha)}$
Student's t, 2 df	$\pi/2^{8/2}$	0	3/8
Student's t, 4 df	$15\pi/64$	0	111/512
Uniform(a, b)	$\frac{(b-a)}{6}$	0	0

Note: $I_x(p, q)$ is the incomplete beta function.

Applications of Moments and L-moments.

A few test of hypothesis methods have been developed using the third and fourth standardized moments and L-moments ratios [76, 44, 18, 81, 35] for testing departures from normality. Pearson and others [76] introduced two tests of using the sample skewness and kurtosis against large scale simulated values obtained for the Normal distribution called the K^2 and R tests, and they compared the power of their tests

against well known tests for normality such as Shapiro-Wilk's, Shapiro-Francia's, and D'Agostino's. An extension of their method was proposed by Seier [81] using normalization transformations of the skewness and kurtosis. Jarque and Bera [44] derived a method for testing the normality, homoscedasticity, and independence of regression residuals by using the Lagrange multiplier procedure and the sample skewness and kurtosis of the residuals. D'Agostino and others [18] discussed the usefulness of using the test of skewness, test of kurtosis, as well as the K^2 test for testing normality especially in lieu of traditional methods such as the chi-squared test and the Kolmogorov-Smirnov test due to better power properties that the former three tests possess. A test on skewness and excess-kurtosis (which is the kurtosis measure corrected for the Normal distribution) for the one-way error component model used for testing normality was proposed by Galvao and others [26]. They developed two new statistics for testing skewness and kurtosis and derived the distributions of the statistics using large sample theory. They also noted that although the test was created for the one-way error component model, this assumption could be relaxed by considering the variance-covariance structure for skewness and kurtosis to accommodate deviations from the one-way error component model.

Elamir and Seheult [22] created a new test of symmetry using L-moments by using the third sample L-moment standardized by the standard error obtained from the sample. They then used a Quantile-Quantile plot to compare the distribution to the quantile values of a given known distribution. Harri and Coble [35] extended the work of Pearson and others [76] by using L-skewness and L-kurtosis in place of their conventional counterparts. L-comoments was applied to robust financial portfolio allocation by Yanou [99] where they used Random Matrix Theory to extract information from the L-variance-covariance matrix.

A great deal of research has been conducted with respect to the application of L-

moments on the field of environmental sciences to summarize data and fit frequency distributions. Some examples include analysis on the flood frequency in the KwaZulu-Natal province as conducted by Kjeldsen and others [47], low streamflow analysis in the United States by Kroll and Vogel [51], analysis on the flow-duration-frequency behavior of British rivers based on small sample data by Zaidman and others [104], extreme wind quantile estimation using frequency analysis by Goel and others [31], and distribution estimation of Canadian annual minimum steamflow by Yue and Pilon [100]. A more extensive list of examples can be found in Hosking and Wallis [42].

A method of deriving a distribution with maximum entropy by conditioning on the first r L-moments was proposed by Hosking [41]. The Principle of Maximum Entropy states that if nothing is known about a distribution other than the class in which it belongs, then the distribution with the largest entropy should be chosen as the closest fit. By maximizing entropy, the amount of prior information built into the distribution is minimized. Given a continuous and differentiable CDF, F, with PDF, f, that is non-zero within its support and a set of L-moments, Hosking showed that the distribution that gives the maximum entropy has a density-quantile function that is a polynomial of degree r called the polynomial density-quantile (PDQ) which is a function of the quantile function, Q, as listed in Table 5. Here, entropy is defined as

$$H = \int_{-\infty}^{\infty} \{-\log f(x)\} f(x)\, dx.$$

The quantile function Q is defined as the inverse of the CDF, $F(Q(u)) = u, 0 < u < 1$. The quantile function is continuous and differentiable on an open bound $(0, 1)$, and $Q'(u) = \frac{1}{f(Q(u))}$. The density quantile function is then defined as $f(Q(u))$. Hosking illustrated the use of this method by performing a nonparametric estimation of the distribution for suicide rate data [86] and compared it against a kernel method as proposed by Wand and Jones [93]. The distribution of the data was shown to be

skewed right. However, even though the results show that the density fits are similar, the PDQ distribution appears to be less influenced by the larger data values.

Table 5. Maximum entropy distribution under different constraints on L-moments as adapted from [41]

Specified L-moments	**Range of Distribution**		
	$(-\infty, \infty)$	$[L, \infty)$	$[L, U]$
None	No Solution	No Solution	Uniform
λ_1	No Solution	Exponential	Truncated Exponential
λ_1, λ_2	Logistic	PDQ_2[1]	PDQ_2[1]
$\lambda_1, \lambda_2, \lambda_3$	$PDQ3$	PDQ_3	PDQ_3
$\lambda_1, \lambda_2, \lambda_4$	PDQ_4	PDQ_4[2]	PDQ_4
$\lambda_1, \lambda_2, \ldots, \lambda_r$	PDQ_r	PDQ_r	PDQ_r
λ_2	Logistic[3]	No Solution	PDQ_2
$\lambda_2, \ldots, \lambda_r$	PDQ_r[3]	No Solution	PDQ_r
$\lambda_3, \ldots, \lambda_r$	No Solution	No Solution	PDQ_r

PDQ_r denotes a PDQ distribution whose density-quantile function is a polynomial of degree r. PDQ3 is a PDQ_3 distribution whose density-quantile function $f(Q(u))$ is zero at $u = 0$ and 1.

2.3 Summary

This chapter presents a literature review and background on two key components for this dissertation; these are 1) graph theoretic definitions and some of the resulting development with which it is associated and 2) definitions of statistical concepts to include power laws and L-moments. Graph definitions were discussed to include graph and network measures, particularly those relating to the degree. It was also shown that some of the more basic measures such as degree, which is highly correlated to more complicated measures, have the advantage of being more computationally efficient. In addition, it has been shown that higher moments and L-moments of the degree distribution are statistically significant in characterizing the network as opposed to using a collection of various measures as listed in Table 2. These results

[1]Some instances of PDQ_2 are truncated logistic distribution
[2]Symmetric distribution
[3]Location parameter undetermined

are the motivation for the selection of network degree measure and features of the degree distribution as a basis for the developed tests in this research. A discussion on graph generating algorithms was briefly given to demonstrate some of the possible ways to randomly create synthetic graphs with known characteristics. Lastly, graph matching and classification techniques were discussed to underline some of the useful ideas that has emanated from the area particularly that of characterizing a network using its local structures.

The power law is then discussed in further detail specifically as it relates to the degree distribution of the Barabási-Albert graph. An introduction of methods for estimating the continuous power law which includes the LS method on the log of the histogram of the distribution, LS method on the log of the SF, and MLE-nonparametric methods were given. Particularly, the LS method on the log of the PDF is susceptible to the binning that is chosen for the histogram, whereas the LS method on the log of the SF provides a better estimation since the SF is not susceptible to binning. This is followed by explanation of the discrete versions of the power law, the Zipf law and Yule-Simon distribution. For moments and L-moments, a few important definitions were highlighted to help introduce the reader to the concepts. A list of tests of normality using moments and L-moments was then discussed followed by a method of deriving a probability density with maximum entropy for a random sample by defining the set of L-moments. Further, examples of the usefulness of these measures in characterizing and testing various phenomena were examined. These definitions and results motivate why these distributional characterization will be used in the derivation of the test statistics for the test of hypothesis.

III. Network Characterization

Recall that the degree distribution of the Barabási-Albert model is well represented by the Pareto distribution. A test of hypothesis for each of the parameters of the Pareto distribution will now be derived followed by the Union-Intersection test for simultaneously testing those parameters. These parameters link a particular member of the Pareto family to a particular Barabási-Albert graph. A simulation is then conducted in order to calculate the power of the individual tests as well as the Union-Intersection test. These methods are then applied on real world networks in an attempt to classify them as a Barabási-Albert network. Computational methods for bias and variance corrections are provided in order to appropriately apply the tests of hypotheses to network data.

3.1 Test of Hypothesis for the Pareto Distribution

Given a random sample of size N from a $Pareto(m, \beta)$ distribution, it is easily shown through the likelihood function

$$\begin{aligned} L(m|\mathbf{x}, \beta) &= \prod_{i=1}^{N} m^{\beta} \beta x_i^{-\beta-1} I_{[m,\infty)}(x_i) \\ &= m^{N\beta} \beta^N I_{[m,\infty)}\left(x_{(1)}\right) \prod_{i=1}^{N} x_i^{-\beta-1} I_{\left[x_{(1)},\infty\right)}(x_i), \end{aligned}$$

that the MLE for m is $\hat{m}_{MLE} = x_{(1)}$, and for β the MLE is

$$\hat{\beta}_{MLE} = N \left[\sum_{i=1}^{N} \ln \frac{x_i}{x_{(1)}} \right]^{-1} = N \left[\ln \frac{\prod_{i=1}^{N} x_i}{x_{(1)}^n} \right]^{-1}.$$

Consider the MLE for m. Fixing β, $L(m|\mathbf{x}, \beta)$ is monotone increasing in m, but $L(m|\mathbf{x}, \beta) = 0$ if $m > x_{(1)}$, so $\hat{m}_{MLE} = x_{(1)} = W(\mathbf{x})$. Note that $W(\mathbf{x})$ is also a sufficient statistic for X. Assuming that the degrees $X_1, ..., X_N$ are independent, the

PDF for $X_{(1)}$ is

$$\begin{aligned}
f_{X_{(1)}}(x) &= N f_X(x) \left(1 - F_X(x)\right)^{N-1} \\
&= N m^\beta \beta x^{-\beta-1} \left(1 - \int_m^x \alpha^\beta \beta u^{-\beta-1} du\right)^{N-1} \\
&= N m^\beta \beta x^{-\beta-1} \left(1 - m^\beta \beta \left(\frac{-1}{\beta u^\beta}\Big|_{u=m}^{x}\right)\right)^{N-1} \\
&= N m^\beta \beta x^{-\beta-1} \left(\frac{m^\beta}{x^\beta}\right)^{N-1} \\
&= N \beta x^{-\beta N-1} m^{N\beta} \\
&= m^{N\beta} N \beta x^{-\beta N-1} \\
\Rightarrow\ & f_{X_{(1)}}(x) \sim Pareto(m, N\beta).
\end{aligned}$$

Now consider again $L(m|\mathbf{x}, \beta)$. The loglikelihood function is

$$\mathcal{L}(m, \beta|\underset{\sim}{\mathbf{x}}) = N\beta \ln m + N \ln \beta - (\beta + 1) \sum_{i=1}^{N} \ln(x_i)$$

and taking the derivative of $\mathcal{L}(m, \beta|\underset{\sim}{\mathbf{x}})$ with respect to β results in:

$$\begin{aligned}
\frac{\partial}{\partial \beta}\mathcal{L}(\beta|m = x_{(1)}, \underset{\sim}{\mathbf{x}}) &= N \ln x_{(1)} + \frac{N}{\beta} - \sum_{i=1}^{N} \ln(x_i) \equiv 0 \\
\Rightarrow \frac{N}{\beta} &= \sum_{i=1}^{N} \ln(x_i) - N \ln x_{(1)} \\
\Rightarrow \hat{\beta} &= \frac{N}{\sum_{i=1}^{N} \ln(x_i) - N \ln x_{(1)}} \\
&= \frac{N}{\sum_{i=1}^{N} \ln(x_i) - \sum_{i=1}^{N} \ln x_{(1)}} \\
&= \frac{N}{\sum_{i=1}^{N} \{\ln(x_i) - \ln x_{(1)}\}} \\
\Rightarrow \hat{\beta} &= N \left[\sum_{i=1}^{N} \ln \frac{x_i}{x_{(1)}}\right]^{-1}.
\end{aligned}$$

To guarantee that $\hat{\beta}$ is an MLE the second derivative of $\mathcal{L}(m, \beta|\underset{\sim}{\mathbf{x}})$ with respect to β must be less than zero, which it is since

$$\frac{\partial^2}{\partial \beta^2}\mathcal{L}(\beta|m = x_{(1)}, \underset{\sim}{\mathbf{x}}) = -\frac{N}{\beta^2} < 0.$$

Therefore, $\hat{\beta}_{MLE} = N\left[\sum_{i=1}^{N} \ln \frac{x_i}{x_{(1)}}\right]^{-1}$. The MLEs will be used to derive the appropriate tests of hypotheses for both m and β. The tests of hypotheses will be evaluated through consideration of their power function and determination of whether or not the test maintained its size. The power function is defined as follows:

Definition 1 *The power function of a hypothesis test on rejection region $\mathcal{R}$ is the function of θ defined by $\mathbb{P}(\theta) = P_\theta(x \in \mathcal{R})$ [14]. Further, a test with power function $\mathbb{P}(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \mathbb{P}(\theta) = \alpha$, for $0 \leq \alpha \leq 1$.*

The power of each test of hypothesis was examined assuming an $\alpha = .05$ level of significance for each test.

3.1.1 Test for m.

Theorem 1 *Let $\underset{\sim}{\mathbf{X}}$ be a random sample of size N from $Pareto(m, \beta)$ with arbitrary β and let $x_{(1)}$ be an estimate of m. Then a test on the hypothesis $H_0 : m \leq m_1$ vs $H_A : m > m_1$ that rejects H_0 if $x_{(1)} \geq m_1/\alpha^{\frac{1}{N\beta}}$ is a level α test.*

Proof Let $W(\mathbf{x}) = x_{(1)}$, the MLE for m and a sufficient statistic. The PDF for $X_{(1)}$ was shown to be a $Pareto(m, N\beta)$ distribution. Consider a one-sided test for a particular value of m, call it m_1, through the hypothesis $H_0 : m \leq m_1$ vs $H_A : m > m_1$

where β is known. Then, the likelihood ratio test (LRT) can be obtained by

$$\begin{aligned}\lambda^*\left(W\left(\mathbf{x}\right)\right) &= \frac{\max_{m_1} L^*\left(m|W\left(\mathbf{x}\right)\right)}{\max_m L^*\left(m|W\left(\mathbf{x}\right)\right)} \\ &= \frac{m_1^{N\beta} N\beta w^{-\beta N-1}}{x_{(1)}^{N\beta} N\beta w^{-\beta N-1}} \\ &= \left(\frac{m_1}{x_{(1)}}\right)^{N\beta}\end{aligned}$$

which has a rejection region of $\{\mathbf{x} : \lambda^*\left(W\left(\mathbf{x}\right)\right) \leq c\} = \left\{\mathbf{x} : x_{(1)} \geq \frac{m_1}{c^{\frac{1}{N\beta}}}\right\}$. Therefore, a level α test where $\alpha = P\left(X_{(1)} \geq m_1/c^{\frac{1}{N\beta}}\right)$ will reject H_0 if

$$x_{(1)} \geq m_1/\alpha^{\frac{1}{N\beta}}. \tag{16}$$

■

The power of the test on $H_0 : m \leq m_1$ vs $H_A : m > m_1$ was computed for $m \in [1, 7]$ at increments of 0.2 for each $N \in \{2^k : k = 5, 6, \ldots, 15\}$ using 1000 iterations of the test described in Equation (16) with $\alpha = 0.05$. Here, m is the smallest value possible from the Pareto random sample and k is the index used for the sample size. A plot of the power curve is given in Figure 5 for $m \in \{2, 4, 6\}$. Note that the true m_1 for the Barabási-Albert graph is dependent on N for smaller networks such that $m_1 = m\sqrt{\frac{(N-m)}{N}}$. Further, the power curve suggests that for $k \geq 9$ ($k \geq 8$ for $m = 2$), the power of the test converges to a steady state, and the power of the test is relatively poor for $k \leq 8$ ($k \leq 7$ for $m = 2$). In the case where $k \leq 8$ ($k \leq 7$ for $m = 2$), the power slowly drops to α as m approaches m_1 which indicates that the TypeII error in the neighborhood of m_1 is fairly high. Also, since this is a one-sided test, as soon as $m > m_1$, the power quickly approaches zero which is expected. The values for the power at specific values of m are listed in Table A.1.

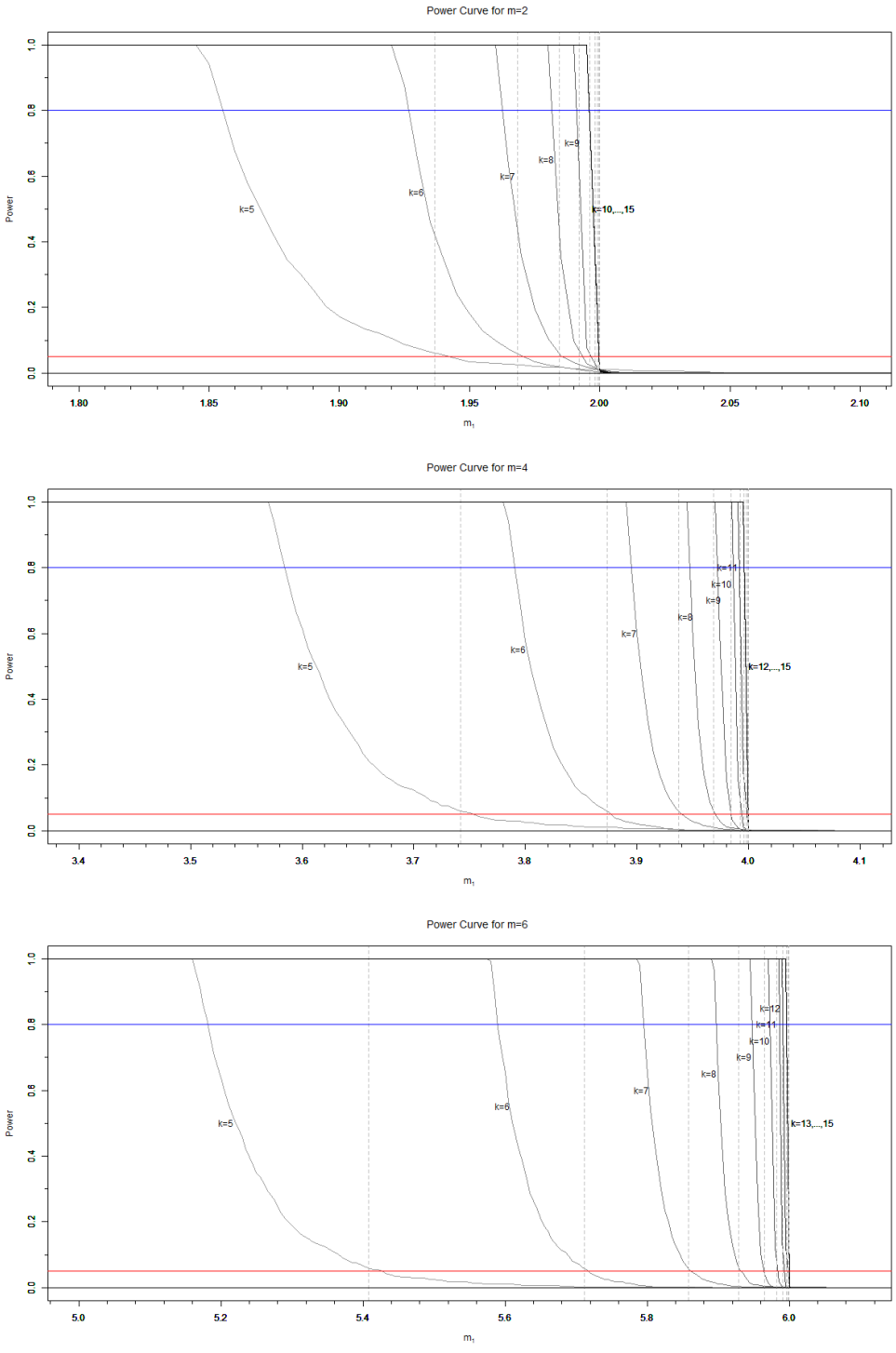


Figure 5. Power curve for the test on m for top) $m = 2$, middle) $m = 4$, and bottom) $m = 6$. Note:Line order from left to right represents $k \in \{5, 6, \ldots, 15\}$, respectively.

3.1.2 Test for β.

Theorem 2 *Let $\underset{\sim}{\mathbf{X}}$ be a random sample of size N from Pareto(m, β) with arbitrary m and define $T(\mathbf{x}) = \ln\left(\prod x_i / x_{(1)}^N\right)$. Then a test on the hypothesis $H_0 : \beta = \beta_0$ vs $H_A : \beta \neq \beta_0$ that rejects H_0 if*

$$T \leq \frac{z_\alpha \sqrt{(N-1)} + (N-1)}{\beta_0} \quad or$$
$$T \geq \frac{z_{1-\alpha} \sqrt{(N-1)} + (N-1)}{\beta_0},$$

is a level α test.

Proof Let $T(\mathbf{x}) = \ln\left(\prod x_i / x_{(1)}^N\right)$, then the MLE for β becomes $\frac{N}{T}$. Consider the test for β with the hypothesis $H_0 : \beta = \beta_0$ vs $H_A : \beta \neq \beta_0$ where m is unknown. Then, $\hat{m} = x_{(1)}$ and the LRT is

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\max_{\beta_0} L(\beta|\mathbf{x})}{\max_{\beta} L(\beta|\mathbf{x})} \\
&= \frac{\prod x_{(1)}^{\beta_0} \beta_0 x_i^{-\beta_0 - 1}}{\prod x_{(1)}^{\frac{N}{T}} \frac{N}{T} x_i^{-\frac{N}{T} - 1}} \\
&= \left(\frac{\beta_0 T}{N}\right)^N \left(\frac{\prod x_i}{x_{(1)}^N}\right)^{\frac{N}{T} - \beta_0} \\
\lambda(\mathbf{x}) &= \left(\frac{\beta_0 T}{N}\right)^N e^{N - \beta_0 T},
\end{aligned}$$

which has a rejection region of $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\} = \left\{\mathbf{x} : \left(\frac{\beta_0 T}{N}\right)^N e^{N - \beta_0 T} \leq c\right\}$. This rejection region is decreasing in T when $\frac{N}{\beta_0} \leq T$ and increasing when $\frac{N}{\beta_0} > T$. It has been shown that $2\beta_0 T$ has a $\chi^2_{2(N-1)}$ distribution [14], therefore a test that rejects H_0 if

$$T \leq \frac{\chi^2_{2(N-1),\alpha/2}}{2\beta_0} \text{ or } T \geq \frac{\chi^2_{2(N-1),1-\alpha/2}}{2\beta_0} \tag{17}$$

is a level α test.

Now consider a simple test of $H_0 : \beta = \beta_0$ vs $H_A : \beta = \beta_1$ where m is unknown. Then, by the Neyman-Pearson Lemma, the Uniformly Most Powerful (UMP) level α test is obtained by rejecting H_0 if $P\left(\frac{f(\mathbf{x}|m,\beta_1)}{f(\mathbf{x}|m,\beta_0)} < d\right) = \alpha$ for some $d \geq 0$. Note that

$$\begin{aligned}\frac{f(\mathbf{x}|m,\beta_1)}{f(\mathbf{x}|m,\beta_0)} &= \frac{\prod x_{(1)}^{\beta_1}\beta_1 x_i^{-\beta_1-1}}{\prod x_{(1)}^{\beta_0}\beta_0 x_i^{-\beta_0-1}} \\ &= \left(\frac{\beta_1}{\beta_0}\right)^N e^{(\beta_0-\beta_1)T}.\end{aligned}$$

For $\beta_0 > \beta_1$ this implies

$$\begin{aligned}d &> \left(\frac{\beta_1}{\beta_0}\right)^N e^{(\beta_0-\beta_1)T} \\ T &< \frac{\ln d + N\ln\left(\frac{\beta_0}{\beta_1}\right)}{(\beta_0-\beta_1)}\end{aligned}$$

and for $\beta_0 < \beta_1$,

$$T > \frac{\ln d + N\ln\left(\frac{\beta_0}{\beta_1}\right)}{(\beta_0-\beta_1)}.$$

Thus, by the fact that $2\beta_0 T \sim \chi^2_{2(N-1)}$, this implies $E\left[2\beta_0 T\right] = 2(N-1)$ and $Var\left[2\beta_0 T\right] = 4(N-1)$. Therefore for large N, a UMP level α test such that $\alpha = P\left(\frac{2\beta_0 T - 2(N-1)}{\sqrt{4(N-1)}} \leq z_\alpha\right)$ will reject H_0 if

$$\begin{aligned}T \leq& \frac{z_\alpha\sqrt{(N-1)} + (N-1)}{\beta_0} \text{ for } \beta_0 > \beta_1 \\ &\text{or} \\ T \geq& \frac{z_{1-\alpha}\sqrt{(N-1)} + (N-1)}{\beta_0} \text{ for } \beta_0 < \beta_1,\end{aligned} \tag{18}$$

where z_α is the α^{th} percentile of the Standard Normal distribution. ■

Note that the test in Equation (18) is not dependent on the value of β_1, is based on the χ^2 distribution, and is identical to the two-sided test in Equation (17). Therefore, the two-sided test, $H_0 : \beta = \beta_0$ vs $H_A : \beta \neq \beta_0$, is identical to the simple test, $H_0 : \beta = \beta_0$ vs $H_A : \beta = \beta_1$. It should also be noted that in cases where m is known, the statistic T becomes $\ln\left(\frac{\prod_{i=1}^{N} x_i}{m^N}\right)$ and the transformation $2\beta_0 T$ has the distribution of χ^2_{2N}.

The power of the test on $H_0 : \beta = 2$ vs $H_A : \beta \neq 2$ for the Barabási-Albert degree distribution was computed for $\beta \in [1, 3]$ at increments of 0.02 for each Barabási-Albert graph with $k \in \{5, 6, \ldots, 15\}$. Similar to the method used for the test for m, 1000 iterations for each Barabási-Albert graph were performed at each increment of 0.02, and the proportion of rejections based on the test described in Equation (18) were computed. A plot of the power curve for this test is given in Figure 6 with values listed in Table 6. It is apparent from Figure 6, that the power of the test improves as k increases. However, for a fixed k, the power curves are identical for different values of m (Figures A.1 and A.2), implying that the power is invariant to m.

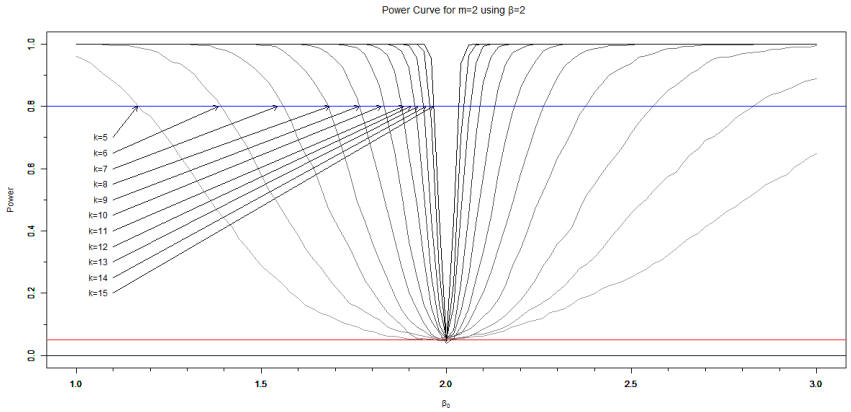


Figure 6. Power curve for the test on β. Note: Lighter shaded lines indicates smaller k

Table 6. Power of the test for Pareto with $\beta_0 = 2$ where $\delta = |\beta - \beta_0|$

	k										
δ	5	6	7	8	9	10	11	12	13	14	15
0	.054	.050	.049	.059	.053	.039	.054	.055	.055	.046	.047
.02	.050	.051	.047	.054	.050	.073	.078	.098	.164	.272	.446
.04	.051	.058	.052	.061	.070	.111	.155	.256	.438	.754	**.956**
.08	.053	.067	.054	.092	.148	.252	.487	.759	**.956**	**1**	**1**
.16	.062	.092	.138	.247	.464	.756	**.966**	**1**	**1**	**1**	**1**
.24	.128	.251	.479	**.807**	**.972**	**1**	**1**	**1**	**1**	**1**	**1**
.32	.267	.533	**.876**	**.995**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.64	.503	**.844**	**.994**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
1	**.960**	**.999**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

3.1.3 Union-Intersection Test.

Theorem 3 *Let $\underset{\sim}{\mathbf{X}}$ be a random sample of size N from a Pareto(m, β). Let $x_{(1)}$ be an MLE for m and define $T = \ln\left(\prod x_i / x_{(1)}^N\right)$. Then a test on the hypothesis*

$$H_0 : \{m \leq m_1 \cap \beta \leq \beta_0 \cap \beta \geq \beta_0\}$$

$$H_A : \{m > m_1 \cup \beta > \beta_0 \cup \beta < \beta_0\}.$$

that rejects H_0 if

$$\left\{ \mathbf{x} : x_{(1)} \geq \frac{m_1}{\alpha^{\frac{1}{N\beta_0}}} \text{ or } T \leq \frac{z_\alpha \sqrt{(N-1)} + (N-1)}{\beta_0} \text{ or } T \geq \frac{z_{1-\alpha} \sqrt{(N-1)} + (N-1)}{\beta_0} \right\}$$

is a level α test.

Proof A Union-Intersection Test (UIT) can be formed if the null hypothesis can be expressed as an intersection. In this case, test for m and β simultaneously by forming

the hypotheses

$$\begin{aligned} H_0 &: \{m \leq m_1 \cap \beta = \beta_0\} = \{m \leq m_1 \cap \beta \leq \beta_0 \cap \beta \geq \beta_0\} \\ H_A &: \{m > m_1 \cup \beta > \beta_0 \cup \beta < \beta_0\}. \end{aligned} \tag{19}$$

Define $C(\mathbf{x}) = \inf_{\gamma \in \{m,\beta\}} \lambda_\gamma(\mathbf{x})$ where $\lambda_\gamma(\mathbf{x})$ is the LRT for the individual tests. Since each $\lambda_\gamma(\mathbf{x})$ is a level α test, then the UIT based on $C(\mathbf{x})$ is a level α test. Therefore, the rejection region for Equation (19) is given by

$$\left\{ \mathbf{x} : x_{(1)} \geq \frac{m_1}{\alpha^{\frac{1}{N\beta_0}}} \text{ or } T \leq \frac{z_\alpha \sqrt{(N-1)} + (N-1)}{\beta_0} \text{ or } T \geq \frac{z_{1-\alpha} \sqrt{(N-1)} + (N-1)}{\beta_0} \right\}$$

and is a level α test. ■

If $\mathbb{P}_C(\boldsymbol{\theta})$ and $\mathbb{P}_\lambda(\boldsymbol{\theta})$ are the power functions for the tests based on C and λ, respectively, then $\mathbb{P}_C(\boldsymbol{\theta}) \leq \mathbb{P}_\lambda(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \{(m, \beta)\}$. Hence the power of the UIT will be bounded by the power of the individual LRTs at the specified level of $\boldsymbol{\theta}$. Therefore, the power of the UIT at each level of $\boldsymbol{\theta}$ is simply the maximum of the power of the each individual test at the particular level. The power of the UIT can be visualized using a surface plot with respect to $(m, \beta, \mathbb{P}_C(\boldsymbol{\theta}))$ on the (x, y, z) axes, respectively.

The power of the UIT is plotted in Figures 7 to 10 for $k = 5, 10$ and $m = 2, 4, 6$ with the associated values included in Table A.2. The power of the UIT is an improvement over the individual tests except when it is stationary at 0.05 for $\beta_0 = 2$ and when m_1 is greater than the true m. Therefore, if a given distribution is truly not from a Pareto distribution, the probability of rejecting will be higher for $\beta_0 = 2$ when m_1 is hypothesized lower than the true value, $m\sqrt{\frac{(N-m)}{N}}$, as opposed to one that is higher. However, if m_1 is hypothesized to be the true value, it does not matter if β_0 is overestimated or underestimated as it will result in roughly the same probability

of rejection. Additionally, similar to the power of the individual tests, the power of the UIT improves as k increases, particularly on the β_0 axis. That is, for smaller k, the combination of m_1 and β_0 affects the power of the test much more than when k is large, at which point only β_0 affects the power of the test.

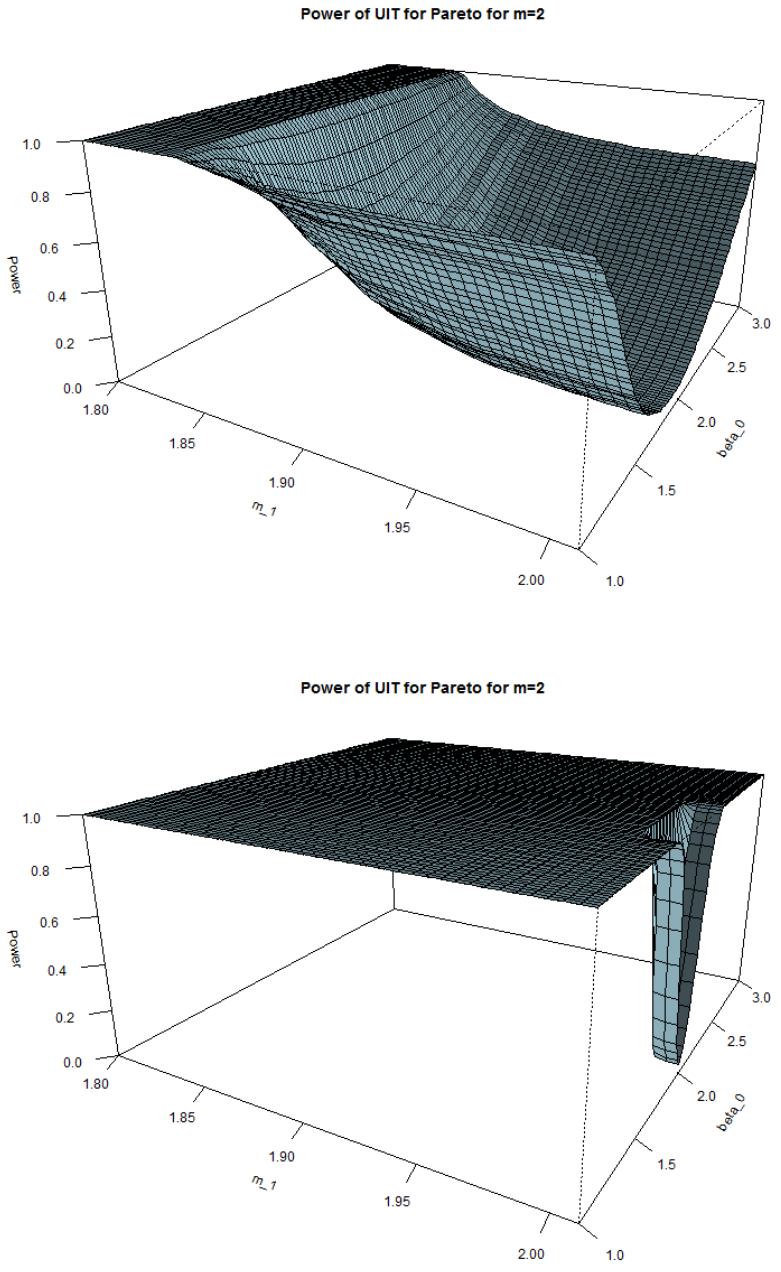


Figure 7. Surface plot of power for the Pareto UIT for $k = 5, 10$ and $m = 2$.

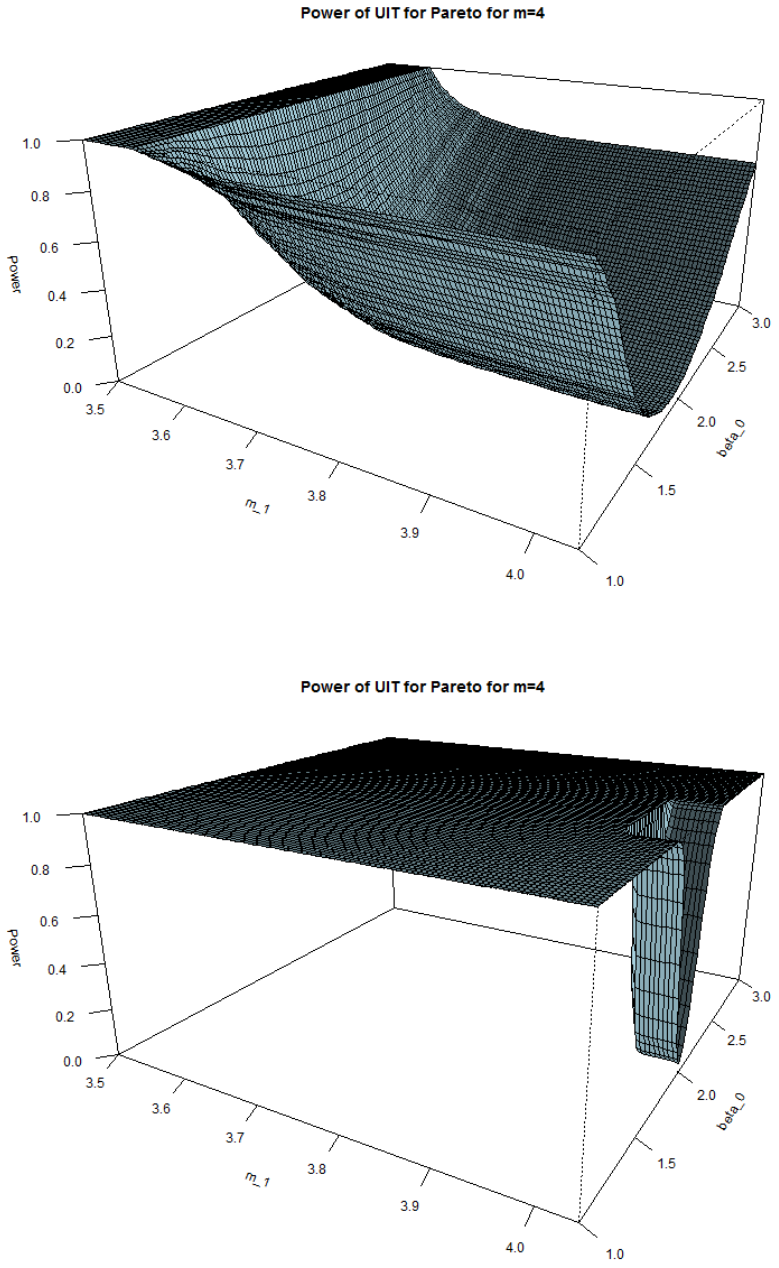


Figure 8. Surface plot of power for the Pareto UIT for $k = 5, 10$ and $m = 4$.

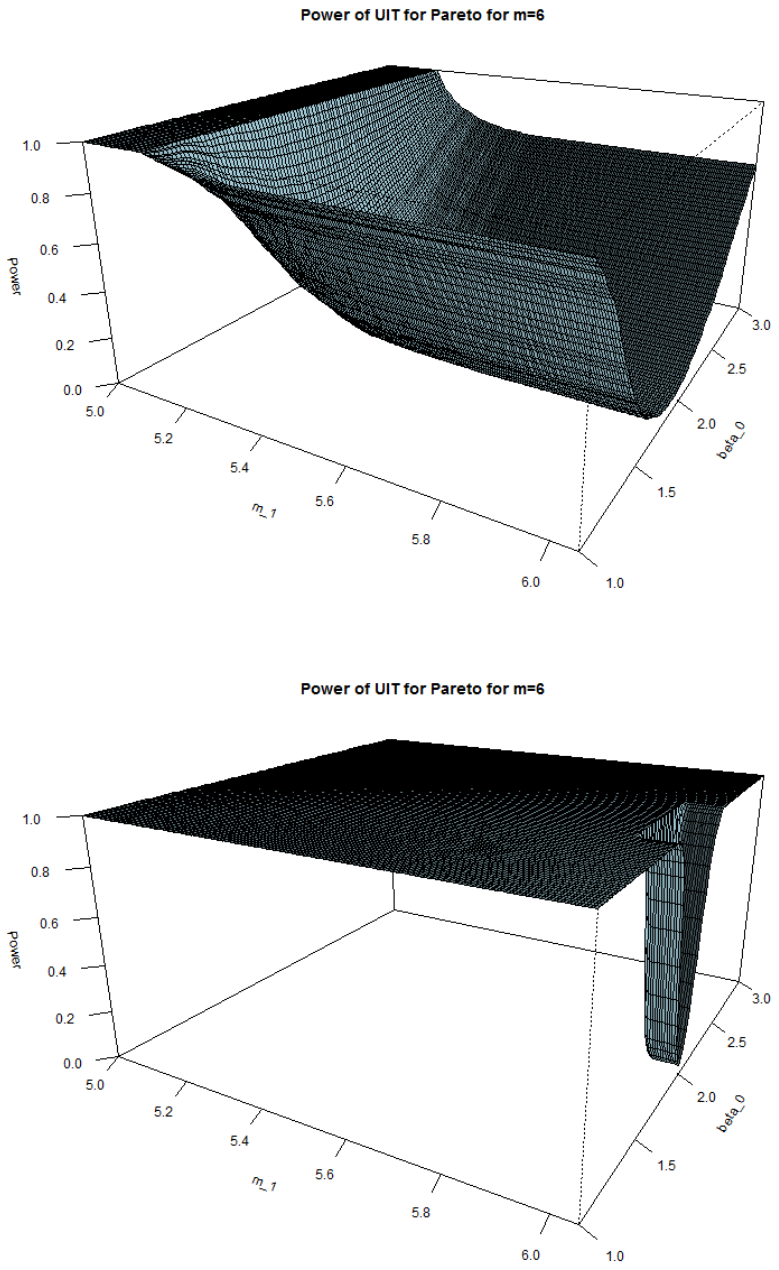


Figure 9. Surface plot of power for the Pareto UIT for $k = 5, 10$ and $m = 6$.

3.2 Test of Hypothesis for the Barabási-Albert Network

Network data was simulated to calculate the power of each test derived in Section 3.1 as applied to the Barabási-Albert network. A dataset comprised of degree

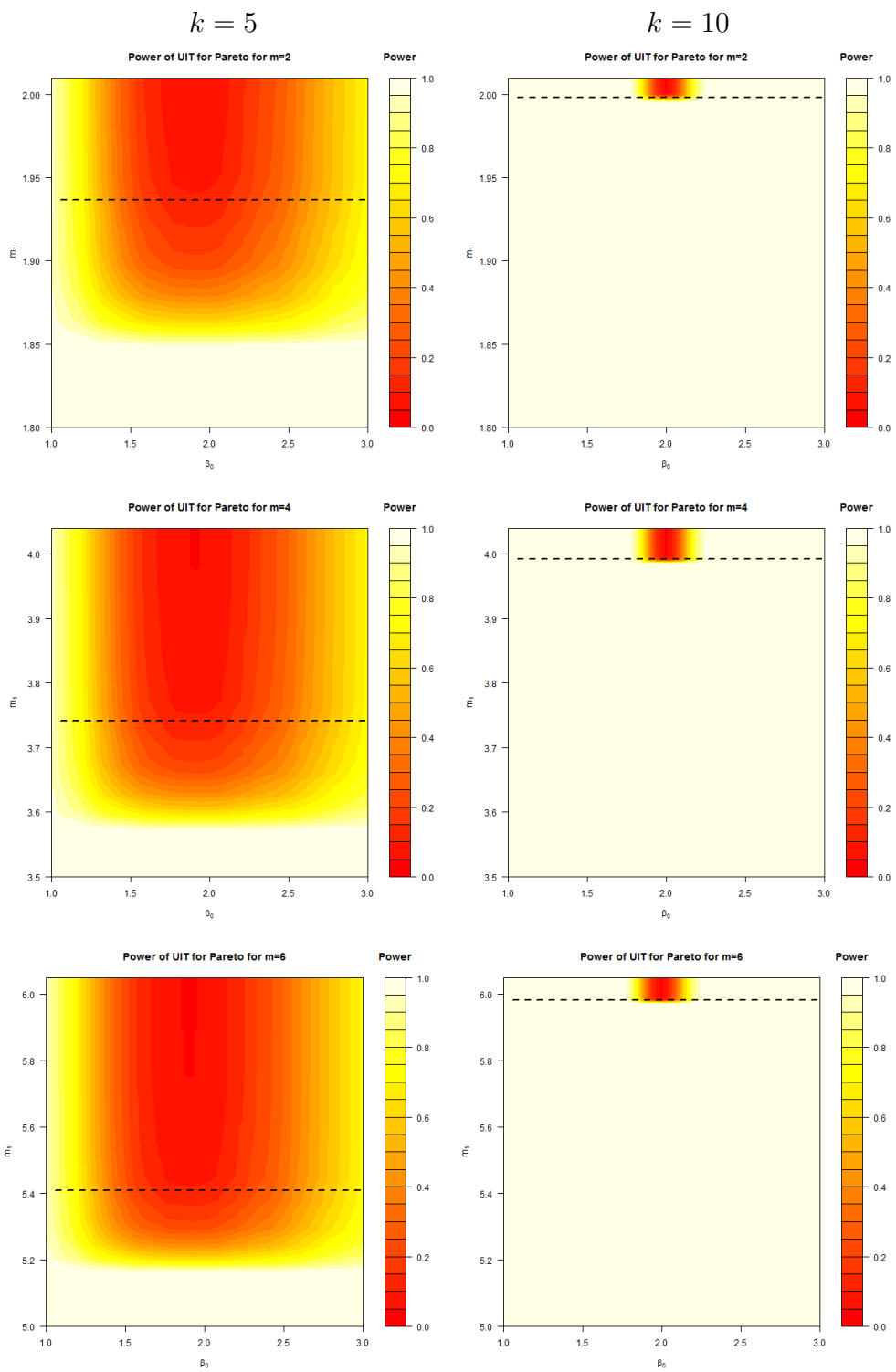


Figure 10. Contour plot of power for the Pareto UIT for $k = 5, 10$ and $m = 2, 4, 6$, from top to bottom, respectively. (Note: Dashed lines are true values for m_1.)

distributions of the simulated Barabási-Albert graph for various parameter and size combinations was generated using the *igraph* package in R [17]. The parameter selection for the simulation is listed in Table 7 where 1000 independent networks were generated for each of the 44 combinations of graph parameter and sizes. These combinations were chosen so that the sizes of the networks examined spanned from small to large, and to also study the behavior of the Barabási-Albert graphs that have nodes with low degrees (i.e. $m^* = 1, 2, 4, 6$). Note that m^* is the value used for network simulation and not to be confused with m which is a parameter of the Pareto distribution.

Table 7. Parameters for network simulation

Parameters	**Size**
$m^* \in \{1, 2, 4, 6\}$	$N = 2^k; k \in \{5, 6, \ldots, 14, 15\}$

Upon inspection of the simulated data, the evolution of the network generation causes the degree of the nodes to be correlated to one another due to the preferrential attachment nature of the Barabási-Albert graph. This is not unexpected as Li and others [59] have observed that the preferential attachment model causes biases in the structure of the graph where high degree nodes are interconnected. The correlation causes the expected value of $Z = 2\beta_0 T$ to be slightly biased and the variance to be much smaller than the expected variance when compared to the $\chi^2_{2(N-1)}$ distribution (Figure 11). Therefore, another set of data was simulated in order to study the behavior of the bias in the expected value and variance of Z. For each (m^*, k) combination, the sample mean and variance of Z was computed from a bootstrap of 1000 iterations and repeated 100 times resulting in a collection of 100 sample means and variances for each combination of graph parameter and size. The ratio of the sample mean and variance over their respective expected values, $\frac{\bar{z}}{E[z]}$ and $\frac{s_z^2}{var[z]}$, indicates that there is an

underlying nonlinear pattern with respect to k (Figure 12). Note that, theoretically, $\frac{\bar{z}}{E[Z]} = \frac{\bar{z}}{2d.f.}$ and $\frac{s_z^2}{var[Z]} = \frac{s_z^2}{4d.f.}$, where $d.f.$ stands for the degrees of freedom.

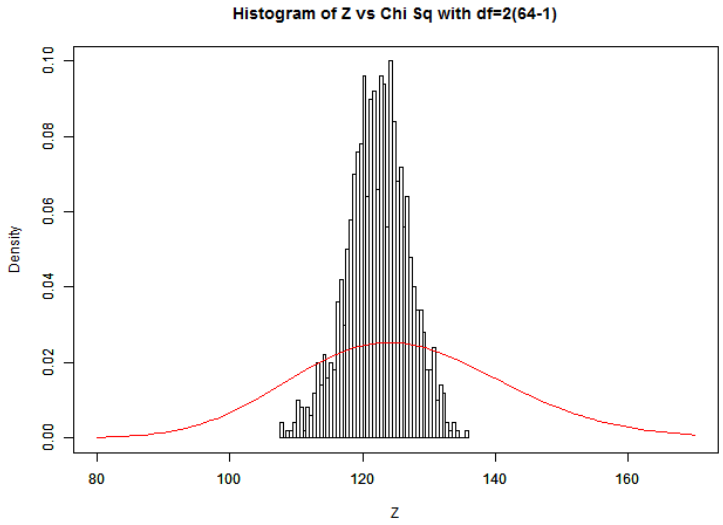


Figure 11. Empirical versus theoretical degree distribution of Z for $m^* = 2$ and $\beta_0 = 2$

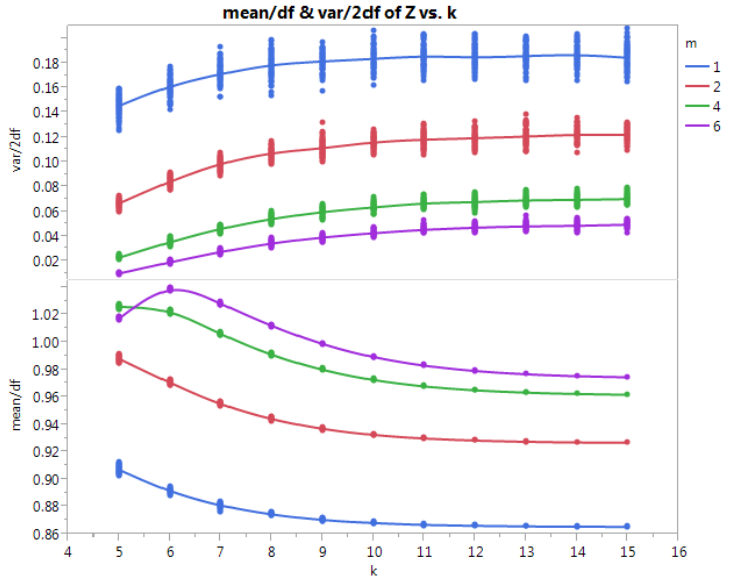


Figure 12. Top: Ratio of $\frac{s_z^2}{Var[Z]}$ versus network size. Bottom: Ratio of $\frac{\bar{z}}{E[Z]}$ versus network size. (Note: Solid lines are the expected ratios based on the Bioexponential and Gompertz models.)

Although there is a bias in the sample mean and variance from the empirical distribution of $2\beta_0 T$, it seems to converge as the network size increases. Furthermore,

for each m, both the ratios, $\frac{\bar{z}}{E[Z]}$ and $\frac{s_z^2}{Var[Z]}$, can be modeled by the 5 parameter Biexponential and 4 parameter Gompertz models as a function of k (all $R^2 \geq .99$) where $\frac{\bar{z}}{E[Z]}$ can be modeled as

$$f(k) = a + b\exp\{-ck\} + d\exp\{-hk\} \tag{20}$$

and $\frac{s_z^2}{Var[Z]}$ can be modeled as

$$g(k) = a + (b-a)\exp\{-exp\{-c(k-d)\}\}. \tag{21}$$

The estimates of the parameters for the correcting scalars for Equations (20) and (21) using $\beta_0 = 2$ were computed using the simulated data and are given in Table 8. Therefore, the level α test from Equation (18) becomes

$$\begin{aligned} T \leq & \frac{z_\alpha\sqrt{(N-1)g(k)} + (N-1)f(k)}{\beta_0} \text{ for } \beta_0 > \beta_1 \\ & \text{or} \\ T \geq & \frac{z_{1-\alpha}\sqrt{(N-1)g(k)} + (N-1)f(k)}{\beta_0} \text{ for } \beta_0 < \beta_1 \end{aligned} \tag{22}$$

which rejects $H_0 : \beta = \beta_0$ for the alternative $H_A : \beta \neq \beta_0$.

Table 8. Parameter estimates for $f(k)$ and $g(k)$.

	$f(k)$					$g(k)$			
m	a	b	c	d	h	a	b	c	d
1	0.8653	313.24	0.6947	-313.66	0.6958	0.1168	0.1859	0.6626	4.7800
2	0.9264	2.6131	0.5890	-5.5817	0.8596	-0.1465	0.1229	0.4391	1.6905
4	0.9610	3.4200	0.5532	-11.951	0.8753	-0.0064	0.0710	0.4558	4.9327
6	0.9733	4.1794	0.5390	-18.966	0.8754	-0.0084	0.0503	0.4259	5.3396

3.3 Power of the Test on m for Simulated Barabási-Albert Network

In this implementation of the test on m, the estimate $x_{(1)}$ will always be equal to $m^* \in \{1, 2, 4, 6\}$ due to the way the Barabási-Albert network is simulated where the smallest degree possible for any generation of the graph is m^*. This fact causes the test to behave differently than it would have with a theoretical distribution as shown in Section 3.1.1. Therefore, consider the test with two possible true values: $m = m^*$ or $m = m^*\sqrt{\frac{(n-m^*)}{n}}$, the theoretical m for a relatively small Barabási-Albert network. The power of the test described by Equation (16) on $H_0 : m \leq m_1$ vs $H_A : m > m_1$ is computed similar to the process as in Section 3.1.1, but at an increment of 0.005 for better resolution. The power curves for the test are plotted in Figure 13 in which the dashed vertical lines are the locations of the true m for each $k \in \{5, 6, \ldots, 15\}$ from left to right, respectively. With $m = m^*$, the power curve suggests that for $k \geq 11$ ($k \geq 9$ for $m^* = 1$ and $k \geq 10$ for $m^* = 2$) the power of the test converges to a steady state similar to the general test for m in Section 3.1.1. However, the power of the test is very poor for $k \leq 10$ ($k \leq 8$ for $m^* = 1$ and $k \leq 9$ for $m^* = 2$) where the power drops to zero even before m_1 approaches m.

If $m = m^*\sqrt{\frac{(n-m^*)}{n}}$, the test has a TypeI error of 100% for $m^* = 4, 6$. Since the dashed lines are to the left of the point where the power decreases from 100% to 0% for $m^* = 4, 6$ and $k \leq 11$. This indicates that H_0 will always be rejected even when H_0 is true. Therefore, when implementing the test on a simulated Barabási-Albert network in such cases, letting $m = m^*\sqrt{\frac{(n-m^*)}{n}}$ instead of $m = m^*$ essentially renders the test unusable and the latter should be used instead.

3.4 Power of the Test on β for Simulated Barabási-Albert Network

Using the appropriate values from Table 8 for $H_0 : \beta = 2$ vs $H_A : \beta \neq 2$, the power of the test in Equation (22) was computed for $\beta \in [1, 3]$ at increments of 0.02

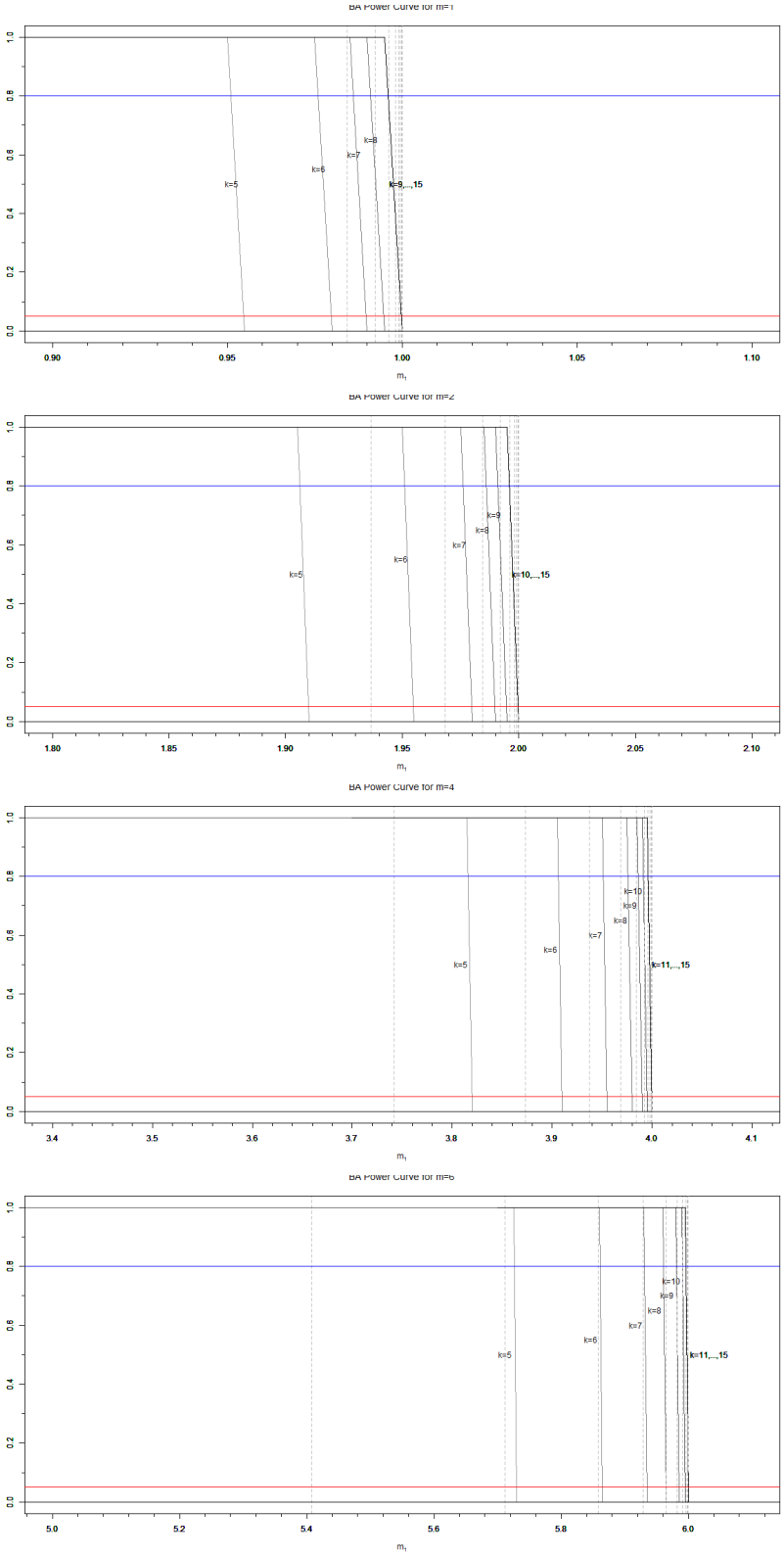


Figure 13. Power curve for the Barabási-Albert (BA) test on m for $m^* \in \{1, 2, 4, 6\}$. (Note: Line order from left to right represents $k \in \{5, 6, \ldots, 15\}$, respectively.)

for each $k \in \{5, 6, \ldots, 15\}$ similar to that of Section 3.1.2. The power curves are plotted in Figure 15 with the actual values listed in Table 9. Again, the power of the test improves as k increases, but unlike the general test for β from Equation (18) and plotted in Figure 6, it is apparent that as m^* increases, the power also improves considerably. Another noticeable difference is that the power increases at a much faster rate across all values of k.

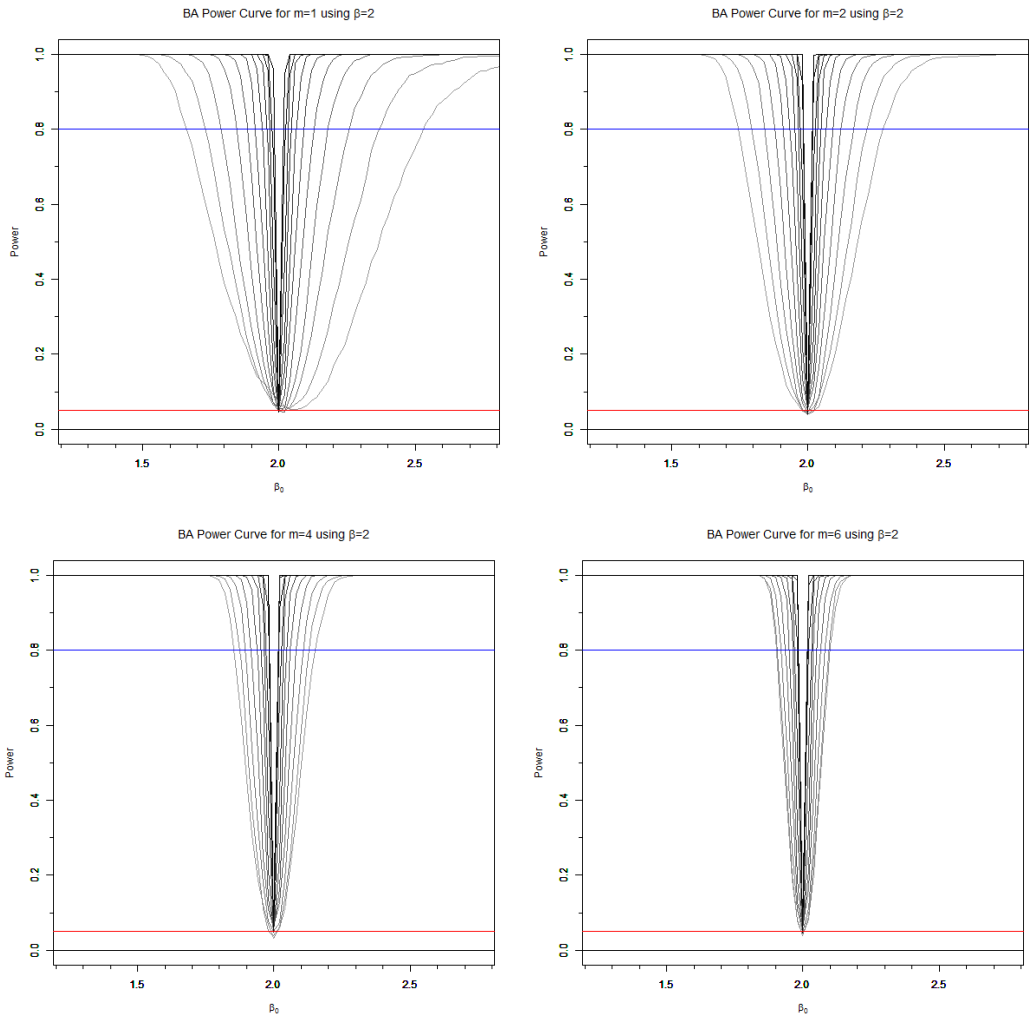


Figure 14. Power curve for the test on β for $m^* \in \{1, 2, 3, 4\}$. (Note: Lines converges towards $\beta = 2$ as k increases.)

The power with the assumption that $\beta_0 = 2.16$ and 2.45 was also examined where the estimates for the correcting scalars were computed using the appropriate β_0 values.

Table 9. Power of the test for $\beta_0 = 2$ where $\delta = |\beta - \beta_0|$

$m^* = 1$	k										
δ	5	6	7	8	9	10	11	12	13	14	15
0	.076	.063	.057	.049	.052	.049	.052	.045	.058	.049	.046
.02	.061	.062	.051	.046	.057	.091	.157	.246	.445	.751	**.948**
.04	.053	.055	.059	.082	.130	.270	.439	.728	**.944**	**.999**	**1**
.06	.052	.066	.082	.128	.266	.484	.760	**.976**	**1**	**1**	**1**
.08	.056	.079	.132	.211	.441	.707	**.940**	**.998**	**1**	**1**	**1**
.1	.062	.095	.189	.322	.603	**.874**	**.993**	**1**	**1**	**1**	**1**
.2	.167	.331	.592	**.862**	**.986**	**1**	**1**	**1**	**1**	**1**	**1**
.3	.357	.610	**.885**	**.988**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.4	.566	**.855**	**.979**	**.999**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

$m^* = 2$	k										
δ	5	6	7	8	9	10	11	12	13	14	15
0	.045	.039	.05	.045	.044	.043	.032	.048	.05	.043	.054
.02	.048	.067	.055	.065	.097	.13	.22	.39	.675	**.92**	**.996**
.04	.062	.078	.109	.153	.235	.408	.702	**.932**	**1**	**1**	**1**
.06	.085	.11	.161	.275	.473	.729	**.947**	**.999**	**1**	**1**	**1**
.08	.115	.177	.246	.444	.704	**.917**	**.999**	**1**	**1**	**1**	**1**
.1	.172	.233	.398	.633	**.876**	**.983**	**1**	**1**	**1**	**1**	**1**
.2	.539	.747	**.899**	**.991**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.3	**.853**	**.957**	**.995**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.4	**.976**	**.995**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

$m^* = 4$	k										
δ	5	6	7	8	9	10	11	12	13	14	15
0	.037	.044	.042	.049	.048	.048	.055	.054	.05	.056	.049
.02	.066	.071	.075	.097	.142	.246	.396	.665	**.915**	**.992**	**1**
.04	.101	.136	.177	.252	.446	.697	**.914**	**.997**	**1**	**1**	**1**
.06	.195	.256	.36	.554	.777	**.952**	**.998**	**1**	**1**	**1**	**1**
.08	.326	.391	.563	.744	**.94**	**.996**	**1**	**1**	**1**	**1**	**1**
.1	.448	.572	.771	**.913**	**.991**	**1**	**1**	**1**	**1**	**1**	**1**
.2	**.95**	**.981**	**.999**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.3	**.998**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.4	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

$m^* = 6$	k										
δ	5	6	7	8	9	10	11	12	13	14	15
0	.057	.053	.052	.051	.042	.044	.059	.05	.054	.045	.039
.02	.073	.081	.109	.127	.198	.341	.55	**.827**	**.98**	**1**	**1**
.04	.193	.209	.284	.416	.632	**.837**	**.982**	**.999**	**1**	**1**	**1**
.06	.4	.424	.567	.749	**.926**	**.989**	**1**	**1**	**1**	**1**	**1**
.08	.604	.644	.773	**.923**	**.997**	**1**	**1**	**1**	**1**	**1**	**1**
.1	.792	**.821**	**.922**	**.986**	**.999**	**1**	**1**	**1**	**1**	**1**	**1**
.2	**.999**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.3	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.4	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

However, for each detectable difference, $|\beta - \beta_0|$, the resulting power did not differ from that observed when $\beta_0 = 2$ which suggests that the power of the test is invariant of the hypothesized β_0 (Figure 15). This implies that regardless of the true exponent of the degree distribution of the Barabási-Albert network, the proposed test is able

to determine a given network as Barabási-Albert if the transformation of the degree distribution, $2\beta_0 T$, under the assumed β_0 follows that of the corrected transformation of the Barabási-Albert test as in Equation (22).

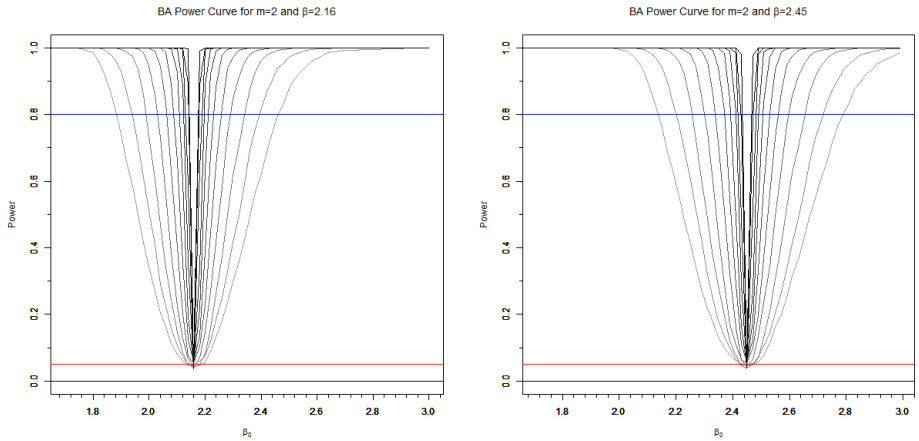


Figure 15. Power curve for the test with $\beta = 2.16$ and $\beta = 2.45$ for $m^* = 2$. (Note: Lines converges towards $\beta = 2$ as k increases.)

3.5 Power of the Union-Intersection test for Simulated Barabási-Albert Network

The power of the UIT for Barabási-Albert networks is an improvement over the individual tests except where it is stationary at 0.05 for when $\beta_0 = 2$ and m_1 is greater than the true value (Figure 16). However, since the degree distribution is discrete, the value of m_1 is set to equal to m^* instead of $m^*\sqrt{\frac{(N-m^*)}{N}}$ because the former is the smallest value that will be observed. Additionally, due to the tighter variance of the empirical distribution of Z, the power improves much faster when compared to the unadjusted UIT test for Pareto parameters in Section 3.1.3. The power also varies with respect to the value of m^* similar to the power of the individual test on β where higher power is observed for larger values of m^*. Interestingly, Figure 16 shows that if $\beta_0 = 2$ is hypothesized for a non-Barabási-Albert network, then a misspecification of m_1 that is lower than the true value of m will result in a higher probability of

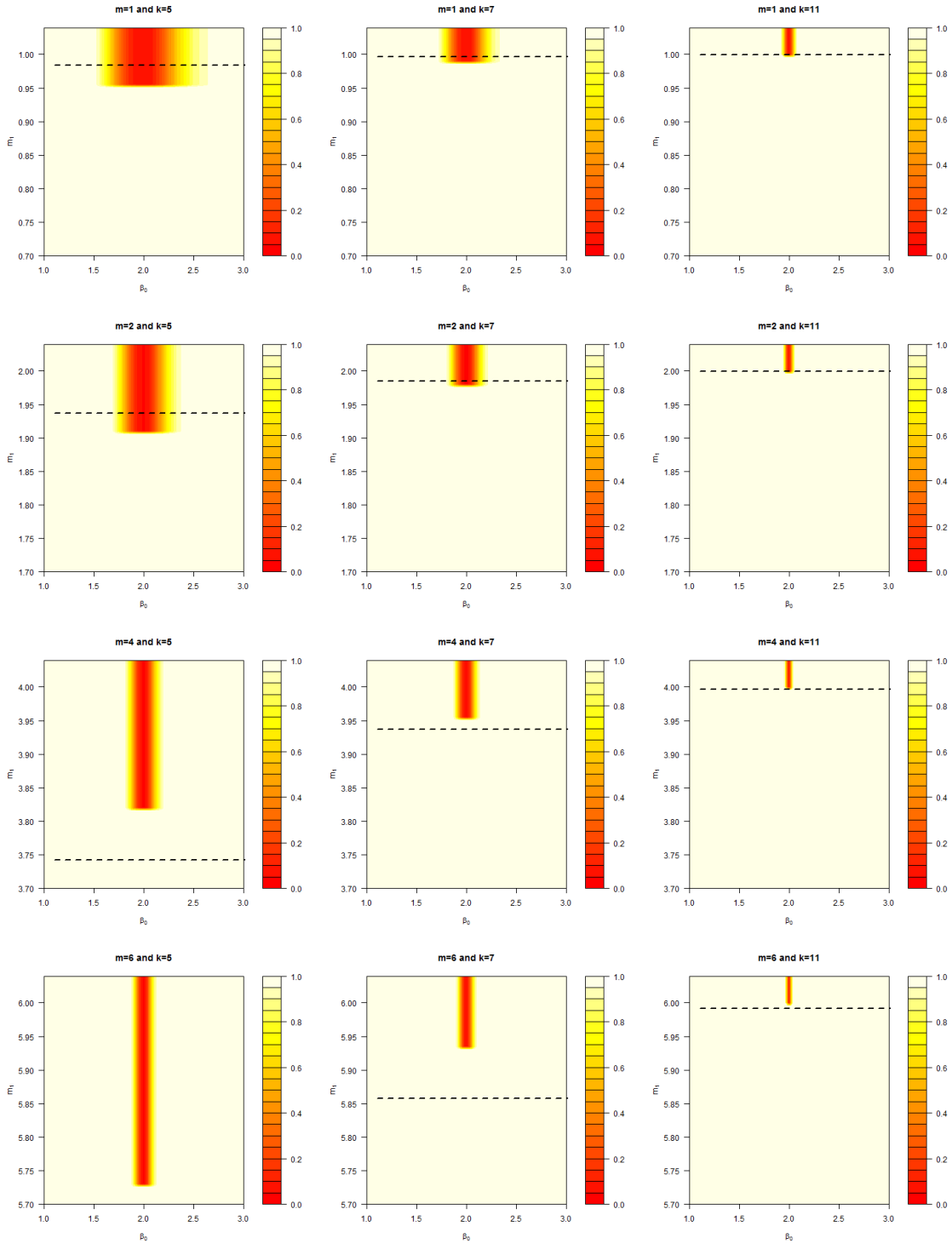


Figure 16. Contour plot of the power for the Barabási-Albert UIT for $k = 5, 7, 11$ from left to right, respectively, and $m^* = 1, 2, 4, 6$ from top to bottom, respectively. (Note: Dashed lines are theoretical values for m_1.)

rejection than an overspecification of m_1. However, if the hypothesized m_1 is equal to the true value, then it does not matter if β_0 is misspecified as it will result in the same probability of rejection. Additionally, similar to the power of the individual tests, the power of the UIT improves as k increases particularly on the β_0 axis. The values for the power of the UIT are included in Table A.3.

3.6 Real World Network Classification

To study the usability of the tests derived in this chapter, classifications on real world datasets are performed in order to see if there are real world networks that may be represented as a Barabási-Albert network based on the degree distribution. Although it is well known that the Barabási-Albert graph models the evolution of a graph that follows the preferential attachment property and is a scale free graph, the converse relationship is less studied. In other words, are there any real world networks that can be represented as a Barabási-Albert network? The datasets compiled are comprised of networks that are believed to be scale free from a variety of fields and also varies in terms of sizes that are available in the literature (Table 10). The datasets were collected by Newman [71] and Leskovec and Krevl [56]. It should be noted that all networks were treated as undirected networks for the analysis.

The UIT from Section 3.5 was then used to test whether or not the degree distribution from a given network is significantly different from that of a Barabási-Albert network of the same size. It is hypothesized that $m_1 = 1$ since it is the lowest theoretical value of the degree in a connected graph, and it is hypothesized that $\beta_0 = 2$ which is the theoretical value for a Barabási-Albert network. The results shows that the networks have parameter values that are significantly different than the values associated with a Barabási-Albert representation (Table 11). The large z-statistic (the $z|m_1 = 1$ column) and small p-value $\big($compared to $z_{\alpha/2=.025} = 1.96\big)$ for each network

Table 10. Real world data description.

Network	**Brief Description**	Type	$k : N = 2^k$	**Reference**
Karate Club	Social network of friendships	Undirected	5.0875	Zachary [102]
Les Miserables	Coappearance of *Les Miserables* characters	Undirected	6.2668	Knuth [48]
Dolphin Social Network	Social network of dolphins	Undirected	5.9542	Lusseau and others [60]
Political Blogs	Hyperlinks between US politics weblogs	Directed	10.5411	Adamic and Glance [3]
Condensed Matter Collaborations	Coauthorships between scientists	Undirected	15.3028	Newman [69]
Astrophysics Collaborations	Coathorships between scientists	Undirected	14.0281	Newman [69]
High-Energy Theory Collaborations	Coauthorships between scientists	Undirected	13.0295	Newman [69]
High-Energy Physics Theory Citations	Network of paper citations	Directed	14.7612	Gehrke and others [27], Leskovec and others [55]
High-Energy Physics Phenomenology Citations	Network of paper citations	Directed	15.0762	Gehrke and others [27], Leskovec and others [55]
Internet	Internet structure	Undirected	14.4870	Newman [71]
Google Webgraphs	Network of hyperlinks between webpages	Directed	19.7401	Leskovec and others [57]
Facebook Social Circles	Social network of friendships	Undirected	11.97978	McAuley and Leskovec [63]

suggests that the degrees of the networks are larger than what is expected from a Barabási-Albert network. Additionally, under the assumption that the degree distribution of the networks is the Pareto distribution, the $\hat{\beta}_{MLE}$ estimates obtained are well below the $\beta_0 = 2$ assumption for Barabási-Albert network, where the closest estimate was found for the Internet dataset with $\hat{\beta}_{MLE} = 1.4351$.

The UIT was then reapplied for further investigation by hypothesizing that $m_1 \in \{2, 4, 6\}$ instead of $m_1 = 1$. However, since m_1 is the lower bound of the support for the distribution, this results in degree values that are outside of the support which will cause T, and consequently the random variable $Z = 2\beta_0 T$, to be biased. Therefore, the UIT on this set of m_1 needs to be performed on the truncated degree distribution

which essentially reduces the original network to a "sub-network" that is associated with the truncated degree distribution. This sub-network contains the main "hubs" of the original network that connect the entire network. In essence, it becomes a test of the network remaining central nodes. Applying the UIT to the sub-networks resulted in the *Les Miserables* and *Dolphin Social Network* to not differ significantly from a Barabási-Albert network with $m_1 = 6$ and $m_1 = 4$, respectively, each having a z-statistic greater than -1.96 (-1.024 and -1.212, respectively (Table 11)).

However, by removing the periphery nodes, the resulting sub-networks only account for 53% and 66% of the original *Les Miserables* and *Dolphin Social Network* networks, respectively. As shown in Figure 17, a substantial proportion of the degree distribution is spread out further to the right of the degree axis than is expected from a $Pareto(m_1 = 1, \beta_0 = 2)$ distribution (red). Note that when comparing the truncated distributions to their respective Pareto distributions (blue), the degrees seem to follow characteristics of the Pareto distribution much better. This result suggests that even though a network might be significantly different than a Barabási-Albert network based on its degree distribution, the sub-network, which contains the central hubs of the original network, could still possess characteristics similar to the Barabási-Albert network and be used for visualization and study.

Table 11. z-statistic and $\hat{\beta}_{MLE}$ for real world networks.

Network	N	$z\|m_1 = 1^*$	$\hat{\beta}_{MLE}$	$z_{best}\|m_1{}^\dagger$	sub-net %
Karate Club	34	25.25	0.7809	7.994|2	97
Les Miserables	77	43.55	0.6911	**-1.024|6**	53
Dolphin Social Network	62	38.07	0.7085	**-1.212|4**	66
Political Blogs	1224	322.67	0.4147	300.90|6	68
Cond Matter Collaborations	39577	1131.38	0.6022	274.76|6	14
Astrophysics Collaborations	16046	932.11	0.4948	520.10|6	57
High-Energy Physics Theory Collaborations	7610	251.54	0.9481	-16.27|6	22
High-Energy Physics Theory Citations	27770	1647.09	0.3896	1059.30|6	76
High-Energy Physics Phenomenology Citations	34546	1904.60	0.3780	1151.31|6	81
Internet	22963	185.33	1.4351	-70.21|2	66
Google Webgraphs	875713	5721.48	0.5703	2427.05|6	52
Facebook Social Circles	4039	807.92	0.3153	649.35|6	89

*All p-values are < 0.0001.
†: **Bold** indicates not significantly different from Barabási-Albert network. $z|m_1 = 1$ is the z-score when m_1 is assumed to be 1, and $z_{best}|m_1$ is the smallest absolute z-score given the associated m_1 assumption.

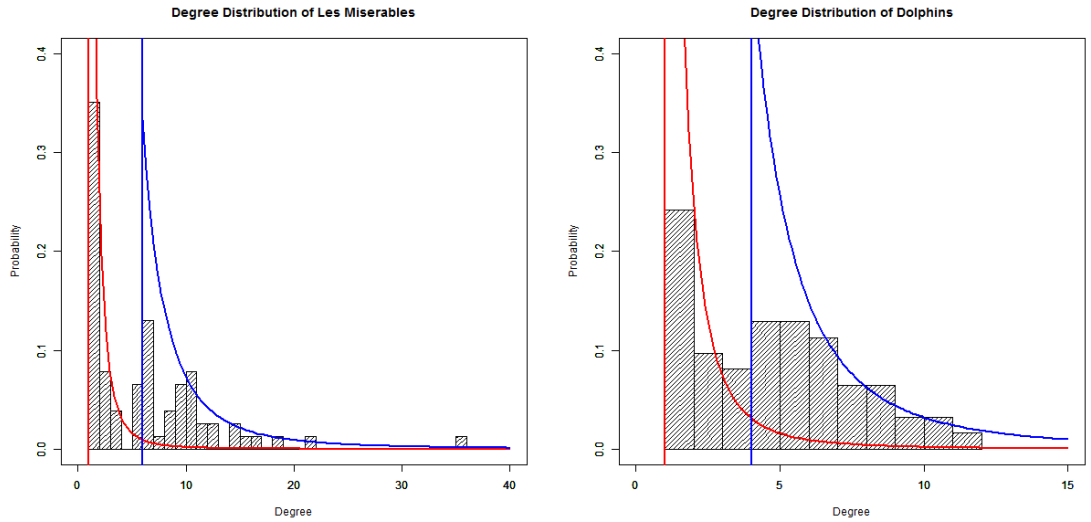


Figure 17. Historgram of degree distributions for Les Miserables and Dolphins and compared to the Barabási-Albert distribution with $m_1 = 1$ (red) and distribution with $m_1 = 6$ and $m_1 = 4$ (blue) for left and right, respectively.

IV. Network Degradation Detection

The second primary objective of this research is to create a test of hypothesis to detect subtle degradation within a Barabási-Albert network attributable to nodal or edge deletion. For this research, degradation is defined as removal of the nodes or edges within the network that changes the structure of the network and its degree distribution. To achieve this objective, L-moments from the degree distribution of the Barabási-Albert network are used for creating a multivariate test of hypothesis. This is then followed by an analysis on the sensitivity of the test to changes within the network based on proportion of edge and nodal deletion to investigate how quickly the test rejects a degrading Barabási-Albert network.

4.1 Empirical Distribution for the L-moments of the Barabási-Albert Degree Distribution

In order to utilize the L-moments as statistical measures for identifying or characterizing a network, their distributions with respect to degree distribution first need to be identified, so that a statistical test of hypothesis can be created. This work examines a nonparametric approach to deriving the theoretical distribution of the L-moments for the Barabási-Albert degree distribution. Due to the graph theoretic property of the Barabási-Albert network, the mean degree for a fixed (m, k) pair is also fixed since the mean degree is a function of the total degree which was shown to be deterministic in Section 2.2.1. Therefore only the empirical distributions of the of the L-scale (λ_2), L-skewness (τ_3), and L-kurtosis (τ_4) of the Barabási-Albert degree distribution are simulated since Hosking [40] suggested that a distribution can be well characterized by including up to the fourth L-moment. The simulation was conducted as followed:

1. Bootstrap each L-moment distribution from 1000 randomly generated Barabási-Albert graphs for each $m \in \{2, 4, 6\}$ and $n \in \{2^k : k = 5, 6, \ldots, 15\}$ parameter combination.

2. Test each distribution for normality using the Shapiro-Wilk and Anderson-Darling tests for univariate normality of the marginals, and using the Royston H test [80] which is a multivariate extension on the Shapiro-Wilk for multivariate normality of the joint distributions.

3. Repeat step 1-2 100 times to obtain the proportion of instances where the L-moment distributions are no different than a normal distribution.

4. Compare results of step 3 to that expected at $\alpha = .05$ level.

Hosking [40] also suggested considering the pairwise combination of L-skewness and L-kurtosis of a distribution in characterizing the distribution from which the L-moments came. The expected λ_2, τ_3 and τ_4 of the Generalized Pareto distribution were derived by Hosking [40] and the expected τ_3 and τ_4 relationship is shown in Figure 18 along with the pairwise empirical L-skewness and L-kurtosis of the Barabási-Albert network. Also included in Figure 18 is the empirical distribution of the L-skewness and L-kurtosis from the $Pareto(m, 2)$ distribution where the sample size corresponds to the network sizes of the Barabási-Albert sampling. Overall, although the L-moments of the Barabási-Albert degree distribution converge towards the expected (τ_3, τ_4) values of the Pareto distribution (point labeled "2" on the Generalized Pareto [Gen Pareto] line in Figure 18), the distributions of the L-moments themselves do not lie on the line of expected values for the Pareto distribution. Another observation is that the separation of the L-moments distribution between the different network sizes becomes more prominent as m increases.

One can then argue that the empirical degree distribution of the Barabási-Albert

graph is different than what was theoretically derived by Barabási and Albert [6] at least for relatively small graphs. Further, it is also apparent that the (τ_3, τ_4) pair may have a bivariate distribution which resembles that of a normal distribution (Figure 18). Therefore, the marginal as well as the joint distributions distributions of the L-moments are tested for normality using the Shapiro-Wilk and Anderson-Darling test for the marginals and the Royston H test for the joint distributions.

Based on the results, it was shown that none of the distributions of λ_2 and τ_4 differ significantly from the normal distribution although it appears that the distributions for τ_3 when $k \leq 8$ are significantly different from the normal distribution (Table 12). The histogram of the L-moments for $k \in \{6, 15\}$ are shown in Figures 19 and 20 where it appears that the normal distribution overlaps the empirical L-moment distributions quite fittingly even for τ_3. The multivariate normality test on the joint distributions suggests that the non-normality of τ_3 seems to affect whether or not the joint distributions are significantly different from the multivariate normal (Table 13). Specifically, it seems that when τ_3 is part of the joint distribution and $k \leq 6$, said joint distribution is significantly different from the multivariate normal more often than when τ_3 is not part of the joint distribution. Nevertheless, the proportion of the sample where the univariate and multivariate distributions involving τ_3 for $k \geq 6$ is not significantly different than normal is in the large majority ($\gg 0.5$), thus it will be included in the investigation for selecting the appropriate L-moments for the test of hypothesis.

All 100 empirical distributions of each L-moment are then compiled together into one distribution of 10^5 random samples to obtain a larger bootstrapped L-moment distribution. From the 10^5 random sample, the mean and covariance structure between the L-moments are estimated which are listed in Tables B.4 and B.5. The estimates are then used to standardize the L-moments and to transform them into the mul-

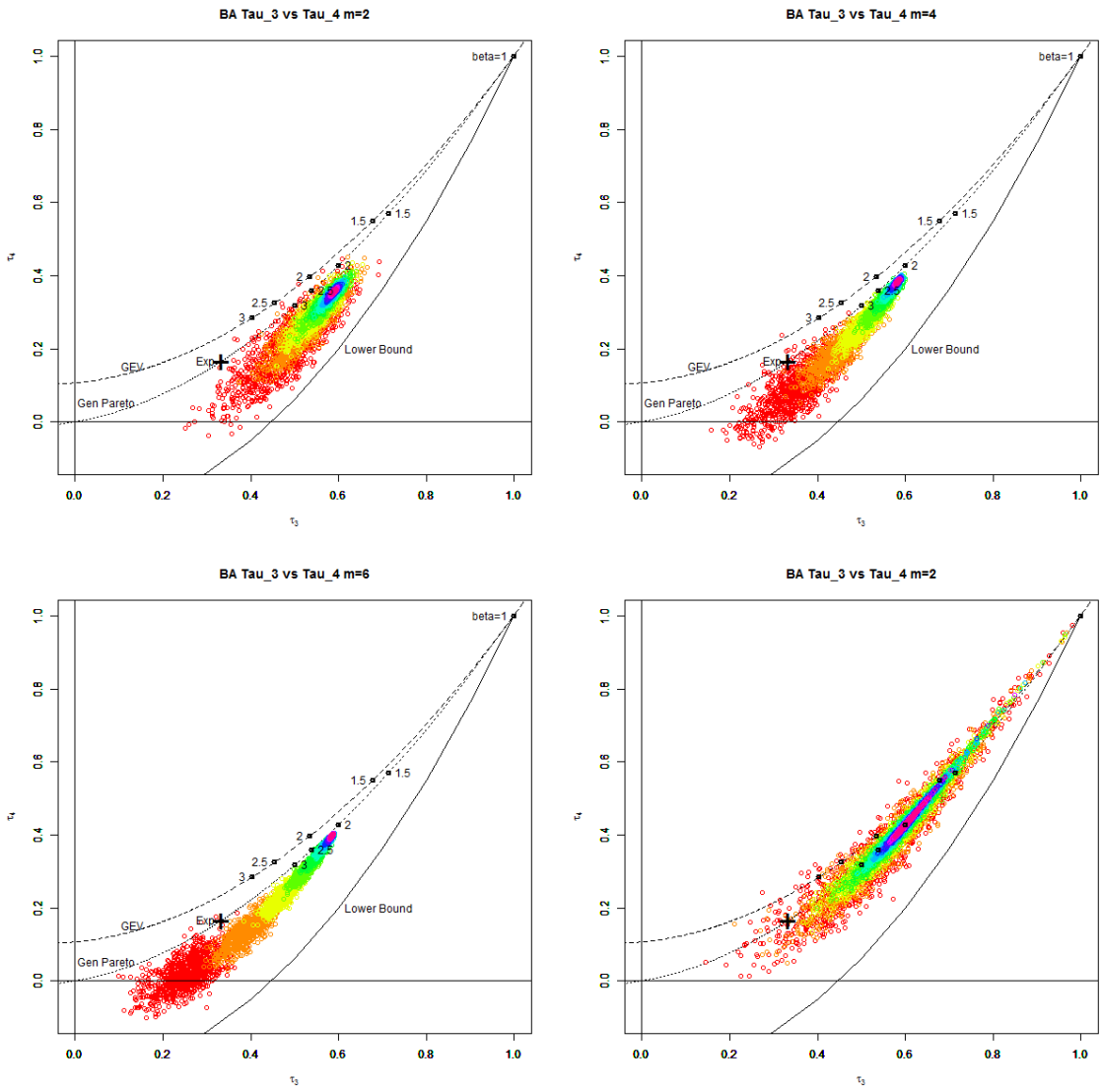


Figure 18. Plot of τ_3 vs τ_4 for $m = 2, 4, 6$ of the Barabási-Albert (BA) degree distribution and the Pareto distribution. (Note: Colors distinguish $k = 5$ (large spread) to $k = 15$ (small spread). Points are expected (τ_3, τ_4) values for the associated β values for the Generalized Pareto (Gen Pareto) and Generalized Extreme Value (GEV) distributions. Exp is the expected (τ_3, τ_4) for the Exponential distribution.)

tivariate standard normal distribution which minimizes the correlation between the L-moments. The transformed multivariate normal distribution will be used to create a statistical test of hypothesis on the values of these L-moments to be used to differentiate between various Barabási-Albert networks.

Table 12. Proportion where distribution of L-moments are not significantly different from the normal distribution

		λ_2		τ_3		τ_4	
k	m	**SW**	**AD**	**SW**	**AD**	**SW**	**AD**
5	2	0.93	0.97	0.40	0.58	0.89	0.88
5	4	0.92	0.94	0.43	0.54	0.83	0.94
5	6	0.96	0.96	0.65	0.77	0.84	0.87
6	2	0.90	0.97	0.66	0.80	0.90	0.91
6	4	0.94	0.92	0.62	0.78	0.93	0.92
6	6	0.91	0.97	0.7	0.78	0.89	0.91
7	2	0.90	0.92	0.74	0.85	0.89	0.95
7	4	0.99	1.00	0.83	0.88	0.96	0.99
7	6	0.89	0.95	0.84	0.86	0.91	0.92
8	2	0.94	0.95	0.92	0.94	0.94	0.98
8	4	0.92	0.94	0.87	0.89	0.92	0.93
8	6	0.97	0.98	0.88	0.92	0.95	1.00
9	2	0.92	0.93	0.87	0.87	0.96	0.98
9	4	0.94	0.94	0.91	0.97	0.93	0.97
9	6	0.97	0.97	0.94	0.97	0.98	0.96
10	2	0.95	0.97	0.95	0.98	0.97	0.94
10	4	0.94	0.97	0.95	0.95	0.95	0.97
10	6	0.98	0.96	0.94	0.96	0.94	0.96
11	2	0.94	0.97	0.97	0.96	0.94	0.92
11	4	0.97	0.95	0.92	0.93	0.95	0.98
11	6	0.94	0.93	0.90	0.93	0.95	0.97
12	2	0.94	0.98	0.94	0.96	0.98	0.94
12	4	0.96	0.93	0.94	0.94	0.96	0.97
12	6	0.95	0.96	0.97	0.98	0.94	0.93

SW: Shapiro-Wilks, **AD**: Anderson-Darling tests for normality.

4.2 Multivariate Standard Normal Distribution in Polar Coordinates

A transformation of the bivariate and trivariate normal distribution from the Cartesian coordinate to the polar coordinate will now be demonstrated. This is performed in order to transform the multivariate L-moments collection into a single value (radius) that can be used as a criteria for the test of hypothesis. Let X and Y be independent and identically distributed (iid) $Normal(0, 1)$ and let $R = \sqrt{X^2 + Y^2}$.

Table 13. Proportion where multivariate distribution of L-moments are not significantly different from the multivariate normal distribution based on the Royston H-test

$\boldsymbol{k}$	$\boldsymbol{m}$	λ_2, τ_3	λ_2, τ_4	τ_3, τ_4	$\lambda_2, \tau_3, \tau_4$
5	2	0.48	0.69	0.49	0.49
5	4	0.49	0.86	0.48	0.57
5	6	0.64	0.81	0.58	0.65
6	2	0.76	0.87	0.83	0.82
6	4	0.61	0.85	0.7	0.68
6	6	0.80	0.94	0.79	0.83
7	2	0.84	0.91	0.88	0.87
7	4	0.89	0.91	0.91	0.90
7	6	0.87	0.95	0.89	0.88
8	2	0.91	0.93	0.94	0.91
8	4	0.93	0.93	0.92	0.92
8	6	0.85	0.89	0.89	0.87
9	2	0.92	0.97	0.93	0.92
9	4	0.94	0.92	0.95	0.94
9	6	0.88	0.92	0.91	0.89
10	2	0.93	0.91	0.91	0.90
10	4	0.92	0.91	0.92	0.88
10	6	0.95	0.94	0.96	0.94
11	2	0.87	0.91	0.89	0.90
11	4	0.95	0.93	0.94	0.94
11	6	0.91	0.89	0.95	0.90
12	2	0.98	0.96	0.95	0.96
12	4	0.92	0.90	0.93	0.92
12	6	0.98	0.97	0.96	0.97

Now, find the value of c such that $P(R > \sqrt{c}) = \alpha$. The joint density of (X, Y) is

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\ &= \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}. \end{aligned}$$

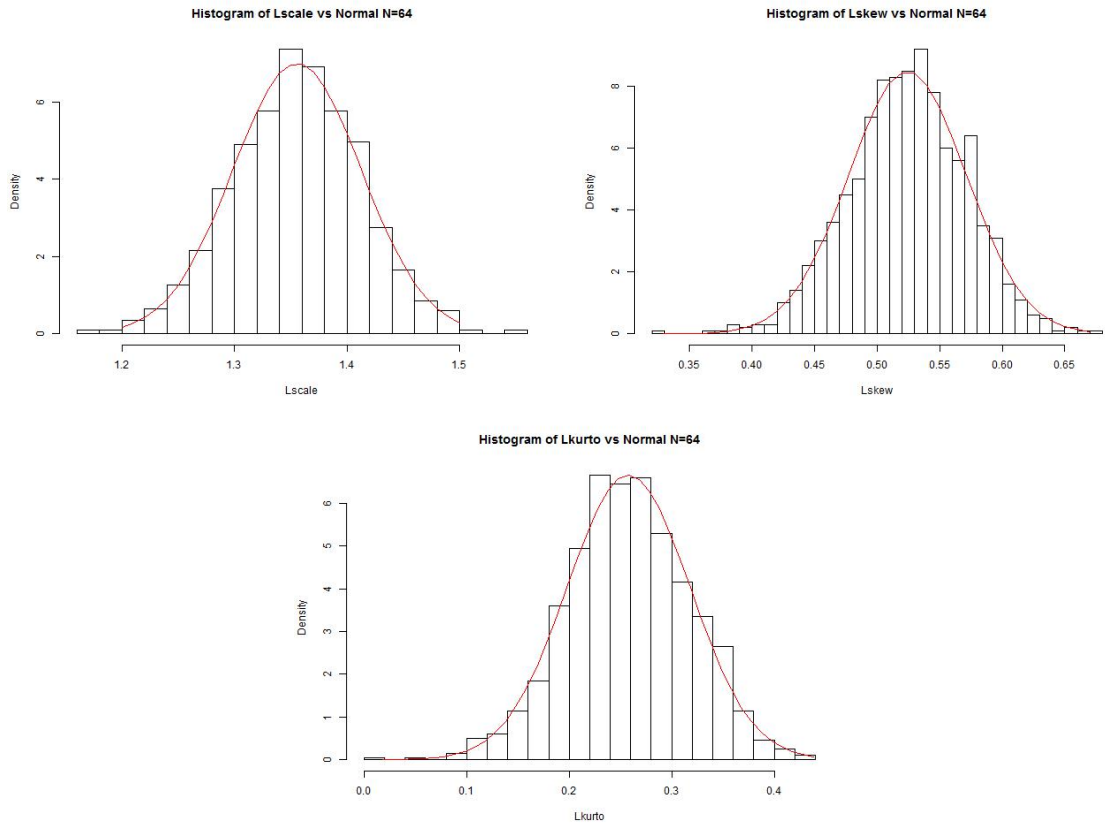


Figure 19. Example histograms of L-moments for $m = 2$ $k = 6$

Let $X = R\cos\Theta$ and $Y = R\sin\Theta$. Then,

$$\begin{aligned} X^2 + Y^2 &= (R\cos\Theta)^2 + (R\sin\Theta)^2 \\ &= R^2\cos^2\Theta + R^2\sin^2\Theta \\ &= R^2(\cos^2\Theta + \sin^2\Theta) \\ X^2 + Y^2 &= R^2. \end{aligned}$$

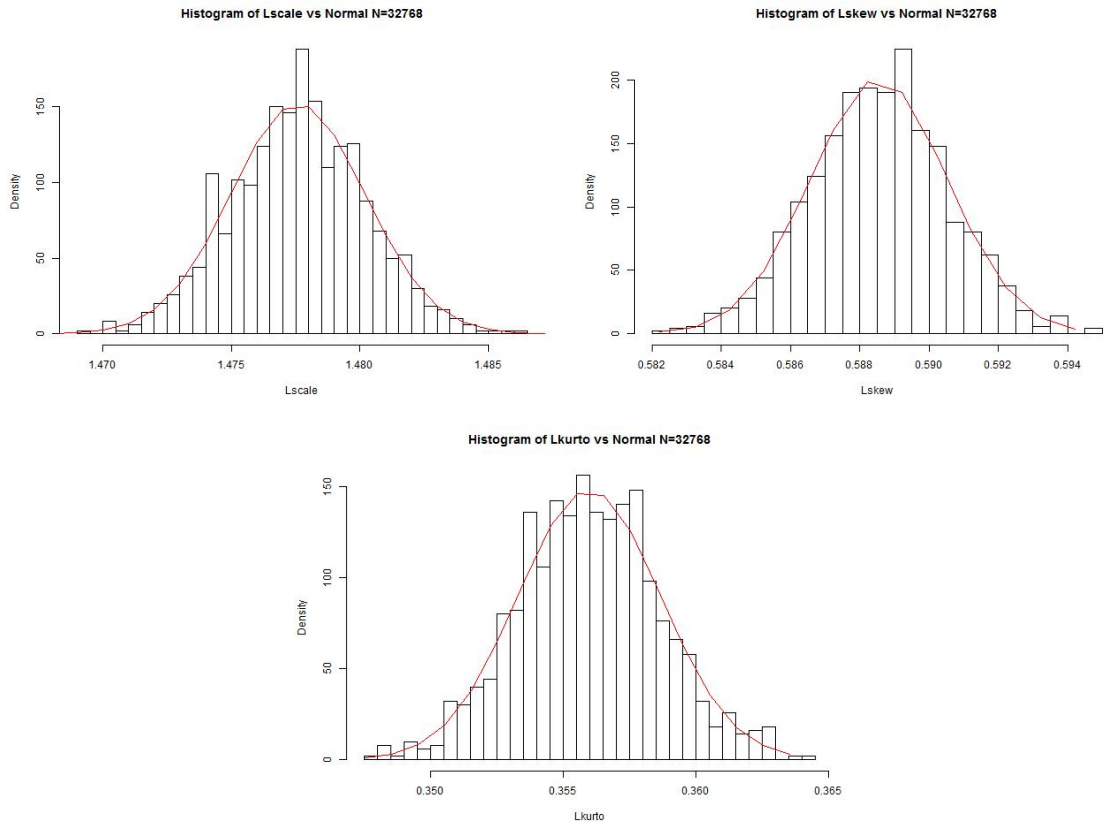


Figure 20. Example histograms of L-moments for $m = 2\ k = 15$

Consequently, the Jacobian is

$$
\begin{aligned}
J &= \begin{vmatrix} \frac{\partial}{\partial r} r\cos\theta & \frac{\partial}{\partial \theta} r\cos\theta \\ \frac{\partial}{\partial r} r\sin\theta & \frac{\partial}{\partial \theta} r\sin\theta \end{vmatrix} \\
&= \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} \\
&= r\cos^2\theta + r\sin^2\theta \\
J &= r.
\end{aligned}
$$

Therefore the joint density of (R, Θ) is

$$f_{R,\Theta}(r,\theta) = \frac{r}{2\pi} e^{-\frac{r^2}{2}}$$

where $r \in [0, \infty)$ and $\theta \in [0, 2\pi)$. Thus, the value of c such that the probability of the marginal of R, $P(R > \sqrt{c}) = \alpha$, is

$$\begin{aligned} 1 - \alpha &= P(R \leq \sqrt{c}) \\ &= \int_0^{\sqrt{c}} \int_0^{2\pi} \frac{r}{2\pi} e^{-\frac{r^2}{2}} d\theta dr \\ 1 - \alpha &= \int_0^{\sqrt{c}} r e^{-\frac{r^2}{2}} dr. \end{aligned}$$

Letting $s = -r^2/2$, then $ds = -rdr$ and

$$\begin{aligned} 1 - \alpha &= \int_{-c/2}^{0} e^s ds \\ &= 1 - e^{-c/2} \\ \alpha &= e^{-c/2} \\ \ln \alpha &= -c/2 \\ c &= -2 \ln \alpha. \end{aligned}$$

Extending the previous derivation to include a third variable Z that is also iid $Normal(0, 1)$, the joint density of (X, Y, Z) then becomes

$$\begin{aligned} f_{X,Y,Z}(x,y,z) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \\ &= \frac{1}{(2\pi)^{3/2}} e^{-\frac{x^2+y^2+z^2}{2}}. \end{aligned}$$

However, now let $X = R\sin\Theta\cos\Phi$, $Y = R\sin\Theta\sin\Phi$, and $Z = R\cos\Theta$ which gives

$$\begin{aligned}
X^2 + Y^2 + Z^2 &= (R\sin\Theta\cos\Phi)^2 + (R\sin\Theta\sin\Phi)^2 + (R\cos\Theta)^2 \\
&= R^2\sin^2\Theta\cos^2\Phi + R^2\sin^2\Theta\sin^2\Phi + \cos^2\Theta \\
&= R^2(\sin^2\Theta\cos^2\Phi + \sin^2\Theta\sin^2\Phi + \cos^2\Theta) \\
X^2 + Y^2 + Z^2 &= R^2,
\end{aligned}$$

and the Jacobian,

$$\begin{aligned}
J &= \begin{vmatrix}
\frac{\partial}{\partial r}r\sin\theta\cos\phi & \frac{\partial}{\partial\theta}r\sin\theta\cos\phi & \frac{\partial}{\partial\phi}r\sin\theta\cos\phi \\
\frac{\partial}{\partial r}r\sin\theta\sin\phi & \frac{\partial}{\partial\theta}r\sin\theta\sin\phi & \frac{\partial}{\partial\phi}r\sin\theta\sin\phi \\
\frac{\partial}{\partial r}r\cos\theta & \frac{\partial}{\partial\theta}r\cos\theta & \frac{\partial}{\partial\phi}r\cos\theta
\end{vmatrix} \\
&= \begin{vmatrix}
\sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\
\sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi \\
\cos\theta & -r\sin\theta & 0
\end{vmatrix} \\
&= 0 + r^2\cos^2\theta\sin\theta\cos^2\phi + r^2\sin^3\theta\sin^2\phi \\
&\quad + r^2\cos^2\theta\sin\theta\sin^2\phi - 0 + r^2\sin^3\theta\cos^2\phi \\
&= r^2\sin\theta(\cos^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi + \cos^2\theta\sin^2\phi + \sin^2\theta\cos^2\phi) \\
&= r^2\sin\theta(\cos^2\theta + \sin^2\theta)(\cos^2\phi + \sin^2\phi) \\
J &= r^2\sin\theta.
\end{aligned}$$

Hence, the joint density of (R, Θ, Φ) is

$$f_{R,\Theta,\Phi}(r,\theta,\phi) = \frac{r^2\sin\theta}{(2\pi)^{3/2}}e^{-\frac{r^2}{2}}$$

where $r \in [0,\infty)$, $\theta \in [0,\pi]$, and $\phi \in [0,2\pi)$. The value of c for the marginal of R

such that $P(R > \sqrt{c}) = \alpha$ is

$$
\begin{aligned}
1 - \alpha &= P(R \leq \sqrt{c}) \\
&= \int_0^{\sqrt{c}} \int_0^{\pi} \int_0^{2\pi} \frac{r^2 \sin\theta}{(2\pi)^{3/2}} e^{-\frac{r^2}{2}} d\phi d\theta dr \\
&= \int_0^{\sqrt{c}} \int_0^{\pi} \frac{r^2 \sin\theta}{\sqrt{2\pi}} e^{-\frac{r^2}{2}} d\theta dr \\
&= \int_0^{\sqrt{c}} \frac{r^2}{\sqrt{2\pi}} e^{-\frac{r^2}{2}} \int_0^{\pi} \sin\theta d\theta dr \\
&= \int_0^{\sqrt{c}} \frac{r^2}{\sqrt{2\pi}} e^{-\frac{r^2}{2}} \left(-\cos\theta|_{\theta=0}^{\pi}\right. dr \\
&= \int_0^{\sqrt{c}} \frac{2r^2}{\sqrt{2\pi}} e^{-\frac{r^2}{2}} dr \\
&= \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{c}} r^2 e^{-\frac{r^2}{2}} dr \\
&= \sqrt{\frac{2}{\pi}} \left(\frac{1}{4} \sqrt{\frac{\pi}{(1/2)^3}} \operatorname{erf}\left(r\sqrt{1/2}\right) - \frac{r}{2(1/2)} e^{-r^2/2} \right|_{r=0}^{\sqrt{c}} \\
&= \sqrt{\frac{2}{\pi}} \left(\frac{1}{4} \sqrt{8\pi} \operatorname{erf}\left(\sqrt{\frac{c}{2}}\right) - \sqrt{c} e^{-c/2} \right) \\
1 - \alpha &= \operatorname{erf}\left(\sqrt{\frac{c}{2}}\right) - \sqrt{\frac{2c}{\pi}} e^{-c/2} \\
\alpha &= 1 - \operatorname{erf}\left(\sqrt{\frac{c}{2}}\right) + \sqrt{\frac{2c}{\pi}} e^{-c/2}.
\end{aligned}
$$

Therefore, to obtain the value of c for a particular α, one must numerically find the closest c that satisfies $P(R > \sqrt{c}) = \alpha$. Table 14 lists the values of c for select α values. From these result, one can now simultaneously test pairwise and triple combinations of standardized L-moments against their respective multivariate standard normal distribution.

Table 14. Values of c such that $P(R > \sqrt{c}) = \alpha$ for the multivariate normal distribution

α	**c**
0.010	11.3449
0.025	9.3484
0.050	7.8147
0.100	6.2514
0.150	5.3171
0.200	4.6416

4.3 Tests on Degree L-moments for the Barabási-Albert Network

Three tests using the L-moments of the degree distribution of a Barabási-Albert network are proposed that are based on the Standard Normal univariate distribution as well as the Multivariate Standard Normal distribution as described in Section 4.2. The tests are built upon the hypothesis $H_0 : \underset{\sim}{\lambda} \in \mathbf{\Lambda}_{(m,k)}$ vs $H_A : \underset{\sim}{\lambda} \notin \mathbf{\Lambda}_{(m,k)}$ where $\underset{\sim}{\lambda}$ is the collection of L-moment estimates for a given network and $\mathbf{\Lambda}_{(m,k)}$ is the empirical distribution of the L-moments with a mean and covariance structure $(\underset{\sim}{\mu}, \Sigma)$ for a Barabási-Albert graph of size k with parameter m as estimated in Section 4.1 and listed in Tables B.4 and B.5. Consider the standardized $\underset{\sim}{\lambda}$ as $\underset{\sim}{t} = \Sigma^{-\frac{1}{2}}(\underset{\sim}{\lambda} - \underset{\sim}{\mu})'$, and define the test statistic

$$S = \|\underset{\sim}{t}\|_2.$$

Thus, for a given α level for the TypeI error, a test of $H_0 : \underset{\sim}{\lambda} \in \mathbf{\Lambda}_{(m,k)}$ vs $H_A : \underset{\sim}{\lambda} \notin \mathbf{\Lambda}_{(m,k)}$ will reject H_0 if

$$S \geq \sqrt{c} \tag{23}$$

where c is defined in Section 4.2 and provided for specific α values in Table 14. Note that for the univariate case, the rejection criterion is equivalent to the univariate

standard normal distribution, where H_0 is rejected if

$$S \leq z_{\alpha/2} \quad \text{or} \quad S \geq z_{1-\alpha/2}. \tag{24}$$

4.4 Power of the Tests on Degree L-moments

In order to develop a test that is able to detect degradation within the degree distribution of a Barabási-Albert network, it is a necessary condition for the test to first be able to correctly classify the networks with high power. Recall that the test of normality on the L-moments in Section 4.1 showed that the distribution of L-scale (λ_2) was no different than a normal distribution in roughly 90% of the samples or higher. However, the opposite could be said for L-skewness (τ_3) where the distribution of τ_3 was shown to be no different than a normal distribution for only a small proportion of the samples when $k \leq 8$ despite the joint distributions of the L-moments not being significantly different from the multivariate normal for smaller k. Therefore, the power of the test based on Equations (23) and (24) using L-scale (λ_2), L-skewness (τ_3), and L-kurtosis (τ_4), as well as their bivariate and trivariate distributions is investigated.

For each m and k combination, a Barabási-Albert network is generated and assigned as the *target* network. Then its degree distribution, λ_2, τ_3, and τ_4 are computed and compared to that of the estimated distributions from Section 4.1 which are designated as the *class* networks. If the statistic for the *target*, as defined in Section 4.3, falls within the rejection region, then the network will be rejected from being assigned to the *class* network. These steps are outlined in Algorithm 1. The false negative and true negative counts are aggregated to compute the TypeI error and power of the tests for each $(target, k)$ pair, respectively.

The simulation shows that the tests with rejection region from Equations (23) and (24) maintained the appropriate $\alpha = 0.05$ level for all m (Tables 15 to 18).

Algorithm 1 L-moments classification algorithm

```
1: procedure CLASSIFY
2:     for each target ∈ {1, 2, ..., 7} do
3:         for each class ∈ {1, 2, ..., 7} do
4:             for each k ∈ {5, 6, ..., 14} do
5:                 for each boot ∈ {1, ..., 1000} do
6:                     g ← Generate a Barabási-Albert network with m = target
                           and random seed boot
7:                     d ← Compute degree distribution of g
8:                     lmom_{target,k} ← Compute (λ_2, τ_3, τ_4) of d
9:                     normlm ← Normalize lmom_{target,k} with (μ̃_{class,k}, Σ_{class,k})
                                from Tables B.4 and B.5
10:                    R ← statistic based on dimension of normlm
11:                    c ← critical value based on alpha and dimension of normlm
12:                    if target = class then
13:                        if R > c then
14:                            return FALSE NEGATIVE
15:                        else
16:                            return TRUE POSITIVE
17:                        end if
18:                    else if target ≠ class then
19:                        if R > c then
20:                            return TRUE NEGATIVE
21:                        else
22:                            return FALSE POSITIVE
23:                        end if
24:                    end if
25:                end for
26:            end for
27:        end for
28:    end for
29: end procedure
```

The test is also shown to be quite powerful when only considering λ_2 as a statistic for classification, where it is able to correctly reject when $target \neq class$ with a probability of one for $k \geq 7$. However, the power of the test for $k \leq 6$ could be improved upon by the multivariate addition of τ_3 and τ_4 as shown in Tables 15 to 18, despite the test being more prone to misclassification for $m \geq 4$ when $k = 5$. Nevertheless, the trivariate test on $(\lambda_2, \tau_3, \tau_4)$ was able to improve the power of the test for these values of m and k ($m \geq 4$ and $k = 5$).

One implication that can be made from the network classification result of using the L-moments is that the tests are, in essence, tests for detecting changes within the degree distribution with respect to m. Thus, should the network behave in such a way to cause its degree distribution to deviate from the initial network with a particular minimum degree (m), then the trivariate test is especially powerful in detecting such a change, even for smaller networks. However, it is worth noting that the change in m (the minimum degree) is discrete and has very low resolution. Therefore, the performance of the test needs to be evaluated with respect to a more sensitive change within the network. Hence, the trivariate test will be chosen as the one to use for change detection since it is very sensitive to the perturbation within the degree distribution of the Barabási-Albert network due to the combination of the L-moments. However, the univariate test on λ_2 is also used as a baseline for comparison since there are no concerns with non-normality for λ_2, and a test on λ_2 may be considered parsimonious.

4.5 Sensitivity Analysis of Edge and Node Deletion

For detecting how sensitive the test described in Section 4.4 is with respect to nodal or edge deletion within the Barabási-Albert network, the network is degraded at three levels of nodal degrees: 1) nodes with minimum degree m (low degree), 2) nodes with

Table 15. Power of the test using only λ_2

$k = 5$	*Class*						
Target	1	2	3	4	5	6	7
1	0.049	1	1	1	1	1	1
2	1	0.054	1	1	1	1	1
3	1	0.997	0.050	0.927	1	1	1
4	1	1	0.929	0.048	0.680	0.988	1
5	1	1	1	0.702	0.047	0.394	0.846
6	1	1	1	0.985	0.468	0.042	0.234
7	1	1	1	0.999	0.856	0.226	0.049

$k = 6$	*Class*						
Target	1	2	3	4	5	6	7
1	0.047	1	1	1	1	1	1
2	1	0.037	1	1	1	1	1
3	1	1	0.041	1	1	1	1
4	1	1	1	0.048	0.999	1	1
5	1	1	1	0.999	0.045	0.980	1
6	1	1	1	1	0.975	0.058	0.947
7	1	1	1	1	1	0.935	0.056

Table 16. Power of the test using (λ_2, τ_3) jointly

$k = 5$	*Class*						
Target	1	2	3	4	5	6	7
1	0.062	1	1	1	1	1	1
2	1	0.06	1	1	1	1	1
3	1	1	0.037	1	1	1	1
4	1	1	1	0.054	0.999	1	1
5	1	1	1	0.994	0.058	0.951	1
6	1	1	1	1	0.926	0.05	0.779
7	1	1	1	1	1	0.761	0.045

$k = 6$	*Class*						
Target	1	2	3	4	5	6	7
1	0.046	1	1	1	1	1	1
2	1	0.048	1	1	1	1	1
3	1	1	0.038	1	1	1	1
4	1	1	1	0.044	1	1	1
5	1	1	1	1	0.047	1	1
6	1	1	1	1	1	0.047	1
7	1	1	1	1	1	1	0.056

degree equal to the median degree (medium degree), and 3) nodes with degree in the top 1% of the network (high degree). For each level of degree, an investigation on how

Table 17. Power of the test using (λ_2, τ_4) jointly

$k = 5$	$Class$						
$Target$	1	2	3	4	5	6	7
1	0.049	1	1	1	1	1	1
2	1	0.048	1	1	1	1	1
3	1	0.999	0.049	0.958	1	1	1
4	1	1	0.949	0.061	0.661	0.995	1
5	1	1	1	0.734	0.064	0.399	0.89
6	1	1	1	0.99	0.425	0.056	0.248
7	1	1	1	1	0.859	0.218	0.044

$k = 6$	$Class$						
$Target$	1	2	3	4	5	6	7
1	0.049	1	1	1	1	1	1
2	1	0.045	1	1	1	1	1
3	1	1	0.04	1	1	1	1
4	1	1	1	0.049	1	1	1
5	1	1	1	1	0.046	0.99	1
6	1	1	1	1	0.981	0.048	0.964
7	1	1	1	1	1	0.948	0.055

Table 18. Power of the test using $(\lambda_2, \tau_3, \tau_4)$ jointly

$k = 5$	$Class$						
$Target$	1	2	3	4	5	6	7
1	0.077	1	1	1	1	1	1
2	1	0.069	1	1	1	1	1
3	1	1	0.054	1	1	1	1
4	1	1	1	0.068	1	1	1
5	1	1	1	1	0.067	1	1
6	1	1	1	1	0.997	0.049	0.989
7	1	1	1	1	1	0.948	0.051

$k = 6$	$Class$						
$Target$	1	2	3	4	5	6	7
1	0.065	1	1	1	1	1	1
2	1	0.053	1	1	1	1	1
3	1	1	0.052	1	1	1	1
4	1	1	1	0.056	1	1	1
5	1	1	1	1	0.058	1	1
6	1	1	1	1	1	0.07	1
7	1	1	1	1	1	1	0.048

the test reacts to both edge deletion and node deletion by varying the proportion of deletion $p \in \{0.01, 0.02, \ldots, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5\}$ is conducted. The algorithms

for both methods of deletion are outlined in Algorithms 2 and 3. Algorithms 2 and 3 are then used within Algorithm 4 to compute the power of the test as a function of p for each (m, k) combination. In essence, the power of the test is computed for each $target$ by deleting edges or nodes with the appropriate p and seeing how well the tests with rejection region from Equation (23) rejects the $target$ from each $class$. Note that for node deletion, deleting a particular node has the consequence of deleting all edges connected to said node. However, the resulting number of edges deleted by the two different processes are equivalent for low and medium degree nodes. Suppose that only nodes with degree equal to d are affected and suppose that the probability of deletion is p. Let ν be the number of nodes with degree equal to d. Thus, for edge deletion, the number of edges affected is $e_v = \nu d$ and the number of edges deleted is $e_v p = \nu dp$. Similarly, the number of nodes deleted for node deletion is $v_1 = \nu p$ and the number of edges affected is $v_1 d = \nu pd$. Therefore, the difference between edge deletion and node deletion then becomes the subtle distinction of whether the edges are deleted randomly out of all affected edges or whether the edges are deleted randomly albeit in a more concentrated fashion as a function of node selection. However, the number of edges deleted becomes more stochastic for high degree level since the degrees of the affected nodes vary.

4.5.1 Edge Deletion.

Characteristics of Edge Deletion.

The characteristics of the network in terms of resulting isolates, components, and the network's clustering coefficient with respect to the deletion process are first investigated. Isolates are any nodes with degree zero resulting from the deletion process, and components are disconnected subgraphs that are themselves connected which resulted from the deletion process. It should be noted that when the minimum

Algorithm 2 Edge deletion algorithm

procedure EDGEDELETION
    $g \leftarrow$ Barabási-Albert network
    $degree \leftarrow$ degree distribution of g
    $delType \leftarrow$ low, medium, or high degree levels
    $p \leftarrow$ proportion of deletion
    if $delType =$ low **then**
        $v \leftarrow$ all nodes with $degree = \min\{degree\}$
    else if $delType =$ medium **then**
        $v \leftarrow$ all nodes with $degree = \text{median}\{degree\}$
    else if $delType =$ high **then**
        $v \leftarrow$ all nodes with $degree \geq 99^{th}$ percentile of degree distribution
    end if
    $e_v \leftarrow$ all edges connected to v
    $e \leftarrow$ randomly selected e_v with probability p
    remove e from g
    remove any isolates (i.e. nodes with degree zero) from g
end procedure

Algorithm 3 Node deletion algorithm

procedure NODEDELETION
    $g \leftarrow$ Barabási-Albert network
    $degree \leftarrow$ degree distribution of g
    $delType \leftarrow$ low, medium, or high degree levels
    $p \leftarrow$ proportion of deletion
    if $delType =$ low **then**
        $v \leftarrow$ all nodes with $degree = \min\{degree\}$
    else if $delType =$ medium **then**
        $v \leftarrow$ all nodes with $degree = \text{median}\{degree\}$
    else if $delType =$ high **then**
        $v \leftarrow$ all nodes with $degree \geq 99^{th}$ percentile of degree distribution
    end if
    $v_1 \leftarrow$ randomly selected v with probability p
    remove v_1 from g
    remove any isolates (i.e. nodes with degree zero) from g
end procedure

Algorithm 4 L-moments change detection algorithm

```
1: procedure DetectChange
2:     delType ← degree levels low, medium, or high
3:     for each m ∈ {1, 2, ..., 7} do
4:         for each k ∈ {5, 6, ..., 14} do
5:             for each p ∈ {0.01, 0.02, ..., 0.1, 0.15, 0.2, 0.3, 0.4, 0.5} do
6:                 for each boot ∈ {1, ..., 1000} do
7:                     g ← Generate a Barabási-Albert network with parameter m
                           of size k and random seed boot
8:                     g_1 ← Alter g with Algorithm 2 or Algorithm 3 using delType
                           and p
9:                     d ← Compute degree distribution of g_1
10:                    lmom_{m,k} ← Compute (λ_2, τ_3, τ_4) of d
11:                    normlm ← Normalize lmom_{m,k} with (μ̃_{m,k}, Σ_{m,k}) from Ta-
                              bles B.4 and B.5
12:                    R ← statistic based on dimension of normlm
13:                    c ← critical value based on alpha and dimension of normlm
14:                    if R > c then
15:                        return TRUE NEGATIVE
16:                    else
17:                        return FALSE POSITIVE
18:                    end if
19:                end for
20:            end for
21:        end for
22:    end for
23: end procedure
```

degree is 1 ($m = 1$), the clustering coefficient is always zero due to the fact that only one edge can be added as the network grows thus a triad formation is not possible. When considering the number of isolates and components caused by edge deletion, it is very apparent that as the minimum degree (m) increases, the networks become less affected by the deletion process (Tables 19 and 20) resulting in fewer isolates and fewer components. It is also apparent that the required network size k and proportion of deletion p that would result in isolates and components becomes larger as m increases. This is expected since most of the nodes when $m = 1$ have only a single edge that connects them to another node. Thus, any deletion will likely cause some isolates that will eventually lead to the network being fragmented, resulting in more components.

Table 19. Summary of isolates caused by edge deletion

m	Degree Level	k	p	# of Isolates 95% CI	
				(*lowest*)	(*highest*)
1	low	all	all	$(6, 8)$	$(4876, 4956)$
	medium	$6, \ldots, 14$	all	$(6, 8)$	$(4876, 4956)$
	high	$9, \ldots, 14$	all	$(6, 19)$	$(553, 688)$
2	low	$7, \ldots, 14$	$0.04, \ldots, 0.5$	$(5, 13)$	$(1813, 1910)$
	medium	$9, \ldots, 14$	$0.15, \ldots, 0.5$	$(5, 15)$	$(537, 615)$
	high	$13, 14$	$0.4, 0.5$	$(6, 21)$	$(17, 39)$
3	low	$8, \ldots, 14$	$0.15, \ldots, 0.5$	$(7, 17)$	$(716, 792)$
	medium	$10, \ldots, 14$	$0.3, \ldots, 0.5$	$(9, 22)$	$(224, 282)$
4	low	$9, \ldots, 14$	$0.3, \ldots, 0.5$	$(5, 15)$	$(288, 345)$
	medium	$13, 14$	0.5	$(11, 71)$	$(28, 134)$
5	low	$11, \ldots, 14$	$0.3, \ldots, 0.5$	$(10, 25)$	$(118, 158)$
	medium	14	0.5	none	$(11, 28)$
6	low	$12, \ldots, 14$	$0.4, 0.5$	$(8, 23)$	$(46, 75)$
7	low	$13, 14$	0.5	$(7, 21)$	$(18, 38)$

Further examination of the number of components caused by the deletion process on the network has some implication in real world applications where it might be of interest to study the vulnerability of a particular network either for interdiction or protection. Figures 21 and 22 show that even for the smallest k, edge deletion

Table 20. Summary of components resulted from edge deletion

m	Degree Level	k	p	# of Components 95% CI	
				(*lowest*)	(*highest*)
1	medium	5	$0.4, 0.5$	none	$(1, 2)$
	high	all	all	$(1, 2)$	$(823, 962)$
2	medium	all	$0.04, \ldots, 0.5$	$(1, 2)$	$(39, 66)$
	high	$10, \ldots, 14$	$0.2, \ldots, 0.5$	$(1, 2)$	$(1, 4)$
3	medium	$7, \ldots, 14$	$0.2, \ldots, 0.5$	$(1, 2)$	$(2, 10)$
4	medium	$11, \ldots, 14$	$0.4, 0.5$	$(1, 2)$	$(1, 3)$
5	medium	14	0.5	none	$(1, 2)$

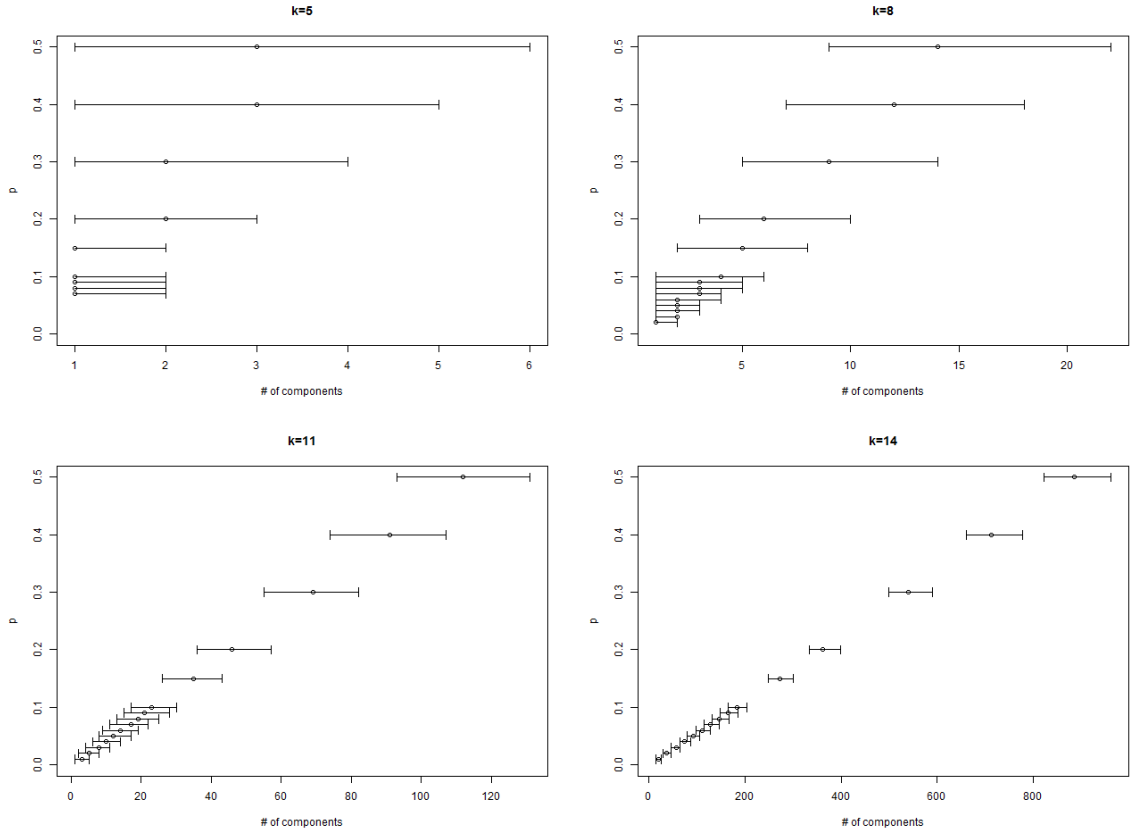


Figure 21. Number of components caused by high degree edge deletion on $m = 1$. (Bars are 95% CI.)

causes the network to be broken apart into multiple components. Although, smaller proportions (p) do not cause the same effect as the minimum degree (m) gets larger as they do when the minimum degree is $m = 1$. However, as k gets larger, smaller

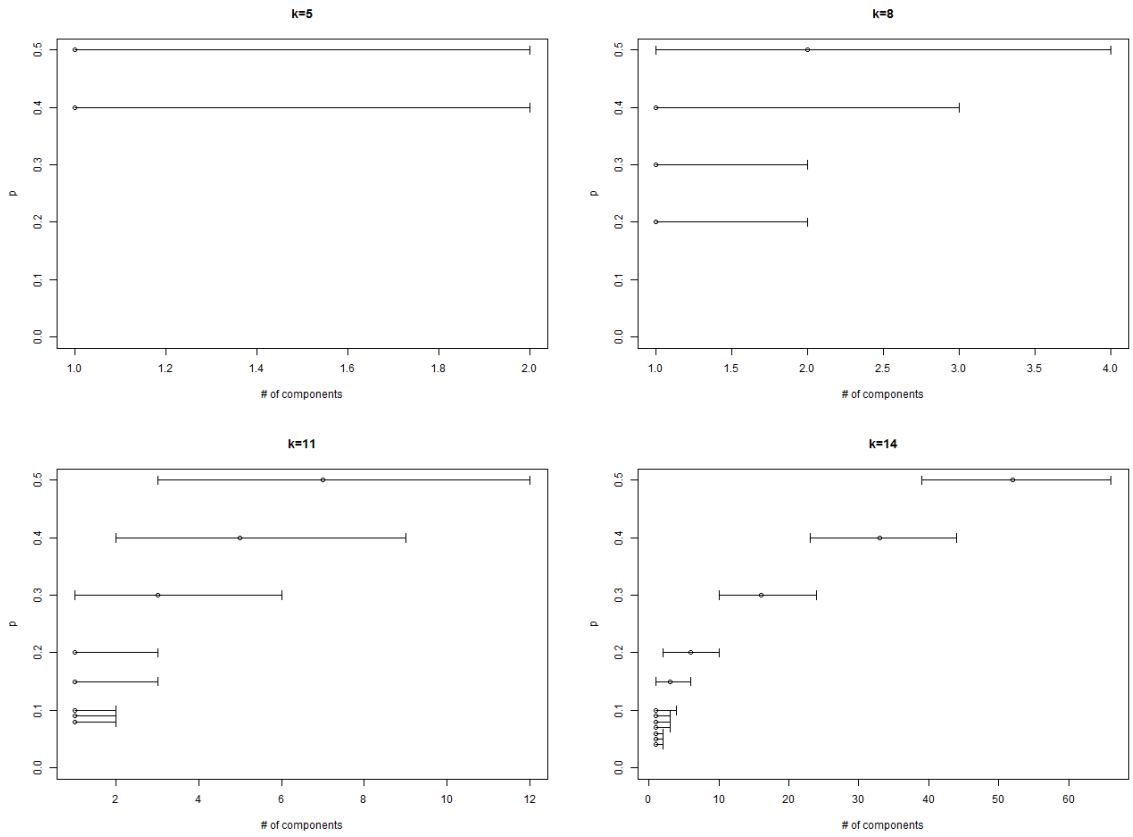


Figure 22. Number of components caused by medium degree edge deletion on $m = 2$. (Bars are 95% CI.)

proportions start to affect both $m = 1$ and $m = 2$. It should be noted that for $m = 1$, it is high degree edge deletion that is producing such result whereas it is medium degree edge deletion for $m = 2$. Other combinations of degree level and m do not provide meaningful results with respect to the number of isolates and components.

When studying the change in the clustering coefficient caused by the edge deletion process, only edge deletion on high degree nodes resulted in a significant change in the clustering coefficient as the proportion of deletion, p, increases, and this is true for all $m \neq 1$ (Figures 23, C.3 and C.4). However, the size k at which the clustering coefficient becomes significantly different as a function of p varies with the minimum degree, m. It seems that the clustering coefficient becomes significantly smaller as p increases, but only for $k \geq 9$, and the size k at which this happens decreases as

m increases. This result is counter to what is observed from the results on isolates and components, where it was shown that as m increases, the networks become less affected by edge deletion on high degree nodes. This suggests that the clustering coefficient is able to capture a characteristic of the network as a function of edge deletion that was not able to be captured by the number of isolates and components alone. Overall, edge deletion seems to have created isolates uniformly across all sizes of networks, but only created components on medium and high degree levels. Even so, the clustering or triadic structure of the network are not affected unless the high degree nodes are affected.

Power of the Test on Edge Deletion.

For low degree and high degree edge deletion, as the proportion of edges deleted (p) increases, the test on λ_2 becomes more likely to reject the degraded network from being classified as the network from which it originated, but the power drops unexpectedly when p becomes too large for $m = 2$ (Figures C.7 to C.9). Figures 24 and 25 show the distribution for L-scale with respect to p for $m = 2$, in which the distribution increases (decreases) and then decreases (increases) for low (medium) degree, explaining the patterns in the power curves. However, the multivariate test of $(\lambda_2, \tau_3, \tau_3)$ outperforms the univariate test of λ_2 in all cases and is able to maintain its power even when p becomes large. For medium degree edge deletion, although the multivariate test outperforms the univariate test, the power is less stable when compared to both low degree and high degree where there is a drop in power around $p = 0.2$ for $k \geq 8$. This seems to suggest that it is harder to detect subtle degradation within the network with respect to edge deletion when only the medium (median) degree nodes are affected.

It is standard practice to consider a power of 0.8 as being sufficient to deem a test

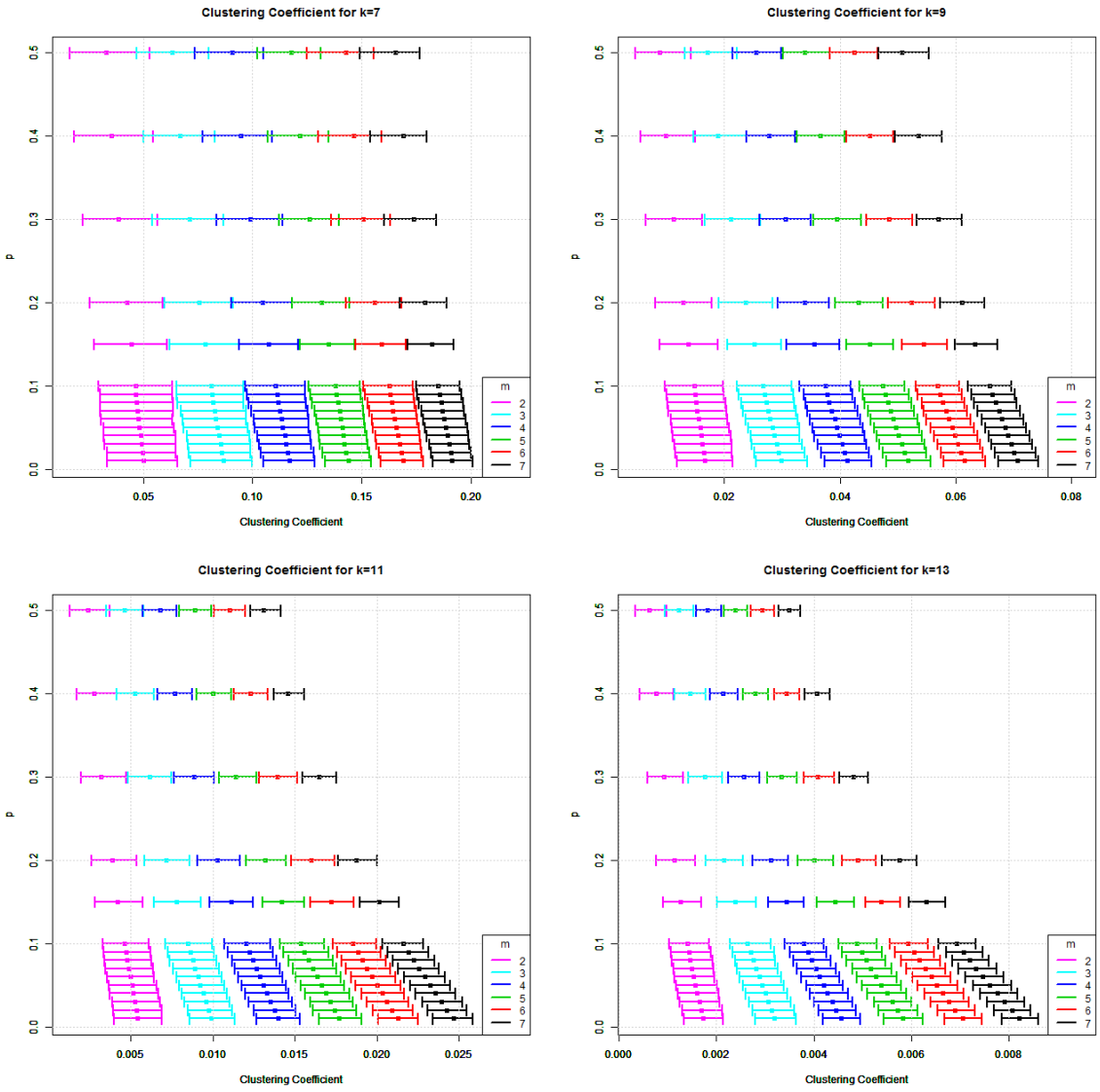


Figure 23. Clustering coefficients of networks after edge deletion on high degrees. (Bars are 95% CI.)

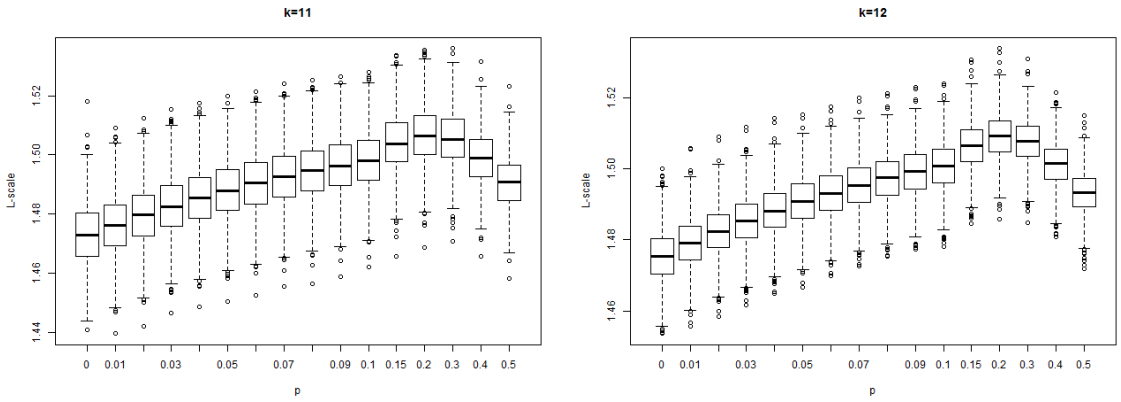


Figure 24. Boxplot of λ_2 for $m = 2$ of low degree deletion

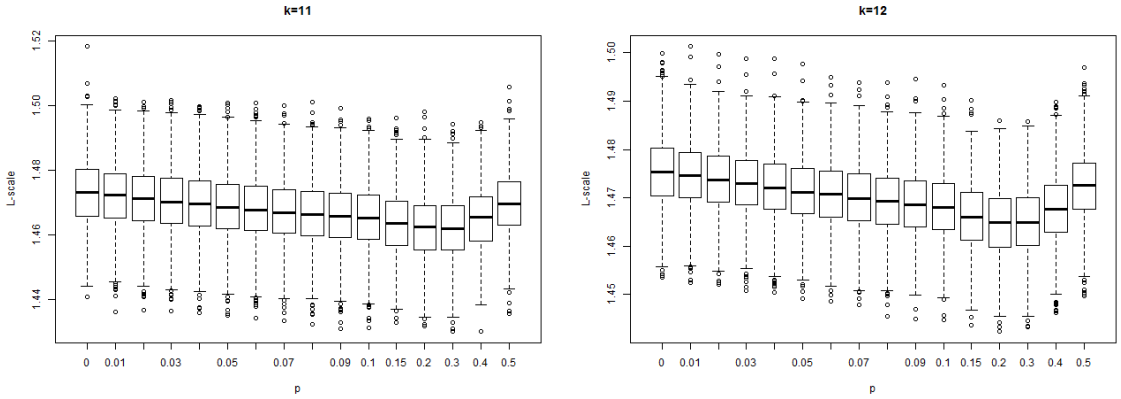


Figure 25. Boxplot of λ_2 for $m = 2$ of medium degree deletion

as being a good test. Therefore, the minimum network size (k) that is required for a test to achieve 80% power versus the proportion of deletion (p) is compared between the univariate and multivariate test. As shown in Table 21 and Figure C.10, the univariate test using λ_2 could not achieve the desired power for some values of p, but the multivariate test was able to do so for all cases except for when $m = 1$ and $p < 0.1$ in low and medium degree deletion. As p increases, the required k becomes smaller which suggests that the proportion of deletion has to be high when the network is small for the test to be able to detect the degradation. This is also evident when looking at the smallest proportion of edge deletion required to achieve 80% power as shown in Table 21. Although the trivariate test could not achieve 80% power when

m is larger for $k = 5$, it outperforms the univariate test in general. This makes sense since there are less edges affected in the smaller networks and a small p might not cause any deletion. Overall, considering deletion of only minimum degree nodes (low degree), 20% or greater of edges must be deleted before the test using λ_2 could detect a change in the network. However, the trivariate test using $(\lambda_2, \tau_3, \tau_4)$ can detect the deletion from only a 1%-15% proportion of deletion for $m \geq 2$. When considering deletion of high degree nodes, the test using $(\lambda_2, \tau_3, \tau_4)$ can detect degradation of large networks with as little as 2% of edges being deleted.

Table 21. Smallest proportion of edge deletion, p, required to achieve 80% power

		$k = 5$		$k = 7$		$k = 10$	
Degree Level	m	λ_2	$(\lambda_2, \tau_3, \tau_4)$	λ_2	$(\lambda_2, \tau_3, \tau_4)$	λ_2	$(\lambda_2, \tau_3, \tau_4)$
Low	1	X	0.30	X	0.30	0.40	0.30
	3	X	0.04	X	0.02	0.20	0.01
	5	X	0.07	0.50	0.03	0.20	0.01
	7	X	0.15	0.50	0.04	0.20	0.01
Medium	1	X	0.40	X	0.40	0.40	0.30
	3	X	X	X	0.2	X	0.15
	5	X	X	X	0.40	0.50	0.30
	7	X	X	X	0.40	X	0.30
High	1	0.50	0.30	0.30	0.09	0.15	0.02
	3	X	0.40	0.40	0.30	0.10	0.07
	5	X	X	0.30	0.30	0.08	0.06
	7	X	X	0.30	0.30	0.07	0.06

Implication of Edge Deletion.

There are a few implications from the results of edge deletion that can be made for networks that can be characterized as Barabási-Albert networks. Edge deletion can be illustrated as destroying the connectors within a network. For example, edge deletion in the context of road networks between places of interest can be thought of as destroying or seizing the roads connecting those places. In social context, it can

be thought of as intercepting or blocking means of communications between individuals, or it can even be thought of as ruining the relationships between individuals. Depending on the objective, one might be more interested in degrading a network, or in detecting whether or not the network is being degraded and reacting accordingly.

The most obvious implication that can be made is that nodes of networks with lower degree are at risk of being isolated or fragmented into multiple sub-networks if the connections between nodes are deleted, especially between those with minimum or median degrees. However, for these networks, the clustering within the remaining connected nodes is not affected, and this implies that the connection within the remaining sub-networks is intact. On the other hand, for networks whose minimum degree (m) is larger, the more effective method of degrading the network is to delete connections for nodes with high degree, and although this will not degrade the network into sub-networks, it will reduce the clustering within the network and reduce its connectivity. These results are summarized in Table 22.

Table 22. Recommended degree level at which to perform edge deletion that results in isolates, components, and changes in clustering

Characteristic Affected	m (minimum degree)						
	1	2	3	4	5	6	7
Isolates	Low, Medium, High	Low, Medium	Low, Medium	Low	Low	none	none
Components	High	Medium, High	Medium	Medium	none	none	none
Clustering	none	High	High	High	High	High	High

On the opposite end, if one is concerned with detecting degradation within the network, then this can be achieved with good power using the trivariate test on the L-moments of the degree distribution. This is especially true for detecting degradation caused by edge deletion of nodes with minimum and median degree where it was shown

that the trivariate test outperforms the univariate test. Additionally, although the univariate test performs comparably to the trivariate test for detecting edge deletion on high degree nodes, the trivariate test is better at detection for networks with smaller minimum degree and size (Table 21).

4.5.2 Node Deletion.

Characteristics of Node Deletion.

When considering the number of isolates and components caused by node deletion, it is very apparent that node deletion is not as destructive on the networks as edge deletion since only a few combinations of degree level, k, and p resulted in isolates or components. Similar to edge deletion, as m increases, the networks become much less affected by node deletion (Tables 23 and 24). However, only Barabási-Albert networks with $m = 1, 2$ are affected by node deletion, and $m = 1$ is more affected than $m = 2$, since the resulting number of isolates and components is larger by magnitudes. Again, this can be accounted to the small degrees in $m = 1$ as mentioned in Section 4.5.1.

Table 23. Summary of isolates caused by node deletion process

m	Degree Level	k	p	# of Isolates 95% CI	
				$(lowest)$	$(highest)$
1	high	$9, \ldots, 14$	$0.02, \ldots, 0.5$	$(6, 25)$	$(532, 700)$
2	medium	$12, \ldots, 14$	$0.3, \ldots, 0.5$	$(7, 23)$	$(16, 40)$
	high	$13, 14$	$0.4, 0.5$	$(6, 22)$	$(16, 40)$

Table 24. Summary of components resulted from node deletion process

m	Degree Level	k	p	# of Components 95% CI	
				$(lowest)$	$(highest)$
1	high	all	all	$(1, 5)$	$(699, 895)$
2	medium	$10, \ldots, 14$	$0.2, \ldots, 0.5$	$(1, 2)$	$(1, 3)$
	high	$10, \ldots, 14$	$0.2, \ldots, 0.5$	$(1, 2)$	$(1, 4)$

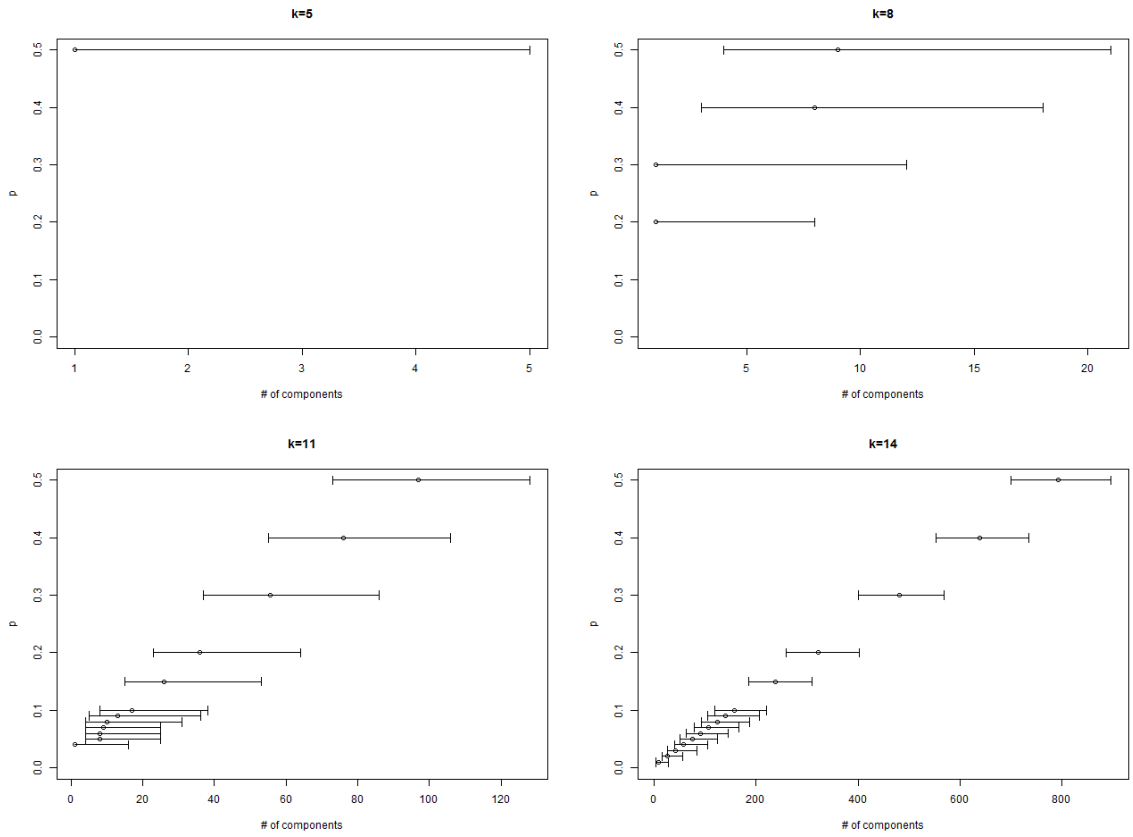


Figure 26. Number of components caused by high node deletion on $m = 1$. (Bars are 95% CI.)

Further investigation on the number of components caused by node deletion shows that only high degree level of $m = 1$ is heavily affected, and a higher proportion of deletion (p) is needed to cause any fragmentation of the network when k is small (Figure 26). However, when k is large, the resulting number of components when nodes are deleted is comparable to that of edge deletion, which again suggests that node deletion causes the same level of degradation in terms of number of components despite having only affected less item than edge deletion (i.e. less nodes versus less edges required for deletion).

Similar to edge deletion, investigation on the clustering coefficient of the networks with respect to node deletion suggests that clusters within the network are only affected by node deletion on high degree nodes. However, unlike edge deletion, only

networks with $m \geq 3$ are affected (Figure 27) and the effects are not as prominent as those for edge deletion. When $m = 2$, the clustering coefficient does not significantly change as more nodes are deleted (i.e. p increases) even when the network is large ($k = 14$) (Figure 28). Again, this result is the opposite from those of isolates and components. Thus, it seems that although node deletion on high degree results in a large number of isolates and components when the minimum degree is small ($m \leq 2$), the clustering of the network is not affected. On the other hand, when the minimum degree is not small ($m \geq 3$), the network is able to stay connected, but the remaining nodes now become less clustered as shown by the decreasing clustering coefficient.

Power of the Test on Node Deletion.

The test using λ_2 appears to have very low power for medium degree node deletion with $m \geq 2$ where it could barely achieve 80% power unless p and k are very large (Figure C.12). Medium degree node deletion also appears to be the only level of node deletion where the test using $(\lambda_2, \tau_3, \tau_3)$ definitely outperforms the univariate test with the exception of $m = 1$ for which the test using λ_2 appears to be on par for $k \geq 10$. For low and high degree node deletion, both tests seem to be equal in power. Nevertheless, for low degree node deletion, the multivariate test outperforms the univariate test when k is small or when m is large whereas in the high degree case, the multivariate test is noticeably better when $k = 8, 9, 10$ and $m \leq 3$. Neither test was useful for high degree node deletion when $k \leq 7$.

When comparing the minimum network size k that is required for a test to achieve 80% power versus the proportion of deletion p for node deletion, several observations are made. For low degree node deletion, the univariate test could not achieve the desired power unless $k \geq 10$, and even though the multivariate test performs better, it experiences the same problem when $m \geq 3$. For medium degree node deletion, the

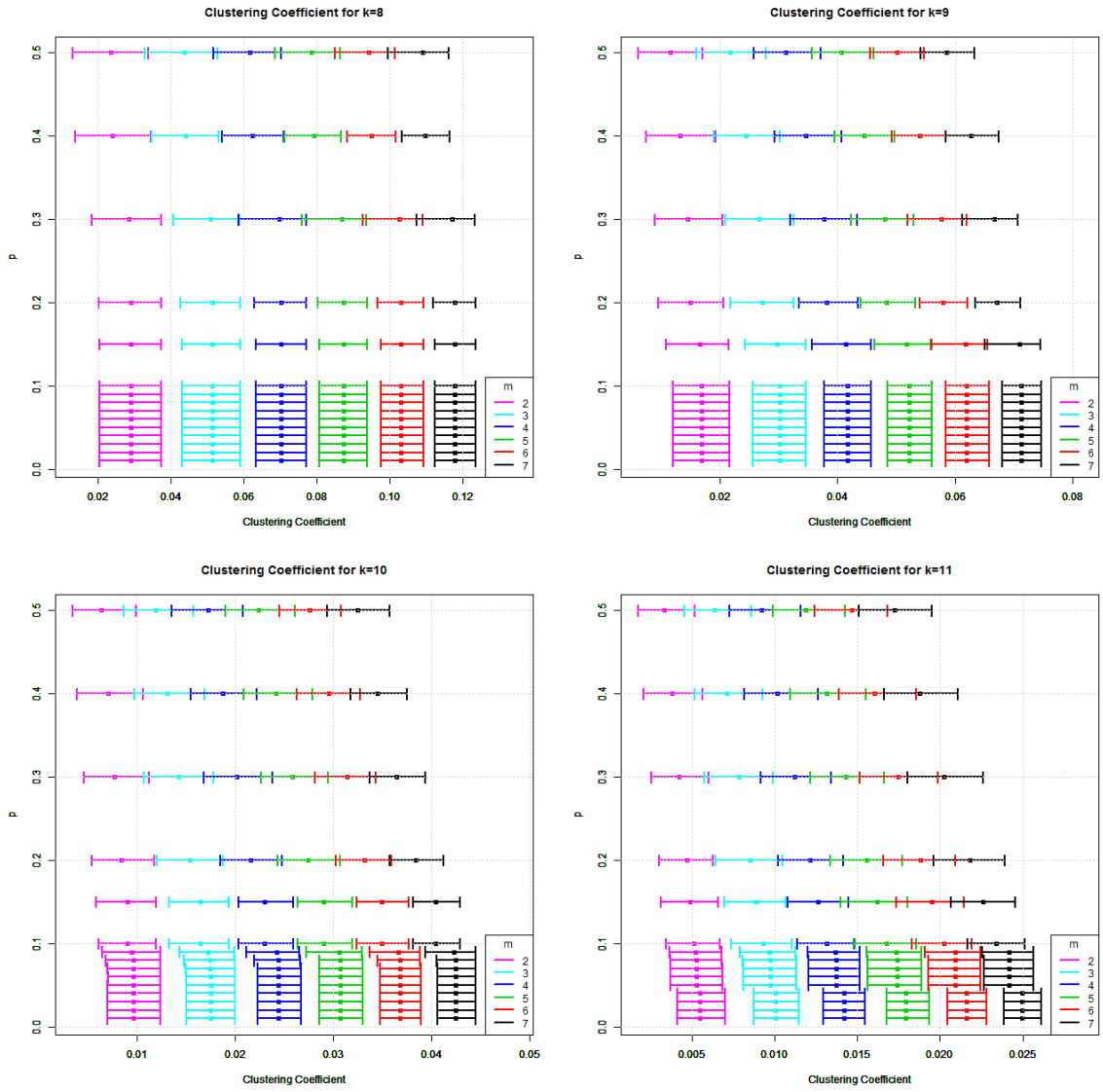


Figure 27. Clustering coefficients of networks after node deletion on high degrees. (Bars are 95% CI.)

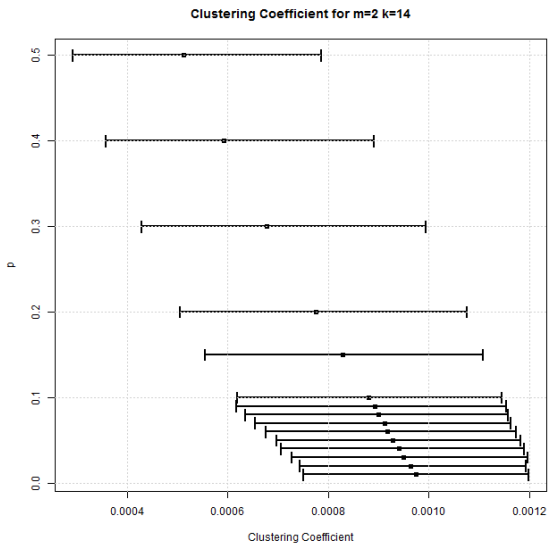


Figure 28. Clustering coefficients of networks after node deletion on high degrees when $m = 2$ and $k = 14$. (Bars are 95% CI.)

univariate test could not achieve the desired power for most cases and even when it does, the required network size is very large ($k \geq 10$) and the proportion required is very parge ($p \geq .20$). Lastly, the performance of both tests are comparable for high degree node deletion in terms of minimum required size ($k \geq 7$), and the multivariate test is only slightly better for $m \leq 3$ ($k \geq 7$). One overall observation that applies to all three degree node deletion levels is that, unlike the drastic drop in the required network size observed in edge deletion, the change for node deletion is gradual, requiring higher network sizes for change detection. Furthermore, the proportion of nodes affected must be higher than the proportion of edges affected to detect a change. These results imply that to detect degradation in a network, or to degrade a network, a bigger effect is obtained through a smaller proportion affected with respect to edge deletion versus node deletion.

Table 25. Smallest proportion of node deletion, p, required to achieve 80% power

		$k=5$		$k=7$		$k=10$	
Degree Level	m	λ_2	$(\lambda_2, \tau_3, \tau_4)$	λ_2	$(\lambda_2, \tau_3, \tau_4)$	λ_2	$(\lambda_2, \tau_3, \tau_4)$
Low	1	X	0.30	X	0.30	0.40	0.30
	3	X	0.04	X	0.02	0.50	0.40
	5	X	X	X	X	X	0.40
	7	X	X	X	X	X	X
Medium	1	X	0.40	X	0.40	0.40	0.30
	3	X	X	X	0.2	X	0.05
	5	X	X	X	0.50	X	0.15
	7	X	X	X	X	X	0.30
High	1	X	X	X	0.50	0.20	0.10
	3	X	X	X	0.50	0.20	0.10
	5	X	X	0.50	0.50	0.20	0.10
	7	X	X	0.50	0.50	0.10	0.10

Implication of Node Deletion.

Similar to edge deletion, there are a few implications from the results for node deletion. However, unlike edge deletion, node deletion can instead be illustrated as destroying the actors or entities of interest within a network. Going back to the context of road networks, a node deletion is analogous to destroying the actual places of interest within the network whereas in a social context, it can be thought of as detaining or removing the specific individuals from the network. Note that in both context, once the entities are removed, all connections between those entities to others in the network are rendered useless. This corresponds directly to edges being removed when removing the nodes in a graph.

It is apparent from the characteristic of node deletion that the nodes of networks with smaller minimum degree (m) are at risk of being isolated or fragmented into multiple sub-networks if the nodes with medium or high degree are deleted. However, the clustering within the remaining connected nodes for this group of network (i.e. networks with smaller minimum degree) is not affected. For networks with larger

minimum degree, the more effective method of degrading the network is to delete nodes with high degree similar to edge deletion since it will reduce the clustering within the network and reduce its connectivity. These results are summarized in Table 26.

Table 26. Recommended degree level at which to perform node deletion that results in isolates, components, or changes in clustering.

Characteristic Affected	m (minimum degree)						
	1	2	3	4	5	6	7
Isolates	High	Medium, High	none	none	none	none	none
Components	High	Medium, High	none	none	none	none	none
Clustering	none	none	High	High	High	High	High

For detecting degradation, the trivariate test is shown to perform better than the univariate test especially for detecting degradation caused by node deletion of nodes with minimum and median degree. However, the univariate test performs comparably to the trivariate test for detecting node deletion on high degree nodes except for specific network sizes.

4.5.3 Summary of Sensitivity Analysis.

In summary, edge deletion affects the network more so than node deletion with respect to the number of isolates and components caused and to changes in clustering. The trivariate test for change detection using $(\lambda_2, \tau_3, \tau_4)$ is also more sensitive to edge deletion than node deletion since it is able to detect the changes at a smaller proportion of deletion (p). In other words, the test is able to detect degradation caused by edge deletion much sooner than if it is caused by node deletion. Recall that, here, degradation is defined as removal of nodes or edges within the network that changes the structure of the network and its degree distribution. However, the results

from the trivarite test are also applicable to growing networks that is a function of its minimum degree. The power of the test from Section 4.4 shows that the trivariate test is able to detect, with good power, if the minimum degree (m) of the network has changed even when the minimum degree is not small.

V. Empirical Degree Distribution of Barabási-Albert Networks

The secondary objective of providing an accurate estimate of a parameter, β, for which the hypothesis tests in the primary objectives are constructed will now be presented. As previously stated, such an estimate has not been conclusively provided within the available literature. Nevertheless, the proper value of β is necessary in order to form the correct hypotheses for the tests described above. Therefore, a simulation of the Barabási-Albert network is performed in order to estimate the parameter empirically using methods as proposed by Newman [70] and Clauset and others [16].

5.1 Network Simulation

Network data was simulated to 1) determine the values of the parameters for the degree distribution when assuming the degree follows a continuous power law (Pareto) distribution, and compare the distribution with the estimated parameters to the theoretical distribution as derived by Barabási and Albert [6], and 2) to determine which form of the degree distribution (Pareto or Yule-Simon) best fits the degree distribution. Two sets of degree distribution data were generated from simulated Barabási-Albert graphs of various parameter and size combinations. The first set of data is used for parameter estimation and comparison for the Pareto distribution whereas the second set of data is used for goodness of fit to determine which distributional form best represents the degree distribution. The simulation was conducted in R using the *igraph* package [17] for network generation and computation of network degrees. The Barabási-Albert algorithm takes in as input the number of nodes, n, and the number of edges added at each iteration, m^*. An optional starting graph input was also included that forces the algorithm to start with having m^* nodes,

each having degree m^* so that the condition in Equation (3) as set by Barabási-Albert [6] is met and to also guarantee a connected graph. The parameter selection for the simulation is listed in Table 27 where 1000 independent networks were generated for each of the 33 combinations of graph parameters and sizes. These combinations were chosen so that the sizes of the networks examined spanned from small to large, and to also study the behavior of the associated Barabási-Albert graphs that have nodes with low degrees (i.e. $m^* = 2, 4, 6$).

Table 27. Parameters for network simulation

Parameters	Size
$m^* \in \{2, 4, 6\}$	$n = 2^k; k \in \{5, 6, \ldots, 14, 15\}$

5.2 MLE and Nonparametric Estimation

Under the assumption that the degree distribution of the Barabási-Albert network follows the Pareto distribution, the MLE-nonparametric method [16] is used to estimate the parameters m and β using $\hat{m}_{MLEnp}$ and $\hat{\beta}_{MLEnp}$, respectively, as described in Section 2.2.2. Additionally, assuming that $\beta = 2$, m was directly estimated by letting $\tilde{m}_{MLEnp}$ be a value that minimizes the KS statistic. Estimation of the parameters is performed using the poweRlaw package in R [29]. Another estimate of m was computed by fixing $\beta = \hat{\beta}_{MLEnp}$ and solving the least square estimate $\hat{m}_{MLEnp2}$. The purpose was to obtain a smaller estimate of m since the method proposed by Clauset and others [16] truncates the support of the distribution. For $X \sim Pareto(m, \beta)$, the SF is given by

$$S_X(x) = m^\beta x^{-\beta}$$

and

$$\ln(S_X(x)) = \beta \ln(m) - \beta \ln(x) \tag{25}$$

which is a linear function of $\ln(x)$. Therefore m and β can be estimated by the method of least squares using Equation (25). Given a network of size n, and the degrees for each of the n nodes of the network denoted by the $n \times 1$ vector, $\underset{\sim}{\mathbf{x}}$, the empirical SF ($\hat{S}_n(x)$) was computed from the empirical CDF for each observed value x as

$$\hat{F}_n(x) = \frac{\text{number of elements in sample} \leq x}{n}$$

and $\hat{S}_n(x) = 1 - \hat{F}_n(x)$. Let

$$\underset{\sim}{\mathbf{y}} = \ln \hat{S}_n(\underset{\sim}{\mathbf{x}}) + \beta \ln(\underset{\sim}{\mathbf{x}}), \quad \mu = \beta \ln(m), \text{ and } \mathbf{X} = \underset{\sim}{\mathbf{1}},$$

then the least squares solutions to $\underset{\sim}{\mathbf{y}} = \mathbf{X}\mu$ when $\beta = \hat{\beta}_{MLEnp}$ is $\tilde{m} = \exp\left(\frac{\bar{\mathbf{y}}}{\hat{\beta}_{MLEnp}}\right)$. For each combination of n and m, a Monte Carlo distribution for $\hat{m}_{MLEnp}, \hat{\beta}_{MLEnp}$, and $\hat{m}_{MLEnp2}$ as well as $\tilde{m}_{MLEnp}$ for when $\beta = 2$ was built from the first set of simulated data. The data was also used to create a 95% bootstrapped confidence interval of the estimates where the estimates were sampled from the 1000 simulated networks.

Results demonstrate that the MLE-nonparametric method is less stable for estimating m than just simply fitting a least square estimates on the SF. As shown in Figure 29, as n increases, the estimates and associated 95% confidence intervals for $\hat{m}_{MLEnp}$ shift upwards and widens even when fixing $\beta = 2$. The overlapping intervals also indicate that $\hat{m}_{MLEnp}$ estimates are not significantly different by m^*, nor are they dependent on m^*. However, the estimate, $\hat{m}_{MLEnp2}$, has a pattern that is more conforming to the theoretical $m = m^*\sqrt{\frac{(n-m^*)}{n}}$ where the values are closer to m and falls within the 95% CI (Figure 29 and Table 28). Since any observed degree that falls outside the support of a hypothetical distribution will result in a likelihood of zero, the $\hat{m}_{MLEnp2}$ estimate is used since it provides a support that better represent

the observed degree distribution as opposed to $\hat{m}_{MLEnp}$.

Table 28. MLE-nonparametric point estimates of m and β.

m^*	$k : n = 2^k$	$\hat{m}_{LS}\|(\beta = 2)$	$\hat{m}_{LS}$	$\hat{\beta}_{LS}$	m	$\hat{m}_{MLEnp}$	$\hat{m}_{MLEnp2}$	$\hat{\beta}_{MLEnp}$
2	5	1.7148	1.4427	1.5475	1.9365	5	1.9762	2.6275
2	6	1.7079	1.4154	1.5229	1.9685	6	1.9769	2.5919
2	7	1.7081	1.4200	1.5336	1.9843	7	1.9465	2.5291
2	8	1.7001	1.4126	1.5419	1.9922	9	1.9346	2.5171
2	9	1.6965	1.4167	1.5544	1.9961	11	1.9147	2.4716
2	10	1.6953	1.4225	1.5626	1.9980	13	1.905	2.4565
2	11	1.6934	1.4235	1.5689	1.9990	16	1.886	2.4181
2	12	1.6927	1.4256	1.5720	1.9995	19	1.8933	2.4327
2	13	1.6919	1.4268	1.5744	1.9998	23	1.8904	2.4287
2	14	1.6914	1.4269	1.5753	1.9999	27	1.8929	2.435
2	15	1.6914	1.4281	1.5769	1.9999	32	1.8927	2.4313
4	5	3.7156	3.5147	1.8101	3.7417	6	4.0697	2.3686
4	6	3.7319	3.4111	1.7241	3.8730	8	4.0085	2.2735
4	7	3.7149	3.3739	1.7008	3.9370	10	3.9683	2.2546
4	8	3.6975	3.3558	1.7079	3.9686	12	3.895	2.1927
4	9	3.6775	3.3428	1.7130	3.9843	14	3.8523	2.1721
4	10	3.6663	3.3445	1.7222	3.9922	17.5	3.8209	2.1581
4	11	3.6582	3.3487	1.7315	3.9961	21	3.821	2.1625
4	12	3.6534	3.3516	1.7376	3.9980	26	3.8154	2.1622
4	13	3.6509	3.3541	1.7421	3.9990	32	3.815	2.1659
4	14	3.6490	3.3554	1.7447	3.9995	38	3.8174	2.1715
4	15	3.6481	3.3568	1.7462	3.9998	46	3.823	2.1785
6	5	5.6962	5.7645	2.0553	5.4083	7	5.8604	2.0934
6	6	5.7858	5.5453	1.8535	5.7118	9	5.9282	2.0921
6	7	5.7683	5.4150	1.7895	5.8577	11	5.8841	2.0678
6	8	5.7323	5.3503	1.7760	5.9293	13	5.8362	2.0589
6	9	5.6955	5.3225	1.7796	5.9647	17	5.7929	2.0596
6	10	5.6706	5.3148	1.7888	5.9824	21	5.7643	2.0574
6	11	5.6537	5.3135	1.7969	5.9912	25	5.7637	2.0708
6	12	5.6444	5.3197	1.8053	5.9956	31	5.7608	2.0785
6	13	5.6380	5.3227	1.8109	5.9978	38	5.7671	2.0861
6	14	5.6333	5.3255	1.8146	5.9989	46	5.7702	2.091
6	15	5.6314	5.3284	1.8174	5.9995	55	5.7809	2.099

The interval estimates of β as shown in Figure 30 seem to be more stable with the only changes coming from the width of the confidence intervals which becomes smaller as n increases. Since these intervals also overlap, we conclude that $\hat{\beta}_{MLEnp}$

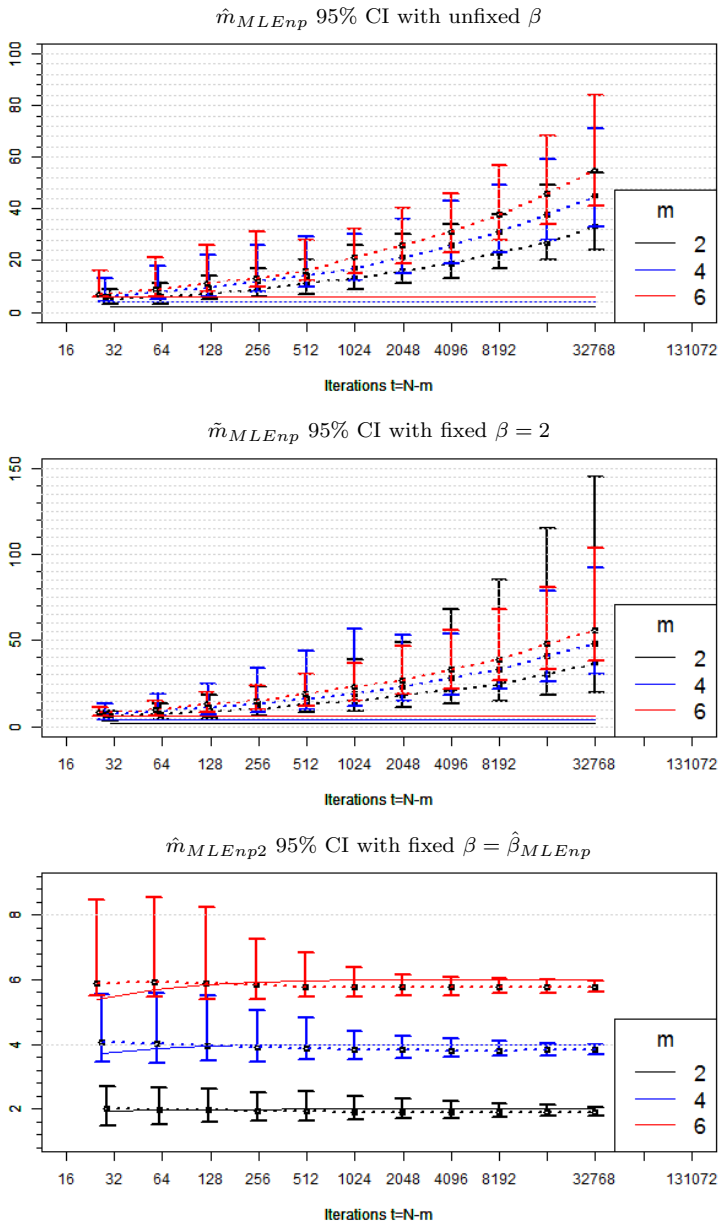


Figure 29. Top: $\hat{m}_{MLEnp}$ estimate with β unfixed. Middle: $\tilde{m}_{MLEnp}$ estimate with fixed $\beta = 2$. Bottom: $\hat{m}_{MLEnp2}$ estimate with fixed $\beta = \hat{\beta}_{MLEnp}$. t is the number of iterations required to complete a graph of size n

is not significantly different for m^*, and even though the median values seem to be stable across n, the confidence intervals seem to be converging to a value higher than $\beta = 2$ which, again, is counter to what is theoretically suggested. However, compared to the least squares estimate, the $\hat{\beta}_{MLEnp}$ values are much closer to the

theoretical value. The result from the MLE-nonparametric approach suggests that the distribution derived by Barabási-Albert [6] is only observed in the tail of the degree distribution as implied by the truncation for the estimation of β. In essence, the best fit was obtained by setting a high $\hat{m}_{MLEnp}$ and truncating the lower end of the distribution and while emphasizing the tail.

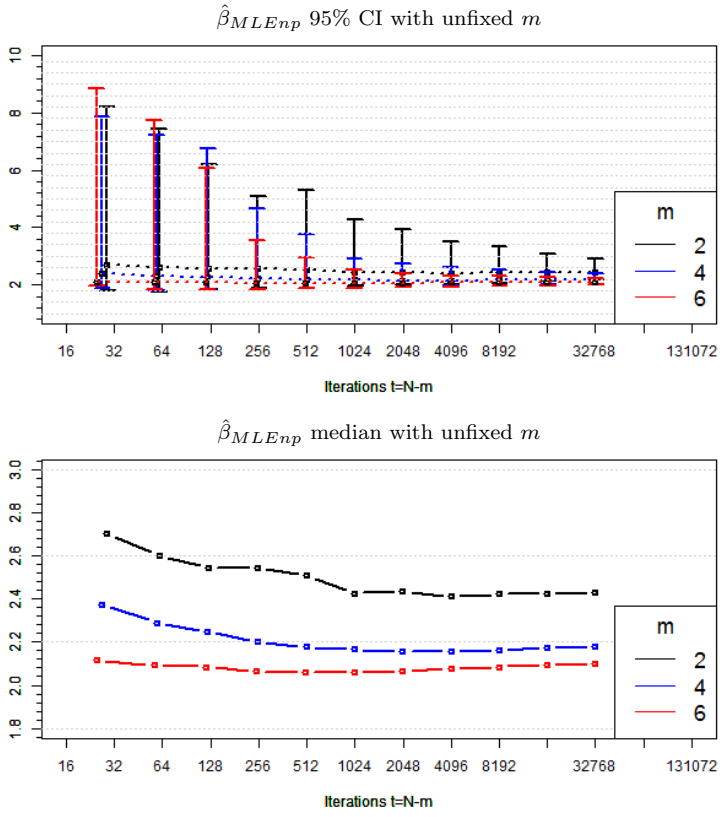


Figure 30. $\hat{\beta}_{MLEnp}$ estimate with m unfixed. t is the number of iterations required to complete a graph of size n

5.3 Goodness of Fit

Several metrics were used to determine which distribution and parameters best fit the degree distribution: mean squared error (MSE) and $-loglikelihood$ computed from the second set of simulated data fitted against various distributional forms (Pareto and Yule-Simon) and the parameter estimates. The Zipf's law was

not compared since the Yule-Simon distribution is considered a better fit under the discrete assumption [70]. Recall that Newman [70] suggested that the exponent for the Barabási-Albert graph is $\beta = 1.2$ as opposed to $\beta = 2$ [6]. In addition, Li and others [59] found that the exponents of the power law distribution from a simulated Barabási-Albert graph for $m^* \in \{4, 6, 8\}$ for size $n = 5000$ are less than $\beta = 2$. Since our own analysis using the estimates from least squares and MLE-nonparametric methods seems to suggest that the exponent varies depending on m^* and n, other parameter selections were included for the goodness of fit comparison. The various selections of β and m for the Pareto and Yule-Simon distribution examined are given in Table 29 and span the range of values suggested in the literature and those found in Section 5.2.

Table 29. (m, β) **combinations for goodness of fit comparisons**

Distribution	β	m
Pareto	$1.2, 1.5, 1.8, 2, 2.1$	$m^*\sqrt{\frac{(n-m^*)}{n}}$
	median for $\hat{\beta}_{MLEnp}$	median for $\hat{m}_{MLEnp}$ and $\hat{m}_{MLEnp2}$
	2	median for $\tilde{m}_{MLEnp}$
Yule-Simon	$1.2, 2$	m^*

Therefore, 1000 estimates for the distribution of the MSE and $-loglikelihood$ for each of the 33 network parameter combinations were used to compare 1) if the least squares and MLE-nonparametric methods of parameter estimation differed significantly in resulting parameter estimates, and 2) which method produces better estimates of m and β via the MSE and $-loglikelihood$. The MSE is defined as the average squared difference between the predicted value of the empirical probability density value for each parameter and the expected density of each parameter with

respect to a hypothesized distribution. A model with maximum likelihood value (or lower $-loglikelihood$) and lowest MSE is preferred.

Since the support of the Pareto distribution is bounded below by m, the joint probability of any random sample with elements that are less than m will result in the likelihood being equal to zero. On the other hand, the MSE is not affected as significantly as the likelihood with respect to the bounded distribution. The likelihood is capturing the exact information for distribution fitting such that any random sample that falls outside of the support of the hypothesized distribution is likely not from said distribution. Therefore, a hierarchical goodness of fit approach is used to test model parameters by first looking at the $-loglikelihood$ to find a set of parameters that gave a significantly better fit, and then of those indistinguishable via the $-loglikelihood$, the MSE was used to distinguish which estimates produced the smallest error.

MSE and $-loglikelihood$ values of the second dataset were computed using the median point estimates from the bootstrapped $\hat{m}$ and $\hat{\beta}$ distribution given in Table 28 in addition to the other parameter selections as outlined in Table 29. The MLE-nonparametric estimate, $\hat{m}_{MLEnp}$, provided a poor estimation of the support for the degree distribution with a truncation point that essentially discounted a huge proportion of the degree distribution since the lower bound set by $\hat{m}_{MLEnp}$ is higher than the possible smallest degree m^* (Table 28). Moreover, as n increased, the point of truncation shifted higher, misspecifying a majority of the true degree density. Therefore, $\hat{m}_{MLEnp2}$ is used instead of $\hat{m}_{MLEnp}$ for comparison against the other estimates of the degree distribution as listed in Table 29.

A comparison of the Pareto and Yule-Simon degree distributions using both β values of 1.2 and 2 for the smaller networks (i.e. $k \leq 8,\ n = 2^k$), suggested that the Pareto distribution is either the best fit for the data or is no worse than the Yule-Simon distribution depending on the n and m^* as shown in Table 30. Here

in Table 30, bold values indicate which MSE values are significantly different from one another (95% CIs do not overlap) with bolded values representing a significantly better goodness of fit. Further, these results demonstrate that $\beta = 1.2$ is not a significantly different fit than the theoretical $\beta = 2$ for the degree distribution with $m^* = 2$ and $k \leq 7$. Therefore, we focus on the Pareto distribution as the assumption for the degree distribution of the Barabási-Albert graph, and a value of $\beta = 2$ for m^* greater than 2 and larger k.

A comparison of the goodness of fit measures for various β values for the Pareto distribution gives some insight into the behavior of the degree distribution as n and m^* varies (Tables D.8 and D.9). β gradually increased from 1.2 as the size of the network increased for all m^*, although the increase in the value of β is not significantly different until $k = 8$ and $k = 9$ for $m^* = 4$ and $m^* = 6$, respectively. However, this may be an artifact of the degree distribution having a higher density in the lower portion of its support, where the likelihood of any estimated parameter whose density is heavier in the lower portion will result in a better likelihood (Figures 31 and 32).

The $-loglikelihood$ confidence interval for the least squares estimate appears to be significantly worse than any of the other theoretical estimates (Table 31) with the exception of $m^* = 6$ for $k = 6$. However, the confidence interval of the MSE for the least squares estimate is either significantly better, or no worse than the best MSE of other theoretical distributions. Although the $-loglikelihood$ confidence intervals of the MLE-nonparametric estimates overlap the confidence intervals of the best fitting theoretical distributions, they do not overlap the $-loglikelihood$ confidence intervals of the least squares estimates (Table 31). Additionally, the least squares method is heavily influenced by the density of the lower portion of the degree distribution, and since this lower portion does not fit the same power law distribution of the tail as shown by the truncation point in the MLE-nonparametric method, the least

Table 30. -loglikelihood and MSE values of the fitted degree distribution by m^*, k, and parameter assumptions

−loglikelihood

k	Distribution (β)	Median $m^*=2$	95% CI $m^*=2$	Median $m^*=4$	95% CI $m^*=4$	Median $m^*=6$	95% CI $m^*=6$
5	Yule (2.0)	58.73	(55.6,61.8)	76.71	(74.7,78.7)	87.15	(85.6,88.5)
5	Yule (1.2)	60.41	(57.9,62.9)	78.99	(77.4,80.6)	89.94	(88.7,91)
5	Pareto (2.0)	**47.88**	(**43.7**,**52.2**)	74.16	(71.6,76.6)	89.13	(87.3,90.7)
5	Pareto (1.2)	**51.18**	(**48.1**,**54.3**)	76.07	(74.2,77.8)	90.2	(88.9,91.4)
6	Yule (2.0)	117.91	(112.5,122.4)	154.83	(151.2,158.4)	177.89	(175,180.6)
6	Yule (1.2)	121.15	(116.7,124.9)	158.98	(156.2,161.8)	182.49	(180.3,184.6)
6	Pareto (2.0)	**93.96**	(**86.7**,**100.0**)	145.08	(140.9,149.5)	174.79	(171.5,178)
6	Pareto (1.2)	**101.32**	(**96.0**,**105.8**)	150.36	(147.3,153.6)	178.78	(176.4,181.1)
7	Yule (2.0)	234.86	(226.8,241.4)	308.74	(302.5,314.7)	355.64	(350.8,360.3)
7	Yule (1.2)	241.44	(234.8,246.8)	317.19	(312.2,321.8)	364.79	(361,368.4)
7	Pareto (2.0)	**184.39**	(**173.6**,**193.3**)	**284.67**	(**277.3**,**291.8**)	**342.65**	(**337.2**,**347.9**)
7	Pareto (1.2)	**200.34**	(**192.4**,**206.9**)	297.26	(291.8,302.5)	353.34	(349.3,357.2)
8	Yule (2.0)	466.75	(454.8,477.5)	613.86	(604.3,622.4)	707.14	(699.9,714.8)
8	Yule (1.2)	480.45	(470.8,489.1)	631.49	(624.1,638.2)	726.34	(720.9,732.3)
8	Pareto (2.0)	**362.68**	(**347.0**,**377.2**)	**560.82**	(**549.6**,**570.9**)	**674.25**	(**666.2**,**683.0**)
8	Pareto (1.2)	396.47	(385,407.1)	588.82	(580.6,596.2)	699.41	(693.5,705.8)

MSE

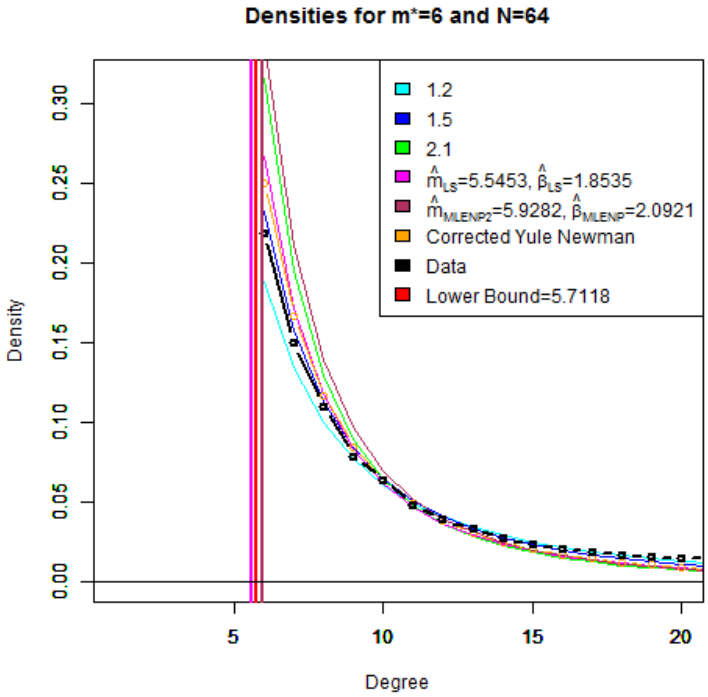
k	Distribution (β)	Median $m^*=2$	95% CI $m^*=2$	Median $m^*=4$	95% CI $m^*=4$	Median $m^*=6$	95% CI $m^*=6$
5	Pareto (2.0)	0.1113	(0.0828,0.1246)	0.0088	(0.0031,0.0151)	0.0030	(0.0012,0.0074)
5	Pareto (1.2)	**0.0094**	(**0.0023**,**0.0255**)	0.0023	(0.0005,0.0083)	0.0028	(0.0010,0.0081)
5	Yule (2.0)	**0.0034**	(**0.0004**,**0.0176**)	0.0028	(0.0007,0.0085)	0.0025	(0.0009,0.0067)
5	Yule (1.2)	**0.0045**	(**0.0010**,**0.0206**)	0.0034	(0.0011,0.0130)	0.0029	(0.0011,0.0085)
6	Pareto (2.0)	0.1242	(0.1074,0.1344)	0.0102	(0.0054,0.0148)	0.0024	(0.0010,0.0045)
6	Pareto (1.2)	**0.0102**	(**0.0037**,**0.0200**)	0.0010	(0.0002,0.0042)	0.0012	(0.0004,0.0039)
6	Yule (2.0)	**0.0021**	(**0.0003**,**0.0102**)	0.0013	(0.0003,0.0051)	0.0010	(0.0003,0.0030)
6	Yule (1.2)	**0.0033**	(**0.0010**,**0.0120**)	0.0025	(0.0009,0.0081)	0.0018	(0.0007,0.0053)
7	Pareto (2.0)	0.1308	(0.1181,0.1385)	0.0111	(0.0074,0.0143)	0.0025	(0.0013,0.0040)
7	Pareto (1.2)	**0.0102**	(**0.0050**,**0.0165**)	**0.0005**	(**0.0001**,**0.0023**)	0.0008	(0.0003,0.0023)
7	Yule (2.0)	**0.0016**	(**0.0002**,**0.0062**)	**0.0007**	(**0.0002**,**0.0027**)	0.0005	(0.0002,0.0018)
7	Yule (1.2)	**0.0029**	(**0.0011**,**0.0091**)	**0.0023**	(**0.0010**,**0.0060**)	0.0016	(0.0007,0.0038)
8	Pareto (2.0)	0.1336	(0.1249,0.1397)	0.0116	(0.0088,0.0141)	0.0026	(0.0017,0.0036)
8	Pareto (1.2)	**0.0100**	(**0.0062**,**0.0145**)	**0.0003**	(**0.0001**,**0.0012**)	0.0006	(0.0002,0.0015)
8	Yule (2.0)	**0.0012**	(**0.0002**,**0.0039**)	**0.0005**	(**0.0001**,**0.0016**)	0.0003	(0.0001,0.0009)
8	Yule (1.2)	**0.0031**	(**0.0014**,**0.0069**)	**0.0023**	(**0.0012**,**0.0049**)	0.0015	(0.0009,0.0029)

Bold indicates grouping of significantly smallest goodness of fit.

squares method is heavily underestimating β. Along with the fact that the MLE-nonparametric estimates are closer to the theoretical values of m and β, this seems to suggest that the MLE-nonparametric approach with untruncated $\hat{m}$ is the more appropriate way to estimate the parameters of the power law distribution of the Barabási-Albert graphs.

Table 31. Goodness of fit for least squares and MLE-nonparametric point estimates

m^*	k	$\hat{m}_{LS}$	$\hat{\beta}_{LS}$	**-loglikelihood Median, (95% CI)**	**MSE Median, (95% CI)**
2	6	1.4154	1.5229	128.446, (122.3829,133.5594)	0.0017, (0.0005,0.0094)
2	10	1.4225	1.5626	2019.9256, (1989.4299,2045.7492)	0.0007, (0.0004,0.0016)
2	15	1.4281	1.5769	64320.948, (64171.6391,64484.4478)	0.0007, (0.0006,0.0008)
4	6	3.4111	1.7241	159.1405, (155.3638,163.1259)	0.0013, (0.0002,0.0053)
4	10	3.3445	1.7222	2532.6096, (2511.0138,2554.2864)	0.0002, (0.0001,0.0006)
4	15	3.3568	1.7462	80473.9143, (80344.5779,80598.8424)	0.0001, (0.0001,0.0002)
6	6	5.5453	1.8535	177.9161, (174.7885,180.9538)	0.00128, (0.00044,0.00346)
6	10	5.3148	1.7888	2873.0657, (2855.4781,2892.2068)	0.00009, (0.00003,0.0003)
6	15	5.3285	1.8174	91197.9158, (91076.6104,91311.3347)	0.00004, (0.00003,0.00006)
m^*	k	$\hat{m}_{MLEnp2}$	$\hat{\beta}_{MLEnp}$	**-loglikelihood Median, (95% CI)**	**MSE Median, (95% CI)**
2	6	1.9769	2.5919	95.3985, (86.7663,102.6784)	0.2948, (0.2850,0.2985)
2	10	1.9050	2.4565	1560.2044, (1519.0710,1595.0360)	0.1845, (0.1807,0.1873)
2	15	1.8927	2.4313	50097.2975, (49898.4811,50315.0100)	0.1690, (0.1684,0.1697)
4	6	4.0085	2.2735	Inf, (Inf,Inf)	0.0278, (0.0111,0.0548)
4	10	3.8209	2.1581	2304.0024, (2278.9483,2329.1504)	0.0105, (0.0092,0.0119)
4	15	3.8230	2.1785	73238.2560, (73088.5620,73382.8479)	0.0110, (0.0108,0.0112)
6	6	5.9282	2.0921	170.2382, (166.8492,173.5299)	0.0043, (0.0023,0.0065)
6	10	5.7643	2.0574	2728.0896, (2708.8076,2749.0749)	0.0018, (0.0014,0.0023)
6	15	5.7809	2.0990	86464.1137, (86330.6857,86588.8671)	0.0021, (0.0020,0.0022)



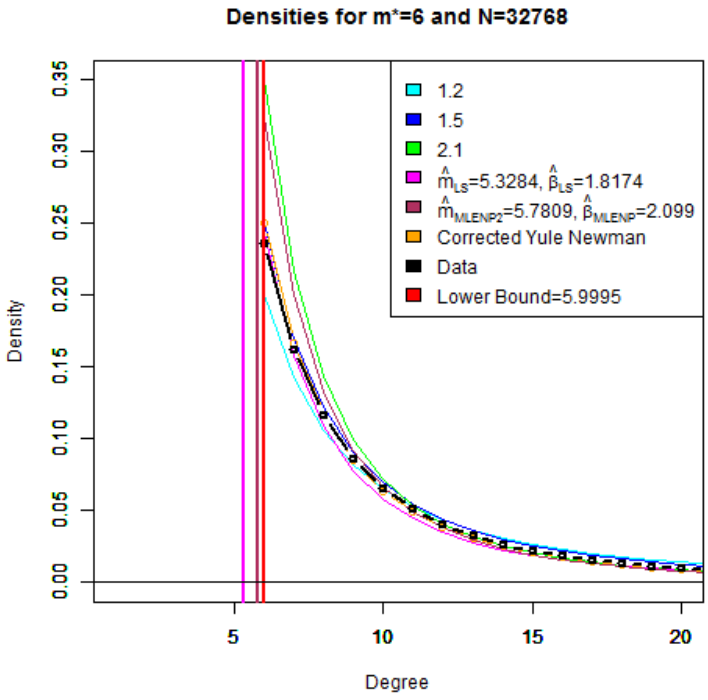


Figure 31. Distribution comparison for $m^* = 6$ with varying N

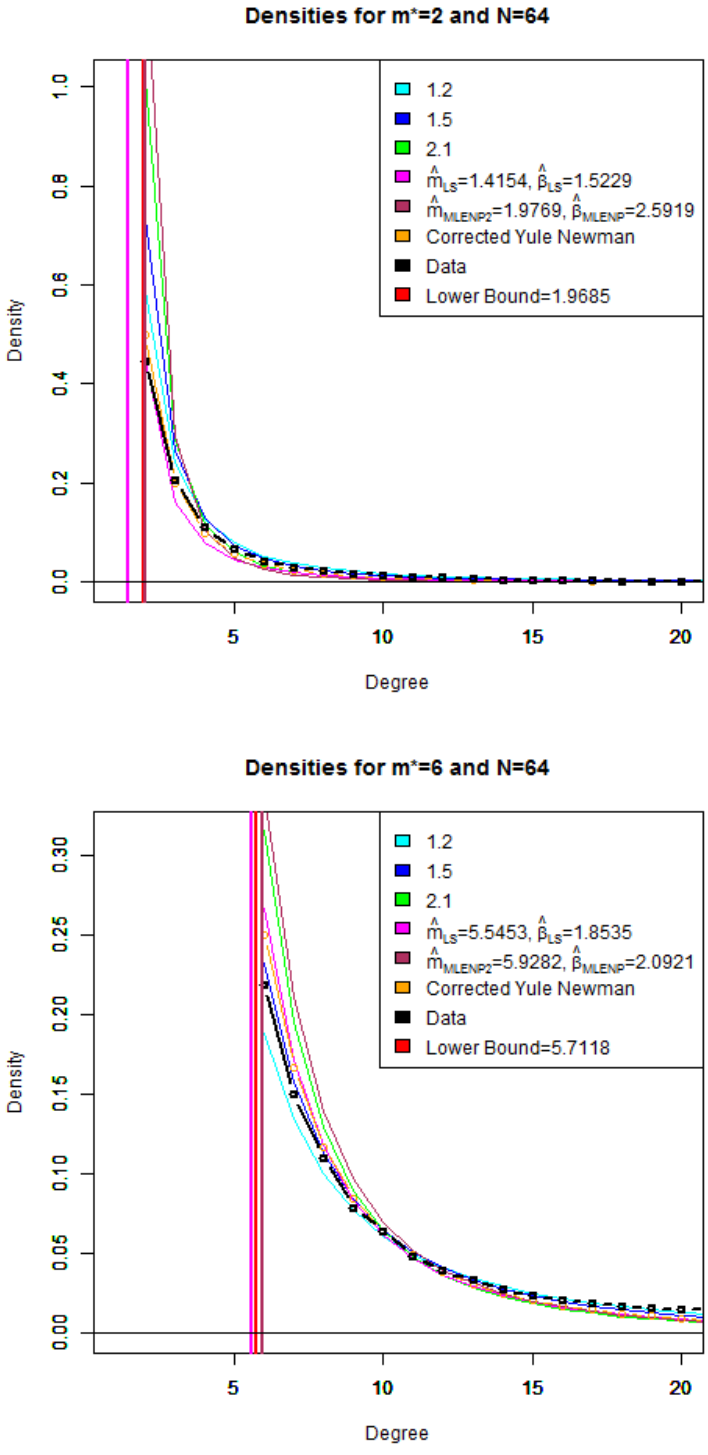


Figure 32. Distribution comparison for $N = 2^6$ with varying m^*

VI. Discussion

The question is raised as to whether or not the true exponent of the empirical degree distribution of the Barabási-Albert graph is equal to the theoretically derived value of $\beta = 2$, especially for relatively small networks with small parameter m^*, the minimum degree of a network node. To use the Barabási-Albert graph and its degree distribution to represent a network of interest, the parameters m and β of the degree distribution must be known or estimated. Previous authors have also made this query, at least in passing, as shown by the works of Newman [70] and Li and others [59], but no detailed investigation on the value of β has been documented to our knowledge. It is noteworthy, however, that some have suggested that the preferential attachment model causes biases on the connection of the high degree nodes in the Barabási-Albert graph [59, 88, 106]. It is conjectured that this is one of the factors contributing to the results of this dissertation which indicates that for relatively small networks, the value of m and β are not equal to the theoretical value as suggested by Barabási and Albert [6] and that m and β vary by m^* and n.

It should be noted that there is a limitation to the least squares estimation for the case of the Pareto distribution. Since a large portion of the density of the Pareto distribution lies in the lower part of its support, the fitted line on the doubly logarithmic SF might be heavily influenced by the lower portion of the density thus underestimating β if that line is not perfectly straight. One consideration is to truncate the data and only consider the tail of the distribution similar to the MLE-nonparametric approach, but realistically, given a real network, there is no prior knowledge as to where that cut off point should be, so the conservative approach is to include the full distribution. Another consideration is to use weighted least squares where the weights are based on proportion of the random variable at the observed value. This will give more influence to the tail of the distribution and account for the underestimation

of the least squares method. Therefore, it is proposed that the appropriate value for the parameters should be bootstrapped using an extended version of the MLE-nonparametric method by Clauset and others [16] from simulated Barabási-Albert networks.

This result comes with a few caveats which need to be addressed. Although the $-loglikelihood$ of the MLE-nonparametric estimates are worse than most of the hypothetical estimates, this could be due to the fact that the lower bound of the Pareto distribution for the MLE-nonparametric is lower than the theoretical lower bound of $m = m^* \sqrt{\frac{(n-m^*)}{n}}$. This causes, for the distribution of the MLE-nonparametric estimate, a large proportion of the probability density to be shifted smaller, and since $m^* > m$, then any distribution with a larger m will obtain a larger $-loglikelihood$ for a given degree. As shown in Figures 31 and 32, the distribution from the $\hat{m}_{MLEnp2}$ tends to have a lower density than the better fitting hypothetical distributions. Therefore, due to this limitation the proposed estimates did not have the best likelihood. However, they advance the research in this area and are the closest estimates when compared to the theoretical values of m and β versus the least squares and hypothetical estimates. The findings in this dissertation supports the notion that the classical theoretic underlining of the degree distribution of the Barabási-Albert model may not apply to smaller networks. The knowledge of the true degree distribution is necessary for characterizing a given network to an appropriate network model.

Nevertheless, a test of hypothesis with the assumption that the degree distribution of the Barabási-Albert graph follows the $Pareto(m, \beta)$ distribution was derived. The power of this test on β suggests that the power increases as m^* increases or as k increases. The test of hypothesis on m can exhibit poor power even in a close neighborhood of the true m for smaller k depending on the true value of m^*, but the power curve reaches a steady state with very high power as k gets larger. Additionally, a

UIT was derived for the Pareto distribution that improved the power compared to the individual tests on β and m. Application to simulated Barabási-Albert networks showed that corrections on the mean and variance of the statistic's distribution are required due to the degree correlation caused by the preferential attachment model. When this test was applied on a few datasets of real world network, it was found that the degrees for most networks examined are larger than what is expected from a Barabási-Albert network and that β is smaller than the theoretical value of 2. However, for some networks, the degree distribution of their associated sub-network is not significantly different than that of the Barabási-Albert network.

The results from the real world datasets are not entirely unexpected. Estimation of β for these networks without the assumption of the Pareto distribution through least square estimation [70] and MLE-nonparametric methods [16] have suggested that they differ from an assumed $\beta = 2$ of the Barabási-Albert network. Regardless, once a value of β is established, the tests, as constructed in this dissertation, were shown to work with good power when the proper values of m and β are provided. Further, it has been suggested that the Barabási-Albert properties of preferential attachment does not exist in real world data [88, 106], and that the degree distribution is a Power law with a cut-off [16]. However, even the ability to represent only the sub-network as a Barabási-Albert graph is very useful in that it still provides a method of modeling the central structures of the network (i.e. the hubs) through the Barabási-Albert model and is beneficial in making observation on a network removing the periphery nodes.

Finally, a test of hypothesis on the sample L-moments of the degree distribution of the Barabási-Albert network for the purpose of detecting degradation within the network structure was also derived. Although the test is based on the empirical degree distribution of the network, the assumption that the network degree follows

the $Pareto(m, \beta)$ distribution where $\beta > 1$ is still required. This is due to the fact that the mean of the distribution has to be defined in order for the second and other higher order L-moments to exist, but the mean is not defined if $\beta \leq 1$. The test on L-scale was shown to have good power when used to test for a Barabási-Albert network, but extending the test into a multivariate test by adding the L-skewness and L-kurtosis improved the power significantly for cases where the network size is small. This initial step of truth classification is important even though it is not the main purpose of the test because for the test to be a test of network degradation with high fidelity, it is necessary for it to be able to classify the network to the ground truth with high power. Additionally, the test also acts as a test for network growth as a function of the m parameter since the test is able to detect with high power if the network's smallest degree, m, has increased.

A sensitivity analysis with respect to network degradation in terms of edge and node deletion was conducted on the test on L-moments. Overall, the test performs very well for edge deletion when either the low or high degree nodes are affected for networks with $m \geq 2$, and the test also performs well for node deletion when either the medium or high degree nodes are affected for networks with $m \geq 2$ with size $k \geq 9$. For edge deletion, it seems that the test performance drops if the proportion of deletion p is in the neighborhood of 0.2 when only the medium degree nodes are affected, but the performance then improves as p increases. For node deletion, the overall performance of the test increases gradually as a function of p, but it is very apparent that the test is practically unusable for smaller networks when only the high degree nodes are affected. This is attributable to the fact that there are only a few number of nodes with high degree for smaller networks and if the proportion of deletion p is small, then it is likely that none of the nodes are deleted at all. These results could be explained when looking at the statistics on the deletion processes

for the various conditions where although edge deletion seems to create some isolates for almost all combinations of degree level, network size, and proportion of deletion, node deletion seems to have only created some isolates on medium and high degree on larger networks with high proportion of deletion. Both degradation processes only affected medium and high degree level with some varying effect with respect to number of components created, but in terms of the degree distribution, only isolates would have caused a drastic change where essentially those proportion of the density are taken out and redistributed thus changing the shape of the distribution. This may explain why the test was more successful for edge deletion and not as much for node deletion.

6.1 Conclusion

As previously discussed, many real world network possess the scale-free property where the degree distribution follows a power law with varying values of β. There is an important implication to the findings in this dissertation. If our results hold and the the value of β is truly dependent on the m^* parameter and is not $\beta = 2$ as proposed, then the degree distribution may be used to model any network that is believed to be scale-free. One can then generate a proxy of the real network by using the appropriate parameter m^* of the Barabási-Albert model that is associated with a given β for the given network size n. A direct application of this finding is in Social Network Analysis where a study by Blaha and others [10] suggested that the most appropriate technique selection for visualizing a network is dependent on various factors such as the network model and the task that is to be performed on the given network. Therefore, if a real network could be characterized to the closest proxy model, then the most appropriate visualization technique that gives the best insight to the structure of the network should be used. For example, if a network

can be characterized by its exponent, β, then the visualization that works best for the corresponding Barabási-Albert model is preferred. Additionally, the real network could also be characterized using the Barabási-Albert model, and hypotheses could be tested on whether or not the network is changing.

Although the real world network classification showed that only the sub-network of two networks can be classified as Barabási-Albert networks, the result is useful nonetheless. It is now possible, at least on the sub-network level, to simulate a network having the same properties as the Barabási-Albert network using the Barabási-Albert network itself. One can then compare the real network and the simulated network through visual comparison or inspection by an analyst, as a form of turing test, to see if the simulated network is a good representation of the real network. Since the simulated network is based on the Barabási-Albert model itself, the two are mathematically similar, so a human visual comparison will be able to complement the similarity provided by the model. Further, being able to examine the structure of the sub-networks implies a way to gain deeper understanding of the main structure of the network without the periphery nodes. Although our test is a good way of comparing the structure of the networks, there might be differences in other aspects of the network, through its visual representation, that could still differentiate the networks. It is now also possible to adopt the result obtained by Blaha and others [10] to visually display the real network based on what is known about the Barabási-Albert network that might give more insight to the real network. Such ability will be very useful especially for the intelligence community where an analyst could gain more insight on the network.

Additionally, the derived test can also be used in real world application to detect possible synthetic networks that are fabricated using a linear preferential attachment model. This idea is akin to the test for the Erdös-Rényi network as described by

Wasserman and Faust [95], except now we are able to not only test whether or not a given network is truly random as modeled by the Erdös-Rényi network, but also whether or not it is an artificial network that was built on the preferential attachment model. Although the motivation for the research was to find a method of linking real world network to a model proxy so that knowledge of the model can be applied to the network for various tasks such as network simulation, being able to detect whether or not a given network is artificial is useful in application such as community detection. Should it prove to be that real world scale-free networks are inherently different than the Barabási-Albert network, then the proposed test can still be used to differentiate whether or not an anonymous network is a real world network or an artificial pseudo-random network of the Barabási-Albert model. In which case, the next step will be to compare whether the method can distinguish between a real world scale-free network and a synthetic scale-free network better than a human analyst. The approach can also be applied to other types of network model with known degree distribution for network characterization.

It is shown that edge deletion affects the network more so than node deletion with respect to the number of isolates and components caused and to changes in clustering. Thus, if the objective is to affect a network in such a way that it changes the network's characteristic, then it is suggested that the focus be placed on the connectors (i.e. edges) within said network as oppose to the entities of the network (i.e. nodes). On the opposite end, however, if the objective is to detect whether or not the network has degraded, then the trivariate test for change detection using $(\lambda_2, \tau_3, \tau_4)$ is able to detect the degradation caused by edge deletion with high power. Recall again that degradation is defined as removal of nodes or edges within the network that changes the structure of the network and its degree distribution.

Although the test on the *Pareto* distribution is specific to the Barabási-Albert

network, the test on the L-moments as designed for monitoring network degradation, on the other hand, is not constrained to that particular model since it is based on the empirical L-moments of the network in question. Even though the Barabási-Albert network is the sole focus of this research, the method can be applied to other network models such as the ones listed in Section 2.1.4. However, in order to apply the test to real world application, a method of characterizing a network in question to a suitable model proxy is required and to our knowledge there are no formal test of hypothesis for characterizing any of the other models that are listed. Once this is established, then the properties associated with the model proxy such as the scale-free and small-world properties can directly be linked to the real network in question and the network can be monitored using the test on L-moments to detect if its structure has degraded to the point that it loses its original properties.

Another advantage to the approach taken for the test on L-moments is that it is also not restricted to the degree distribution of the network. As listed in Section 2.1.2, there are other nodal measures available in the literature that we have not considered. Thus, the same approach can be taken using the listed measures by characterizing the networks using the L-moments from the empirical distribution of these other measures within a given network model. These other measures capture different properties of the network that degree distribution does not and may be able to give a different insight into the the changes within the network. A comprehensive look into all of the other network models and measures is necessary in order to accomplish the goal of creating a network visualization tool that can assist in gaining insight for network analysis which is the overarching motivation for this dissertation.

Appendix A. Power of the Hypotheses Tests

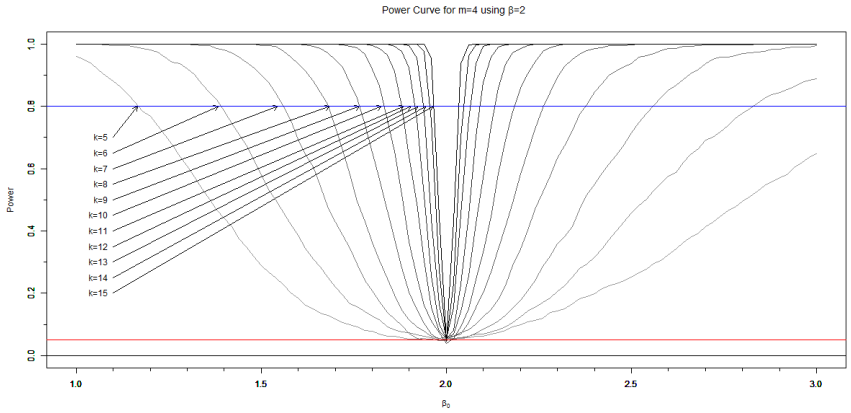


Figure A.1. Power curve for the test on β for $m^* = 4$. Note: Lighter shaded lines indicates smaller k

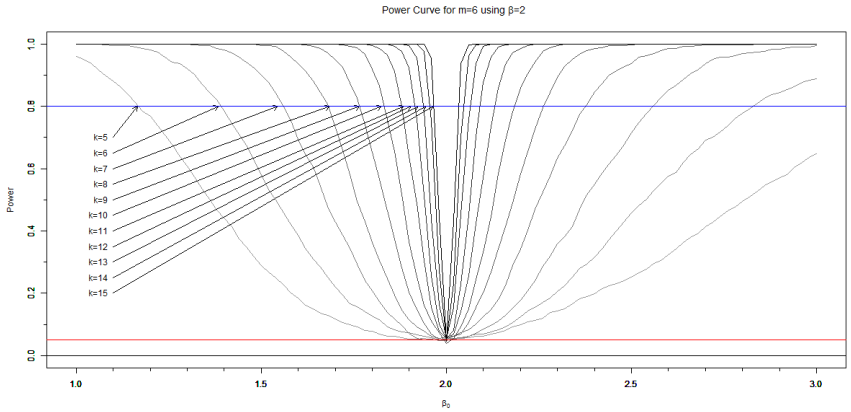


Figure A.2. Power curve for the test on β for $m^* = 6$. Note: Lighter shaded lines indicates smaller k

Table A.1. Power of the test on Pareto for $m_1 = m^*\sqrt{\frac{(n-m^*)}{n}}$ where $\delta = m_1 - m^*$. ϵ is smallest $|\delta|$.

$m^* = 1$						k					
δ	5	6	7	8	9	10	11	12	13	14	15
ϵ	.057	.077	.080	.025	.012	.015	.020	.022	.019	.025	.013
.005	.076	.133	.280	.265	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.01	.111	.253	**.877**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.02	.182	**.901**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.04	.689	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.06	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 2$						k					
δ	5	6	7	8	9	10	11	12	13	14	15
ϵ	.064	.053	.053	.096	.076	.007	.010	.006	.008	.007	.004
.005	.076	.074	.105	.354	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.01	.087	.100	.195	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.02	.123	.180	.633	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.04	.204	.654	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.06	.416	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	**.809**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 4$						k					
δ	5	6	7	8	9	10	11	12	13	14	15
ϵ	.062	.055	.067	.051	.034	.160	.184	.000	.002	.000	.000
.005	.069	.067	.094	.088	.154	**1**	**1**	**1**	**1**	**1**	**1**
.01	.075	.077	.126	.173	.545	**1**	**1**	**1**	**1**	**1**	**1**
.02	.087	.105	.235	.577	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.04	.124	.187	.796	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.06	.154	.365	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	.211	.696	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	.314	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 6$						k					
δ	5	6	7	8	9	10	11	12	13	14	15
ϵ	.058	.060	.052	.054	.041	.122	.144	.124	.001	.000	.000
.005	.062	.069	.066	.081	.103	.622	**1**	**1**	**1**	**1**	**1**
.01	.066	.077	.083	.128	.279	**1**	**1**	**1**	**1**	**1**	**1**
.02	.076	.095	.126	.291	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.04	.092	.136	.297	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.06	.121	.212	.651	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	.138	.334	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	.167	.505	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.2	.563	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Table A.2. Power of the UIT on Pareto for $m_1 = m^*\sqrt{\frac{(n-m^*)}{n}}$ and $\beta_0 = 2$.

	$\delta_\beta = .02$					$\delta_\beta = .16$					$\delta_\beta = .32$				
δ_m	k					k					k				
$m^* = 1$	5	6	7	8	9	5	6	7	8	9	5	6	7	8	9
.01	.11	.25	**.88**	**1**	**1**	.11	.25	**.88**	**1**	**1**	.27	.53	**.88**	**1**	**1**
.02	.18	**.90**	**1**	**1**	**1**	.18	**.90**	**1**	**1**	**1**	.27	**.90**	**1**	**1**	**1**
.04	.69	**1**	**1**	**1**	**1**	.67	**1**	**1**	**1**	**1**	.69	**1**	**1**	**1**	**1**
.06	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 2$															
.01	.09	.10	.20	**1**	**1**	.09	.10	.20	**1**	**1**	.27	.53	**.88**	**1**	**1**
.02	.12	.18	.63	**1**	**1**	.12	.18	.63	**1**	**1**	.27	.53	**.88**	**1**	**1**
.04	.20	.65	**1**	**1**	**1**	.20	.65	**1**	**1**	**1**	.27	.65	**1**	**1**	**1**
.06	.42	**1**	**1**	**1**	**1**	.42	**1**	**1**	**1**	**1**	.42	**1**	**1**	**1**	**1**
.08	**.81**	**1**	**1**	**1**	**1**	**.81**	**1**	**1**	**1**	**1**	**.81**	**1**	**1**	**1**	**1**
.1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 4$															
.01	.08	.08	.13	.17	.55	.08	.09	.14	.25	.55	.27	.53	**.88**	**1**	**1**
.02	.09	.11	.24	.58	**1**	.09	.11	.24	.58	**1**	.27	.53	**.88**	**1**	**1**
.04	.12	.19	**.80**	**1**	**1**	.12	.19	**.80**	**1**	**1**	.27	.53	**.88**	**1**	**1**
.06	.15	.37	**1**	**1**	**1**	.15	.37	**1**	**1**	**1**	.27	.53	**1**	**1**	**1**
.08	.21	.67	**1**	**1**	**1**	.21	.67	**1**	**1**	**1**	.27	.67	**1**	**1**	**1**
.1	.31	**1**	**1**	**1**	**1**	.31	**1**	**1**	**1**	**1**	.31	**1**	**1**	**1**	**1**
$m^* = 6$															
.01	.07	.08	.08	.13	.28	.07	.09	.14	.25	.46	.27	.53	**.88**	**1**	**1**
.02	.08	.10	.13	.29	**1**	.08	.10	.14	.29	**1**	.27	.53	**.88**	**1**	**1**
.04	.09	.14	.30	**1**	**1**	.09	.14	.30	**1**	**1**	.27	.53	**.88**	**1**	**1**
.06	.12	.21	.65	**1**	**1**	.12	.21	.65	**1**	**1**	.27	.53	**.88**	**1**	**1**
.08	.14	.33	**1**	**1**	**1**	.14	.33	**1**	**1**	**1**	.27	.53	**1**	**1**	**1**
.1	.17	.51	**1**	**1**	**1**	.17	.51	**1**	**1**	**1**	.27	.53	**1**	**1**	**1**

Table A.3. Power of the UIT on Barabási-Albert for $m_1 = m^*$ and $\beta_0 = 2$.

	$\delta_\beta = .02$					$\delta_\beta = .10$					$\delta_\beta = .30$				
δ_m	k					k					k				
$m^* = 1$	5	6	7	8	9	5	6	7	8	9	5	6	7	8	9
.01	.06	.06	.05	**1**	**1**	.06	.10	.19	**1**	**1**	**.85**	**.96**	**1**	**1**	**1**
.02	.06	.06	**1**	**1**	**1**	.06	.10	**1**	**1**	**1**	**.85**	**.96**	**1**	**1**	**1**
.04	.06	**1**	**1**	**1**	**1**	.06	**1**	**1**	**1**	**1**	**.85**	**1**	**1**	**1**	**1**
.06	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 2$															
.01	.05	.07	.06	.07	**1**	.17	.23	.40	.63	**1**	**.85**	**.96**	**1**	**1**	**1**
.02	.05	.07	.06	**1**	**1**	.17	.23	.40	**1**	**1**	**.85**	**.96**	**1**	**1**	**1**
.04	.05	.07	**1**	**1**	**1**	.17	.23	**1**	**1**	**1**	**.85**	**.96**	**1**	**1**	**1**
.06	.05	**1**	**1**	**1**	**1**	.17	**1**	**1**	**1**	**1**	**.85**	**1**	**1**	**1**	**1**
.08	.05	**1**	**1**	**1**	**1**	.17	**1**	**1**	**1**	**1**	**.85**	**1**	**1**	**1**	**1**
.1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 4$															
.01	.07	.07	.08	.10	.14	.45	.57	.77	**.91**	**.99**	**1**	**1**	**1**	**1**	**1**
.02	.07	.07	.08	.10	**1**	.45	.57	.77	**.91**	**1**	**1**	**1**	**1**	**1**	**1**
.04	.07	.07	.08	**1**	**1**	.45	.57	.77	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.06	.07	.07	**1**	**1**	**1**	.45	.57	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	.07	.07	**1**	**1**	**1**	.45	.57	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	.07	**1**	**1**	**1**	**1**	.45	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$m^* = 6$															
.01	.07	.08	.11	.13	.20	.79	**.82**	**.92**	**.99**	**1**	**1**	**1**	**1**	**1**	**1**
.02	.07	.08	.11	.13	**1**	.79	**.82**	**.92**	**.99**	**1**	**1**	**1**	**1**	**1**	**1**
.04	.07	.08	.11	**1**	**1**	.79	**.82**	**.92**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.06	.07	.08	.11	**1**	**1**	.79	**.82**	**.92**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.08	.07	.08	**1**	**1**	**1**	.79	**.82**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
.1	.07	.08	**1**	**1**	**1**	.79	**.82**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Appendix B. Mean and Covariance Estimate of (τ_3, τ_4)

Table B.4. Mean and Covariance estimates for (τ_3, τ_4) based on bivariate normal assumption.

m	k	μ_{λ_2}	μ_{τ_3}	μ_{τ_4}	σ_{λ_2}	σ_{τ_3}	σ_{τ_4}	$\sigma_{\lambda_2,\tau_3}$	$\sigma_{\lambda_2,\tau_4}$	σ_{τ_3,τ_4}
1	5	7.15E-01	5.73E-01	2.68E-01	2.16E-03	7.24E-03	1.20E-02	4.06E-03	3.84E-03	8.52E-03
1	6	7.42E-01	5.88E-01	2.93E-01	1.23E-03	3.75E-03	6.19E-03	2.30E-03	2.10E-03	4.50E-03
1	7	7.60E-01	6.03E-01	3.17E-01	5.23E-04	1.54E-03	2.75E-03	9.97E-04	8.74E-04	1.93E-03
1	8	7.68E-01	6.08E-01	3.27E-01	2.86E-04	8.17E-04	1.47E-03	5.39E-04	4.70E-04	1.03E-03
1	9	7.72E-01	6.12E-01	3.33E-01	1.31E-04	3.61E-04	6.30E-04	2.39E-04	2.12E-04	4.46E-04
1	10	7.75E-01	6.14E-01	3.37E-01	6.59E-05	1.86E-04	3.48E-04	1.25E-04	1.08E-04	2.39E-04
1	11	7.76E-01	6.16E-01	3.39E-01	3.72E-05	1.04E-04	1.86E-04	7.06E-05	6.05E-05	1.31E-04
1	12	7.77E-01	6.16E-01	3.40E-01	1.83E-05	4.89E-05	8.44E-05	3.27E-05	2.91E-05	6.01E-05
1	13	7.77E-01	6.16E-01	3.41E-01	8.27E-06	2.23E-05	3.95E-05	1.50E-05	1.32E-05	2.78E-05
1	14	7.77E-01	6.16E-01	3.41E-01	4.27E-06	1.19E-05	2.20E-05	8.03E-06	6.91E-06	1.52E-05
1	15	7.77E-01	6.16E-01	3.41E-01	2.32E-06	6.20E-06	1.08E-05	4.13E-06	3.68E-06	7.65E-06
2	5	1.26E+00	4.79E-01	1.96E-01	6.07E-03	4.78E-03	6.99E-03	3.38E-03	4.88E-03	4.76E-03
2	6	1.35E+00	5.23E-01	2.56E-01	3.46E-03	2.34E-03	3.76E-03	2.08E-03	2.59E-03	2.55E-03
2	7	1.41E+00	5.53E-01	2.99E-01	1.74E-03	1.05E-03	1.76E-03	1.05E-03	1.24E-03	1.19E-03
2	8	1.44E+00	5.68E-01	3.23E-01	8.18E-04	4.98E-04	8.56E-04	5.42E-04	5.88E-04	5.81E-04
2	9	1.46E+00	5.78E-01	3.38E-01	4.24E-04	2.56E-04	4.70E-04	2.85E-04	3.01E-04	3.10E-04
2	10	1.47E+00	5.83E-01	3.47E-01	2.20E-04	1.34E-04	2.45E-04	1.52E-04	1.57E-04	1.62E-04
2	11	1.47E+00	5.86E-01	3.51E-01	1.11E-04	6.32E-05	1.15E-04	7.07E-05	7.65E-05	7.55E-05
2	12	1.48E+00	5.87E-01	3.53E-01	5.72E-05	3.39E-05	6.25E-05	3.86E-05	4.03E-05	4.12E-05
2	13	1.48E+00	5.88E-01	3.55E-01	2.80E-05	1.58E-05	2.91E-05	1.74E-05	1.92E-05	1.89E-05
2	14	1.48E+00	5.88E-01	3.55E-01	1.31E-05	8.19E-06	1.63E-05	9.24E-06	9.37E-06	1.04E-05
2	15	1.48E+00	5.89E-01	3.56E-01	6.43E-06	3.93E-06	7.49E-06	4.52E-06	4.60E-06	4.87E-06
3	5	1.69E+00	4.11E-01	1.43E-01	9.30E-03	3.59E-03	4.97E-03	2.52E-03	4.81E-03	3.39E-03
3	6	1.90E+00	4.84E-01	2.29E-01	5.26E-03	1.58E-03	2.62E-03	1.59E-03	2.41E-03	1.71E-03
3	7	2.03E+00	5.28E-01	2.86E-01	2.79E-03	7.46E-04	1.35E-03	8.80E-04	1.21E-03	8.60E-04
3	8	2.09E+00	5.53E-01	3.24E-01	1.42E-03	3.80E-04	7.17E-04	5.14E-04	6.24E-04	4.58E-04
3	9	2.12E+00	5.66E-01	3.45E-01	7.47E-04	2.00E-04	3.92E-04	2.89E-04	3.30E-04	2.49E-04
3	10	2.14E+00	5.74E-01	3.58E-01	3.88E-04	9.59E-05	1.80E-04	1.42E-04	1.67E-04	1.16E-04
3	11	2.15E+00	5.79E-01	3.65E-01	1.76E-04	4.31E-05	8.60E-05	6.18E-05	7.40E-05	5.32E-05
3	12	2.16E+00	5.81E-01	3.69E-01	9.02E-05	2.25E-05	4.23E-05	3.42E-05	3.89E-05	2.74E-05
3	13	2.16E+00	5.82E-01	3.71E-01	4.65E-05	1.17E-05	2.35E-05	1.75E-05	1.99E-05	1.47E-05
3	14	2.16E+00	5.83E-01	3.72E-01	2.34E-05	5.61E-06	1.10E-05	8.35E-06	9.76E-06	6.91E-06
3	15	2.16E+00	5.83E-01	3.73E-01	1.18E-05	2.86E-06	5.60E-06	4.30E-06	4.97E-06	3.53E-06
4	5	2.05E+00	3.55E-01	9.69E-02	1.26E-02	3.03E-03	3.38E-03	2.05E-03	4.93E-03	2.50E-03
4	6	2.40E+00	4.52E-01	1.98E-01	7.83E-03	1.42E-03	2.15E-03	1.63E-03	2.66E-03	1.47E-03
4	7	2.60E+00	5.08E-01	2.70E-01	3.76E-03	6.00E-04	1.11E-03	7.74E-04	1.17E-03	6.97E-04
4	8	2.72E+00	5.43E-01	3.20E-01	2.08E-03	3.14E-04	5.81E-04	5.45E-04	6.67E-04	3.77E-04
4	9	2.77E+00	5.60E-01	3.48E-01	1.01E-03	1.43E-04	2.73E-04	2.55E-04	3.10E-04	1.74E-04
4	10	2.81E+00	5.71E-01	3.64E-01	5.49E-04	7.63E-05	1.46E-04	1.44E-04	1.70E-04	9.34E-05
4	11	2.82E+00	5.76E-01	3.74E-01	2.87E-04	3.99E-05	8.08E-05	7.37E-05	8.73E-05	5.04E-05
4	12	2.83E+00	5.80E-01	3.79E-01	1.32E-04	1.93E-05	3.97E-05	3.69E-05	4.13E-05	2.48E-05
4	13	2.84E+00	5.81E-01	3.82E-01	7.24E-05	1.01E-05	1.99E-05	1.99E-05	2.25E-05	1.27E-05
4	14	2.84E+00	5.82E-01	3.83E-01	3.43E-05	4.53E-06	9.09E-06	8.66E-06	1.03E-05	5.64E-06
4	15	2.84E+00	5.83E-01	3.84E-01	1.74E-05	2.44E-06	4.97E-06	4.51E-06	5.30E-06	3.10E-06

Table B.5. Mean and Covariance estimates for (τ_3, τ_4) based on bivariate normal assumption.

m	k	μ_{λ_2}	μ_{τ_3}	μ_{τ_4}	σ_{λ_2}	σ_{τ_3}	σ_{τ_4}	$\sigma_{\lambda_2,\tau_3}$	$\sigma_{\lambda_2,\tau_4}$	σ_{τ_3,τ_4}
5	5	2.33E+00	3.05E-01	5.81E-02	1.48E-02	2.45E-03	2.71E-03	1.01E-03	4.42E-03	1.85E-03
5	6	2.85E+00	4.22E-01	1.68E-01	9.48E-03	1.16E-03	1.65E-03	1.28E-03	2.53E-03	1.13E-03
5	7	3.15E+00	4.92E-01	2.54E-01	4.68E-03	4.81E-04	8.68E-04	7.03E-04	1.13E-03	5.47E-04
5	8	3.32E+00	5.33E-01	3.12E-01	2.93E-03	2.59E-04	4.72E-04	4.91E-04	6.85E-04	3.01E-04
5	9	3.42E+00	5.56E-01	3.46E-01	1.44E-03	1.33E-04	2.61E-04	2.69E-04	3.40E-04	1.64E-04
5	10	3.46E+00	5.69E-01	3.68E-01	7.73E-04	7.12E-05	1.33E-04	1.61E-04	1.91E-04	8.65E-05
5	11	3.49E+00	5.75E-01	3.79E-01	3.65E-04	3.10E-05	6.28E-05	6.69E-05	8.31E-05	3.89E-05
5	12	3.50E+00	5.79E-01	3.85E-01	1.81E-04	1.57E-05	3.07E-05	3.44E-05	4.23E-05	1.93E-05
5	13	3.51E+00	5.81E-01	3.89E-01	8.33E-05	7.85E-06	1.66E-05	1.69E-05	1.97E-05	1.02E-05
5	14	3.51E+00	5.82E-01	3.91E-01	4.88E-05	4.18E-06	8.28E-06	9.55E-06	1.14E-05	5.22E-06
5	15	3.51E+00	5.83E-01	3.92E-01	2.28E-05	2.03E-06	4.13E-06	4.51E-06	5.35E-06	2.57E-06
6	5	2.56E+00	2.57E-01	2.72E-02	1.74E-02	2.30E-03	2.07E-03	6.71E-04	4.56E-03	1.46E-03
6	6	3.26E+00	3.92E-01	1.37E-01	1.07E-02	1.01E-03	1.40E-03	1.11E-03	2.40E-03	9.74E-04
6	7	3.68E+00	4.77E-01	2.37E-01	5.97E-03	4.76E-04	7.78E-04	8.78E-04	1.29E-03	5.26E-04
6	8	3.91E+00	5.23E-01	3.02E-01	3.45E-03	2.43E-04	4.48E-04	5.40E-04	6.99E-04	2.91E-04
6	9	4.04E+00	5.51E-01	3.43E-01	1.79E-03	1.13E-04	2.01E-04	2.84E-04	3.52E-04	1.34E-04
6	10	4.11E+00	5.67E-01	3.68E-01	8.67E-04	5.68E-05	1.09E-04	1.39E-04	1.70E-04	7.00E-05
6	11	4.15E+00	5.75E-01	3.81E-01	4.16E-04	2.84E-05	5.88E-05	7.17E-05	8.18E-05	3.68E-05
6	12	4.17E+00	5.80E-01	3.89E-01	2.08E-04	1.36E-05	2.81E-05	3.40E-05	4.00E-05	1.75E-05
6	13	4.18E+00	5.82E-01	3.93E-01	1.16E-04	7.24E-06	1.42E-05	1.86E-05	2.23E-05	9.03E-06
6	14	4.18E+00	5.83E-01	3.96E-01	5.52E-05	3.46E-06	6.76E-06	9.03E-06	1.06E-05	4.35E-06
6	15	4.19E+00	5.84E-01	3.97E-01	3.02E-05	1.84E-06	3.89E-06	4.78E-06	5.66E-06	2.39E-06
7	5	2.73E+00	2.12E-01	-9.54E-04	1.90E-02	2.19E-03	1.81E-03	3.87E-04	4.67E-03	1.23E-03
7	6	3.63E+00	3.65E-01	1.12E-01	1.23E-02	9.26E-04	1.17E-03	1.01E-03	2.43E-03	8.29E-04
7	7	4.18E+00	4.61E-01	2.18E-01	7.39E-03	4.44E-04	6.91E-04	9.07E-04	1.37E-03	4.75E-04
7	8	4.50E+00	5.16E-01	2.93E-01	3.68E-03	1.83E-04	3.17E-04	4.19E-04	6.11E-04	2.07E-04
7	9	4.67E+00	5.47E-01	3.39E-01	2.00E-03	9.69E-05	1.83E-04	2.63E-04	3.31E-04	1.17E-04
7	10	4.76E+00	5.65E-01	3.67E-01	1.11E-03	4.80E-05	9.18E-05	1.31E-04	1.73E-04	5.80E-05
7	11	4.81E+00	5.75E-01	3.83E-01	5.40E-04	2.64E-05	5.24E-05	7.92E-05	9.02E-05	3.35E-05
7	12	4.83E+00	5.80E-01	3.92E-01	2.52E-04	1.09E-05	2.25E-05	2.97E-05	3.83E-05	1.38E-05
7	13	4.84E+00	5.83E-01	3.97E-01	1.32E-04	6.24E-06	1.26E-05	1.82E-05	2.14E-05	7.91E-06
7	14	4.85E+00	5.84E-01	3.99E-01	6.86E-05	3.16E-06	6.38E-06	9.96E-06	1.14E-05	4.00E-06
7	15	4.86E+00	5.85E-01	4.01E-01	3.55E-05	1.55E-06	3.15E-06	4.35E-06	5.52E-06	1.94E-06

Appendix C. Statistics and Plots for Degradation Detection

Table C.6. Edges affected and edges deleted from edge deletion

m	k	Degree Level	Edges Affected 95% CI	Deleted 95% CI $(p = 0.01), (p = 0.5)$	
1	5	low	$(15, 22)$	$(0, 0)$	$(7, 11)$
		medium	$(0, 23)$	$(0, 0)$	$(0, 11)$
		high	$(6, 17)$	$(0, 0)$	$(3, 8)$
	14	low	$(9752, 9913)$	$(97, 99)$	$(4876, 4956)$
		medium	$(9752, 9913)$	$(97, 99)$	$(4876, 4956)$
		high	$(2854, 3322)$	$(28, 33)$	$(1426, 1661)$
2	5	low	$(20, 34)$	$(0, 0)$	$(10, 17)$
		medium	$(0, 36)$	$(0, 0)$	$(0, 18)$
		high	$(10, 26)$	$(0, 0)$	$(5, 13)$
	14	low	$(14716, 15068)$	$(147, 150)$	$(7358, 1534)$
		medium	$(10041, 10605)$	$(100, 106)$	$(5020, 5302)$
		high	$(5868, 6464)$	$(58, 64)$	$(2934, 3232)$
3	5	low	$(24, 42)$	$(0, 0)$	$(12, 21)$
		medium	$(0, 24)$	$(0, 0)$	$(0, 20)$
		high	$(13, 34)$	$(0, 0)$	$(6, 17)$
	14	low	$(17868, 18111)$	$(178, 183)$	$(8934, 9172)$
		medium	$(12772, 13540)$	$(127, 135)$	$(6386, 6770)$
		high	$(8952, 9662)$	$(89, 96)$	$(4476, 4831)$
4	5	low	$(24, 48)$	$(0, 0)$	$(12, 24)$
		medium	$(0, 49)$	$(0, 0)$	$(0, 24)$
		high	$(15, 40)$	$(0, 0)$	$(7, 19)$
	14	low	$(20028, 20656)$	$(200, 206)$	$(10014, 10328)$
		medium	$(11652, 15780)$	$(116, 157)$	$(5826, 7890)$
		high	$(12057, 12884)$	$(120, 128)$	$(6028, 6442)$
5	5	low	$(25, 55)$	$(0, 0)$	$(12, 27)$
		medium	$(0, 49)$	$(0, 0)$	$(0, 24)$
		high	$(17, 51)$	$(0, 0)$	$(8, 25)$
	14	low	$(21575, 22360)$	$(215, 223)$	$(4876, 4956)$
		medium	$(13244, 14322)$	$(132, 143)$	$(6622, 7161)$
		high	$(15165, 16125)$	$(151, 161)$	$(7582, 8062)$
6	5	low	$(24, 60)$	$(0, 0)$	$(12, 30)$
		medium	$(0, 56)$	$(0, 0)$	$(0, 28)$
		high	$(19, 60)$	$(0, 0)$	$(9, 30)$
	14	low	$(22800, 23664)$	$(228, 236)$	$(11400, 11832)$
		medium	$(14640, 15904)$	$(146, 159)$	$(7320, 7952)$
		high	$(18283, 19327)$	$(182, 193)$	$(9141, 9663)$
7	5	low	$(21, 56)$	$(0, 0)$	$(10, 28)$
		medium	$(0, 60)$	$(0, 0)$	$(0, 30)$
		high	$(21, 63)$	$(0, 0)$	$(10, 31)$
	14	low	$(23765, 24710)$	$(237, 247)$	$(11882, 12355)$
		medium	$(13310, 14690)$	$(133, 146)$	$(6655, 7345)$
		high	$(21415, 22489)$	$(214, 224)$	$(10707, 11244)$

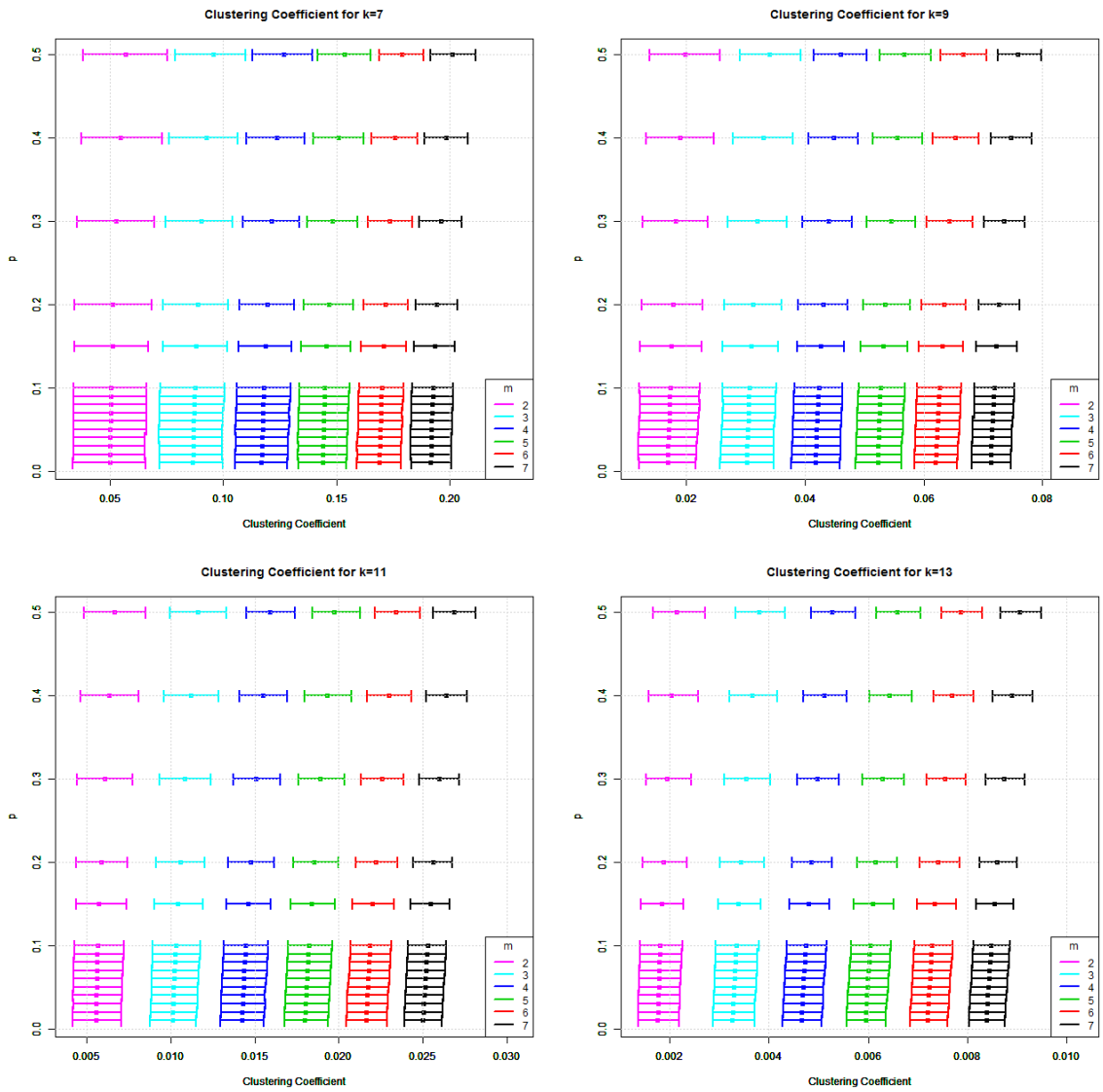


Figure C.3. Clustering coefficients of networks after edge deletion on high degrees. (Bars are 95% CI.)

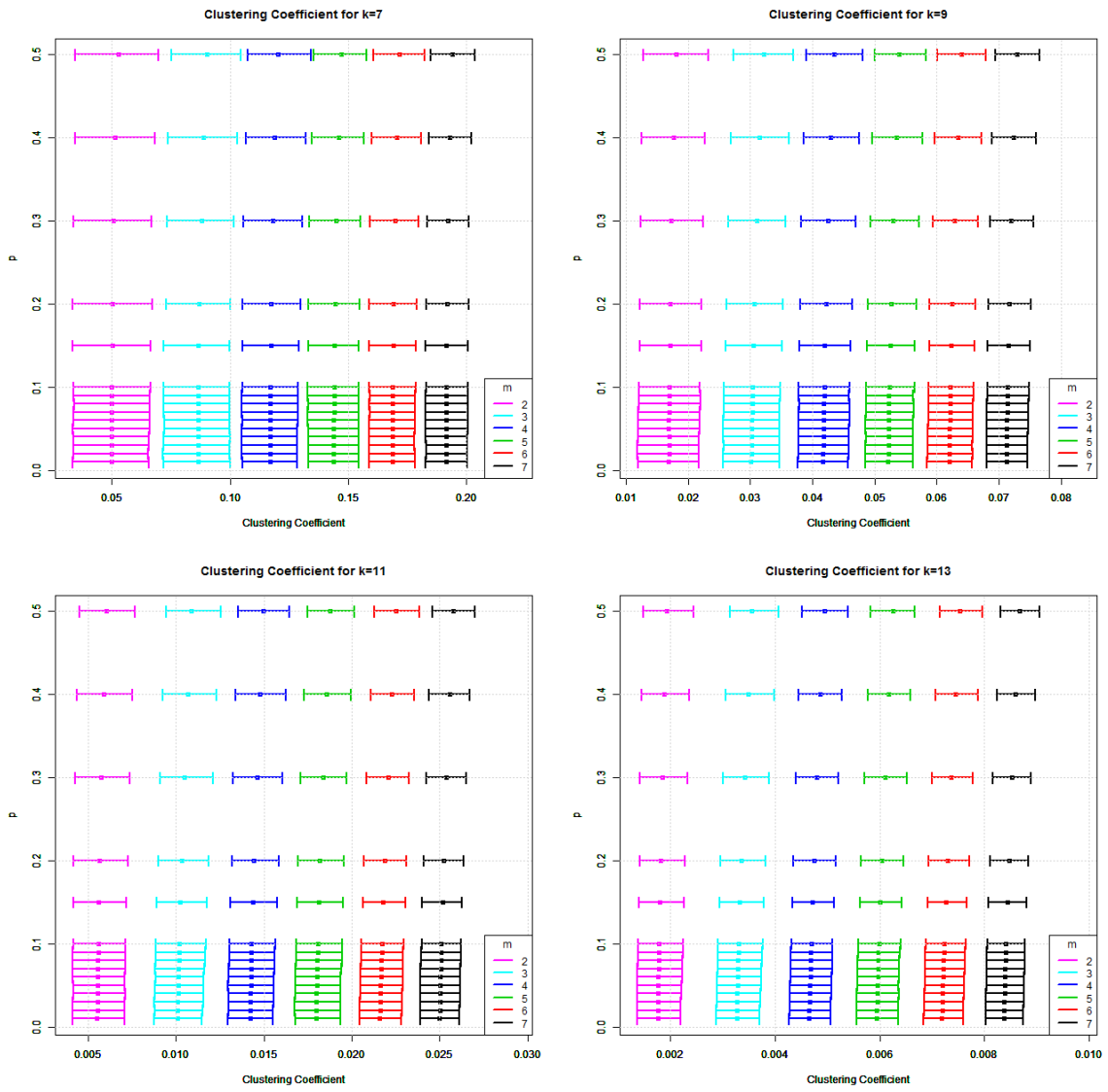


Figure C.4. Clustering coefficients of networks after edge deletion on high degrees. (Bars are 95% CI.)

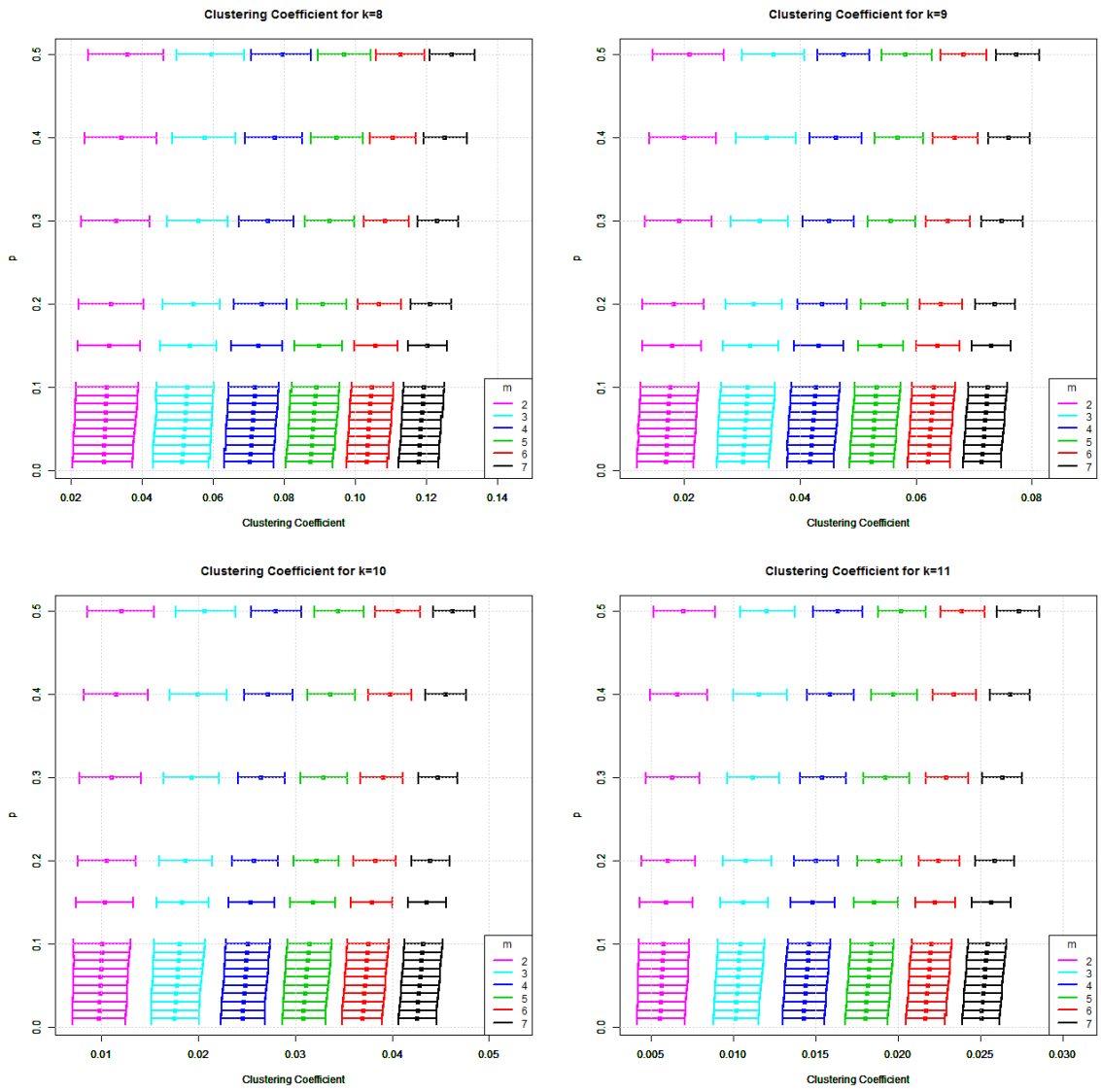


Figure C.5. Clustering coefficients of networks after node deletion on low degrees. (Bars are 95% CI.)

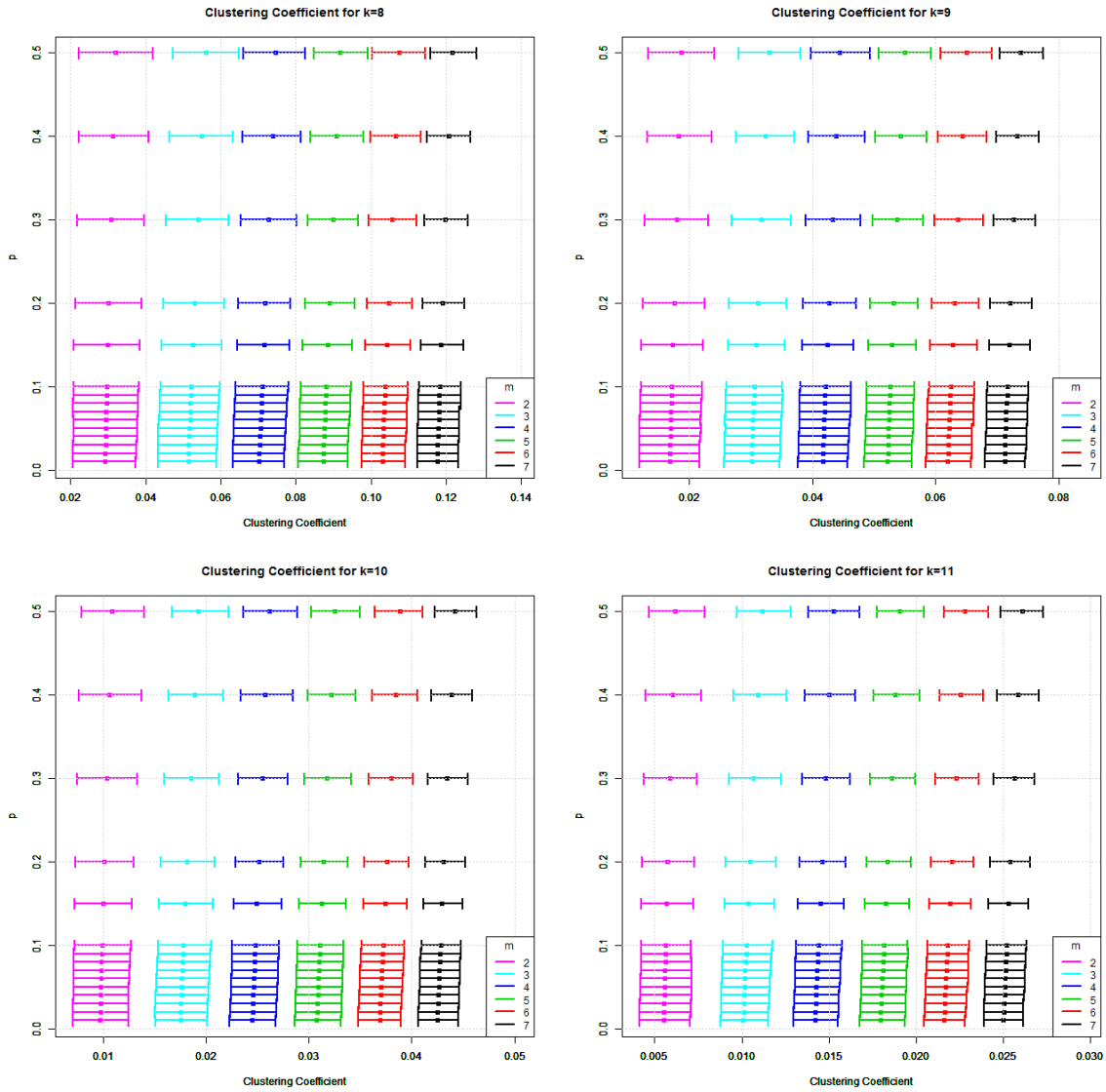


Figure C.6. Clustering coefficients of networks after node deletion on medium degrees. (Bars are 95% CI.)

Table C.7. Nodes affected and edges deleted from node deletion

m	k	Degree Level	Nodes Affected 95% CI	Deleted 95% CI $(p = 0.01), (p = 0.5)$	
1	5	low	$(15, 22)$	$(0, 0)$	$(7, 11)$
		medium	$(0, 22)$	$(0, 0)$	$(0, 11)$
		high	$(1, 3)$	$(0, 0)$	$(0, 8)$
	14	low	$(9752, 9913)$	$(97, 99)$	$(4876, 4956)$
		medium	$(9752, 9913)$	$(97, 99)$	$(4876, 4956)$
		high	$(166, 207)$	$(11, 47)$	$(1362, 1709)$
2	5	low	$(10, 17)$	$(0, 0)$	$(10, 16)$
		medium	$(0, 17)$	$(0, 0)$	$(0, 18)$
		high	$(1, 2)$	$(0, 0)$	$(0, 12)$
	14	low	$(7358, 7534)$	$(146, 150)$	$(7358, 7534)$
		medium	$(3347, 3535)$	$(99, 105)$	$(5019, 5301)$
		high	$(164, 187)$	$(21, 113)$	$(2749, 3410)$
3	5	low	$(8, 14)$	$(0, 0)$	$(12, 21)$
		medium	$(0, 10)$	$(0, 0)$	$(0, 20)$
		high	$(1, 2)$	$(0, 0)$	$(0, 15)$
	14	low	$(5956, 6115)$	$(177, 183)$	$(8934, 9171)$
		medium	$(3193, 3385)$	$(124, 134)$	$(6384, 6768)$
		high	$(164, 179)$	$(31, 162)$	$(4159, 5117)$
4	5	low	$(6, 12)$	$(0, 0)$	$(12, 24)$
		medium	$(0, 9)$	$(0, 0)$	$(0, 24)$
		high	$(1, 3)$	$(0, 0)$	$(0, 18)$
	14	low	$(5007, 5164)$	$(200, 204)$	$(10012, 10328)$
		medium	$(1942, 3156)$	$(114, 155)$	$(5826, 7890)$
		high	$(164, 176)$	$(42, 221)$	$(5556, 6834)$
5	5	low	$(5, 11)$	$(0, 0)$	$(10, 25)$
		medium	$(0, 7)$	$(0, 0)$	$(0, 24)$
		high	$(1, 3)$	$(0, 0)$	$(0, 20)$
	14	low	$(4315, 4472)$	$(215, 220)$	$(10785, 11180)$
		medium	$(1892, 2046)$	$(126, 140)$	$(6622, 7161)$
		high	$(164, 173)$	$(52, 303)$	$(7054, 8532)$
6	5	low	$(4, 10)$	$(0, 0)$	$(12, 30)$
		medium	$(0, 6)$	$(0, 0)$	$(0, 27)$
		high	$(1, 3)$	$(0, 0)$	$(0, 22)$
	14	low	$(3800, 3944)$	$(228, 234)$	$(11400, 11832)$
		medium	$(1830, 1988)$	$(144, 152)$	$(7320, 7952)$
		high	$(164, 172)$	$(62, 332)$	$(8440, 10320)$
7	5	low	$(3, 8)$	$(0, 0)$	$(7, 28)$
		medium	$(0, 6)$	$(0, 0)$	$(0, 30)$
		high	$(1, 3)$	$(0, 0)$	$(0, 23)$
	14	low	$(3395, 3530)$	$(231, 245)$	$(11879, 12355)$
		medium	$(1331, 1469)$	$(130, 140)$	$(6650, 7340)$
		high	$(164, 171)$	$(73, 367)$	$(9938, 12020)$

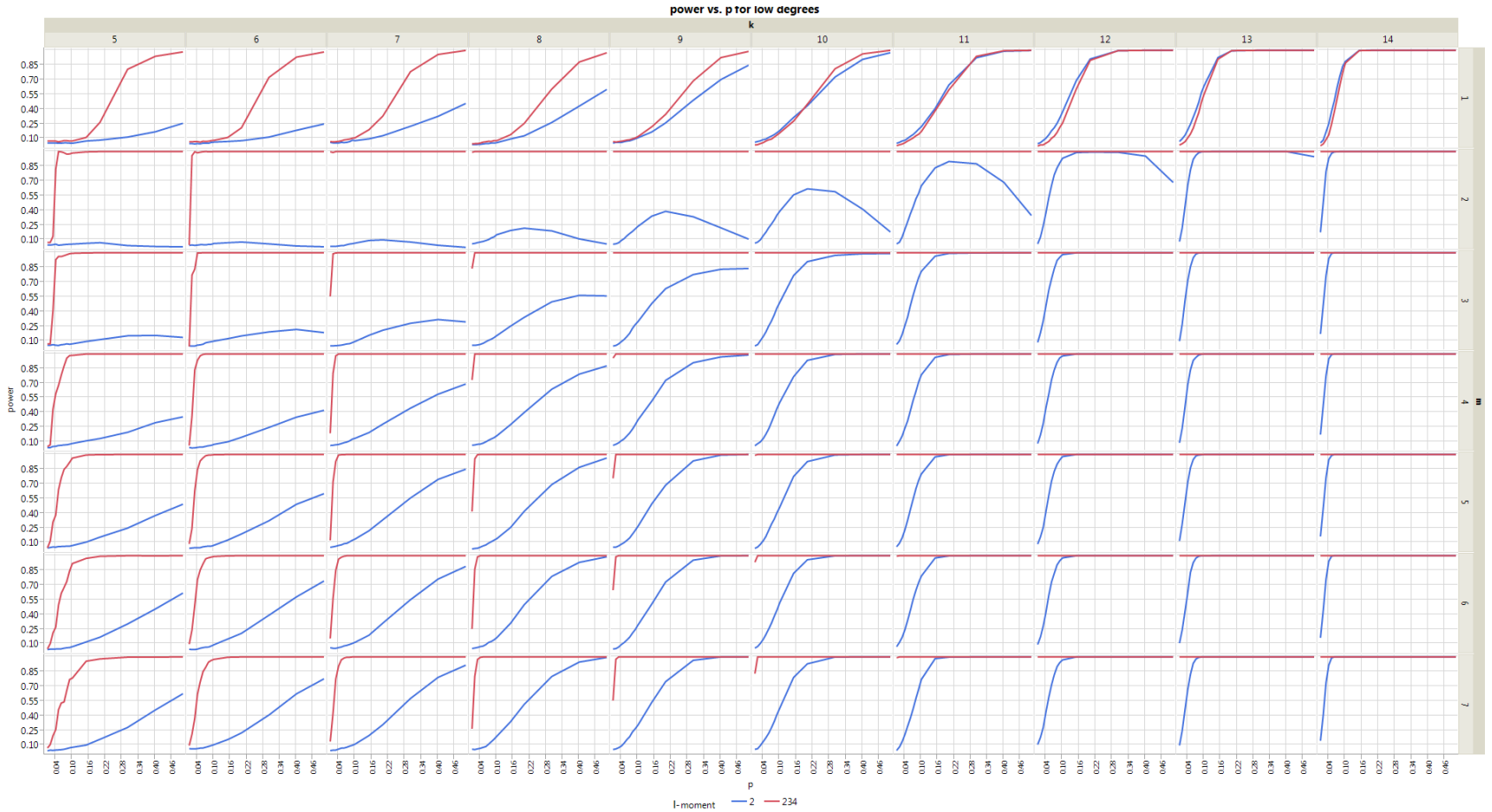


Figure C.7. Power vs. proportion of deletion p for low degree of edge deletion

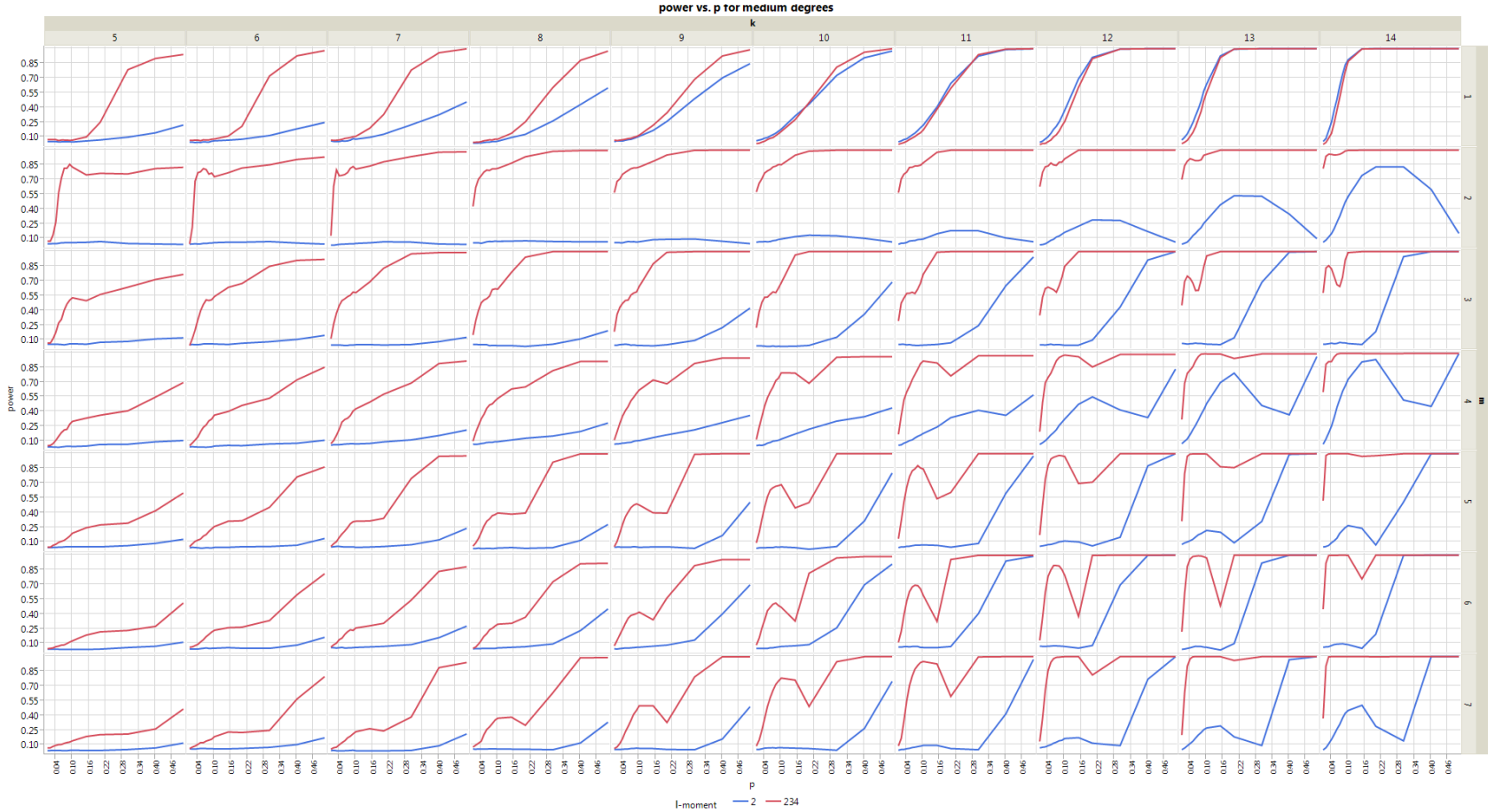


Figure C.8. Power vs. proportion of deletion p for medium degree of edge deletion

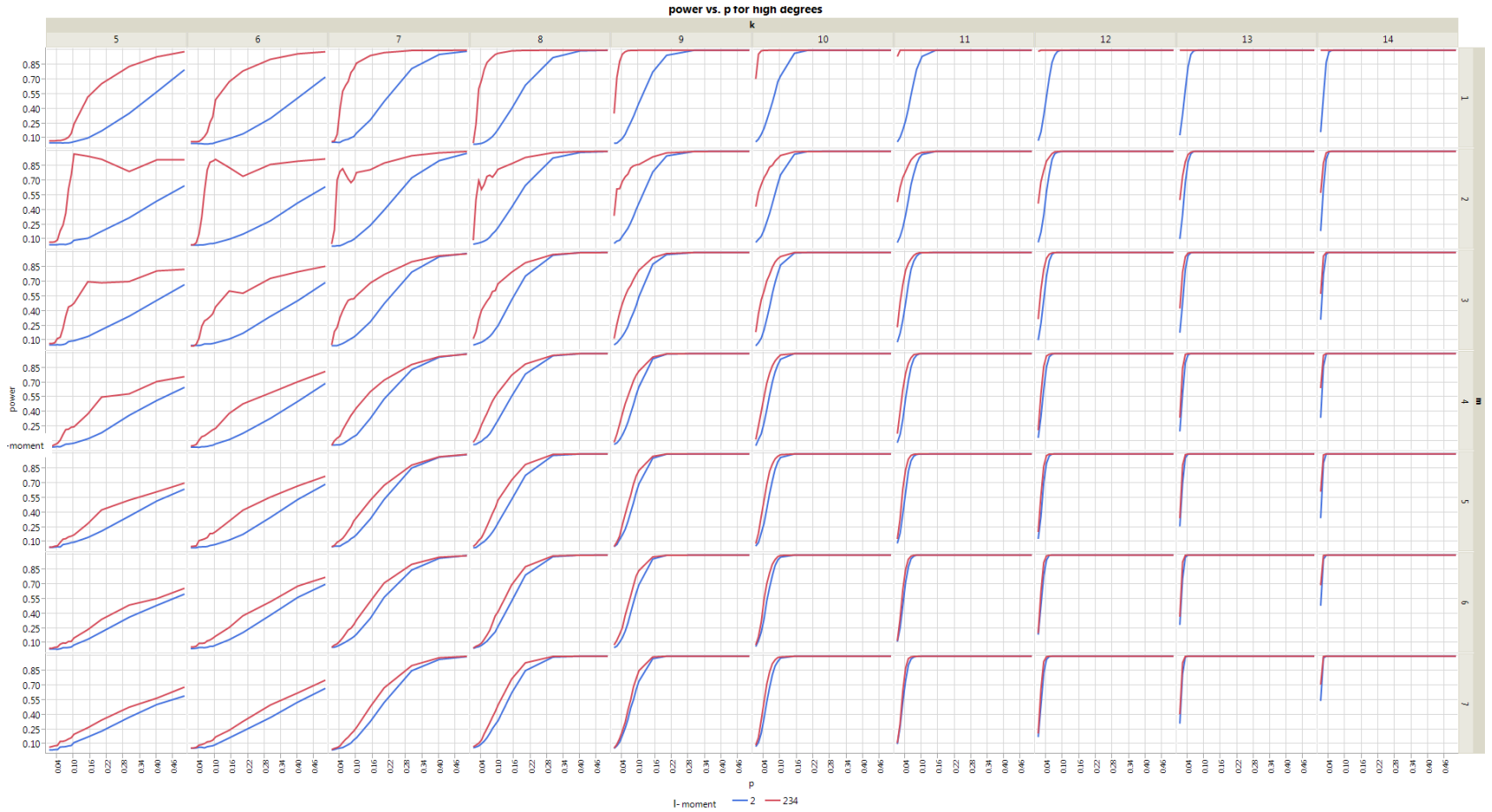


Figure C.9. Power vs. proportion of deletion p for high degree of edge deletion

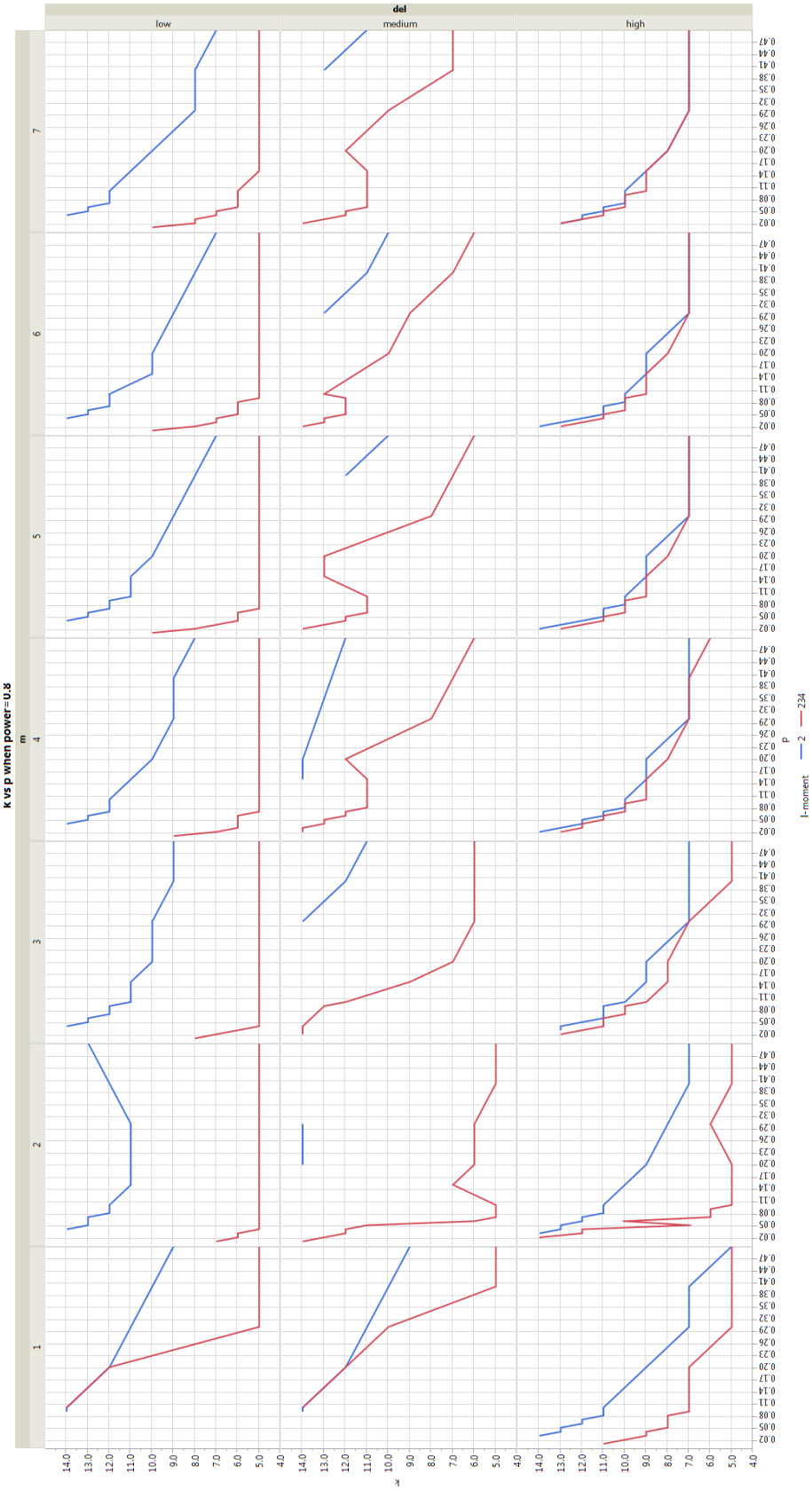


Figure C.10. k **vs proportion of edge deletion** p **to achieve** $power \geq 0.8$

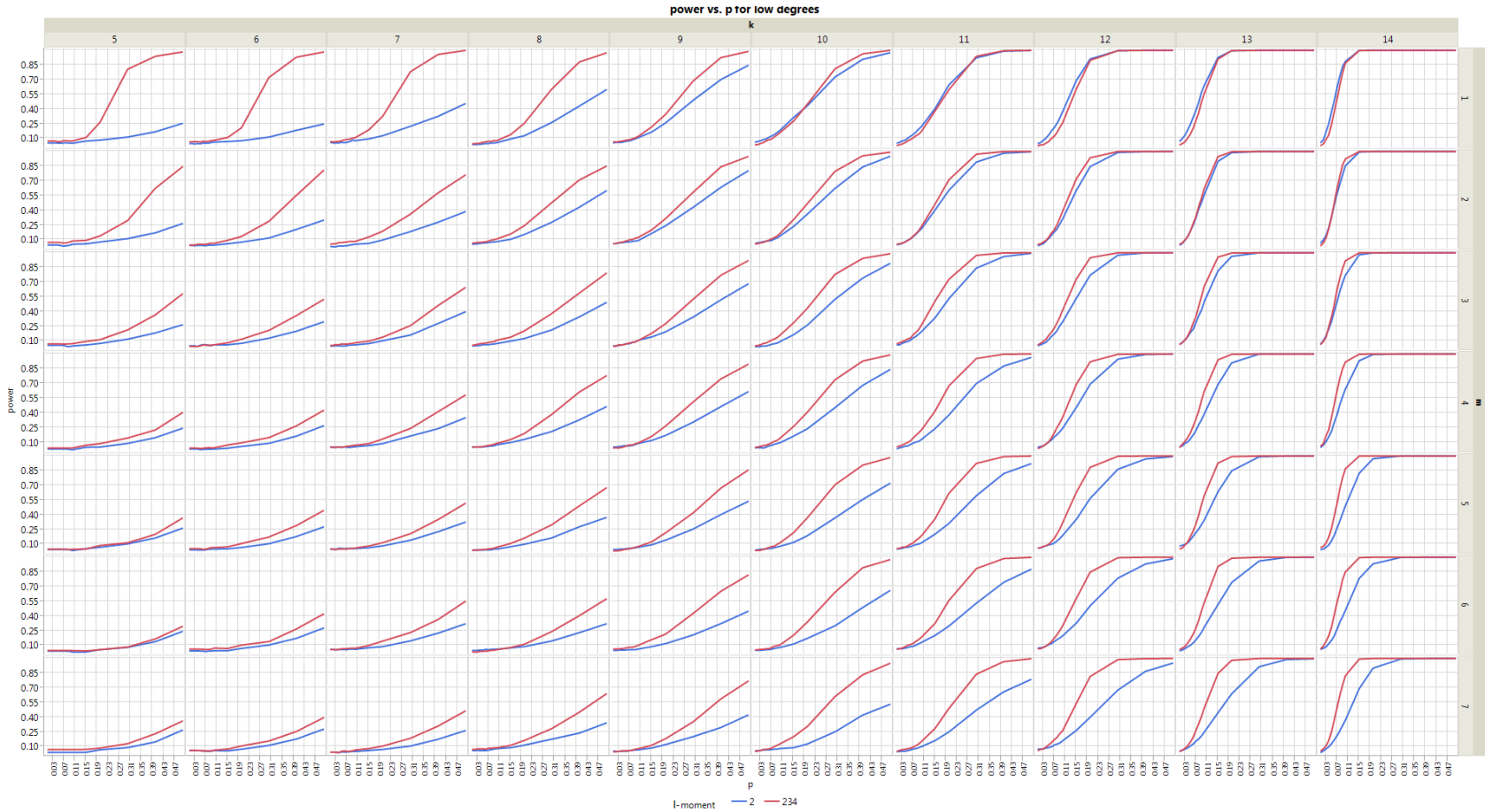


Figure C.11. Power vs. proportion of deletion p for low degree of node deletion

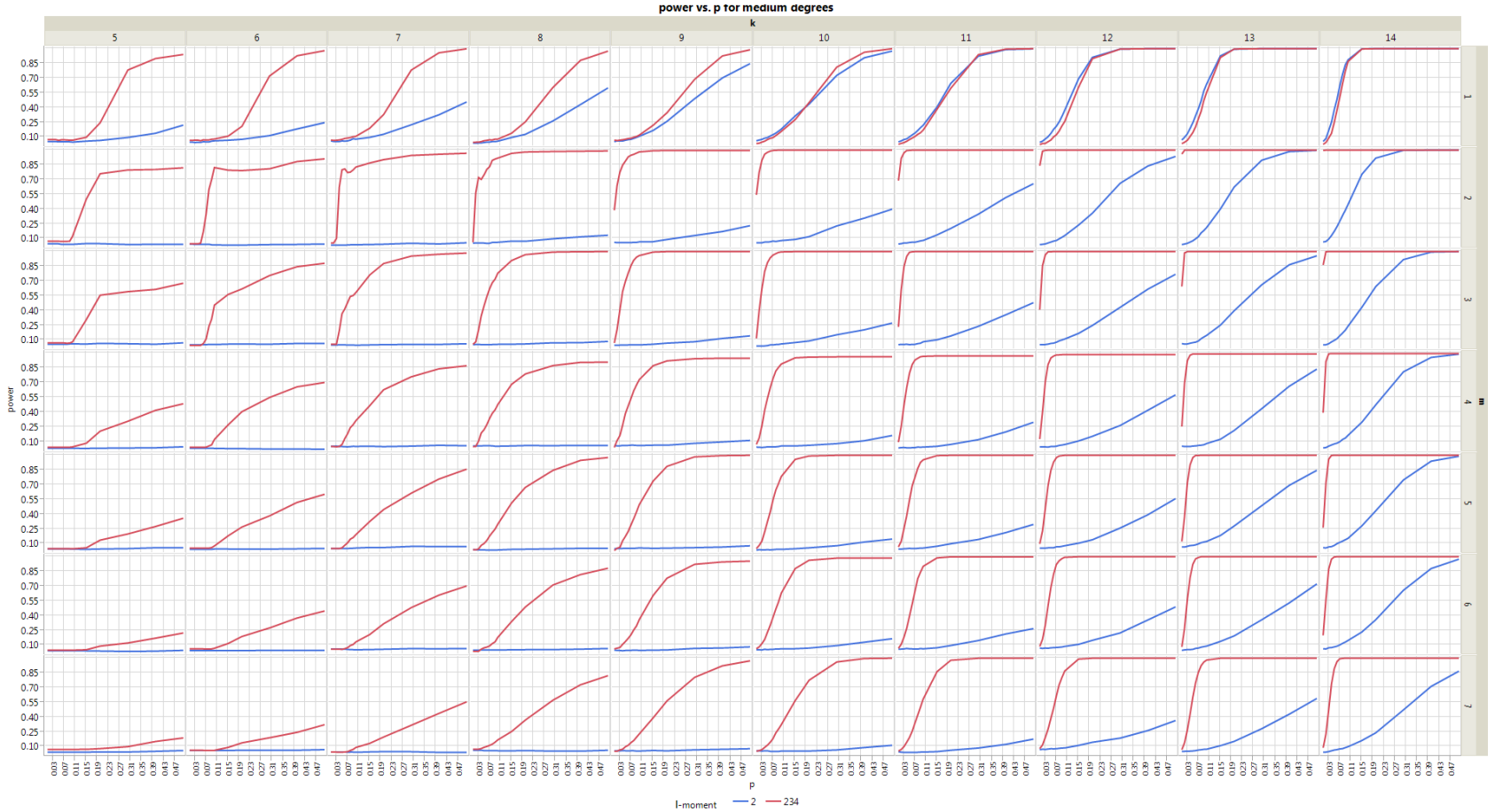


Figure C.12. Power vs. proportion of deletion p for medium degree of node deletion

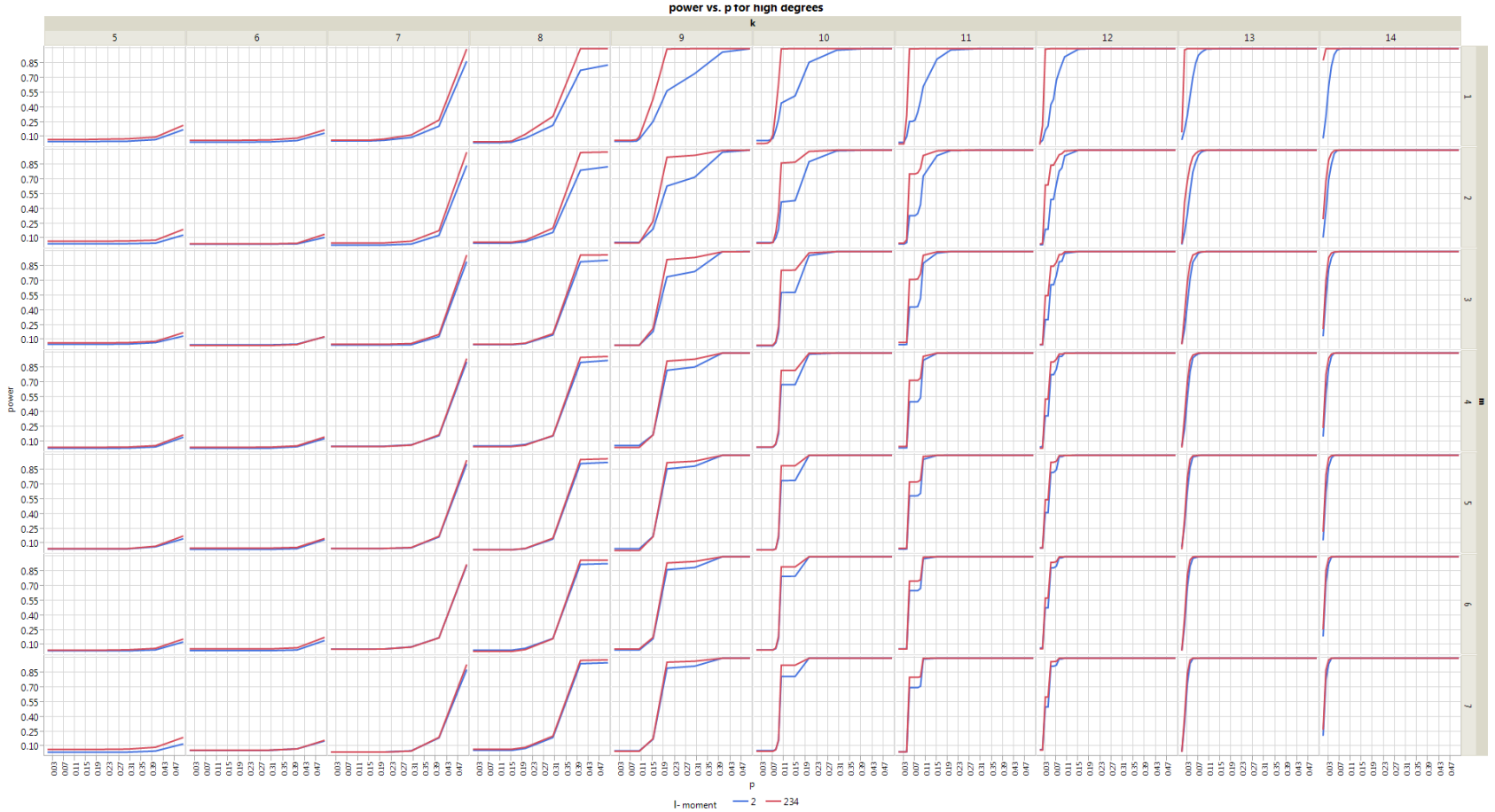


Figure C.13. Power vs. proportion of deletion p for high degree of node deletion

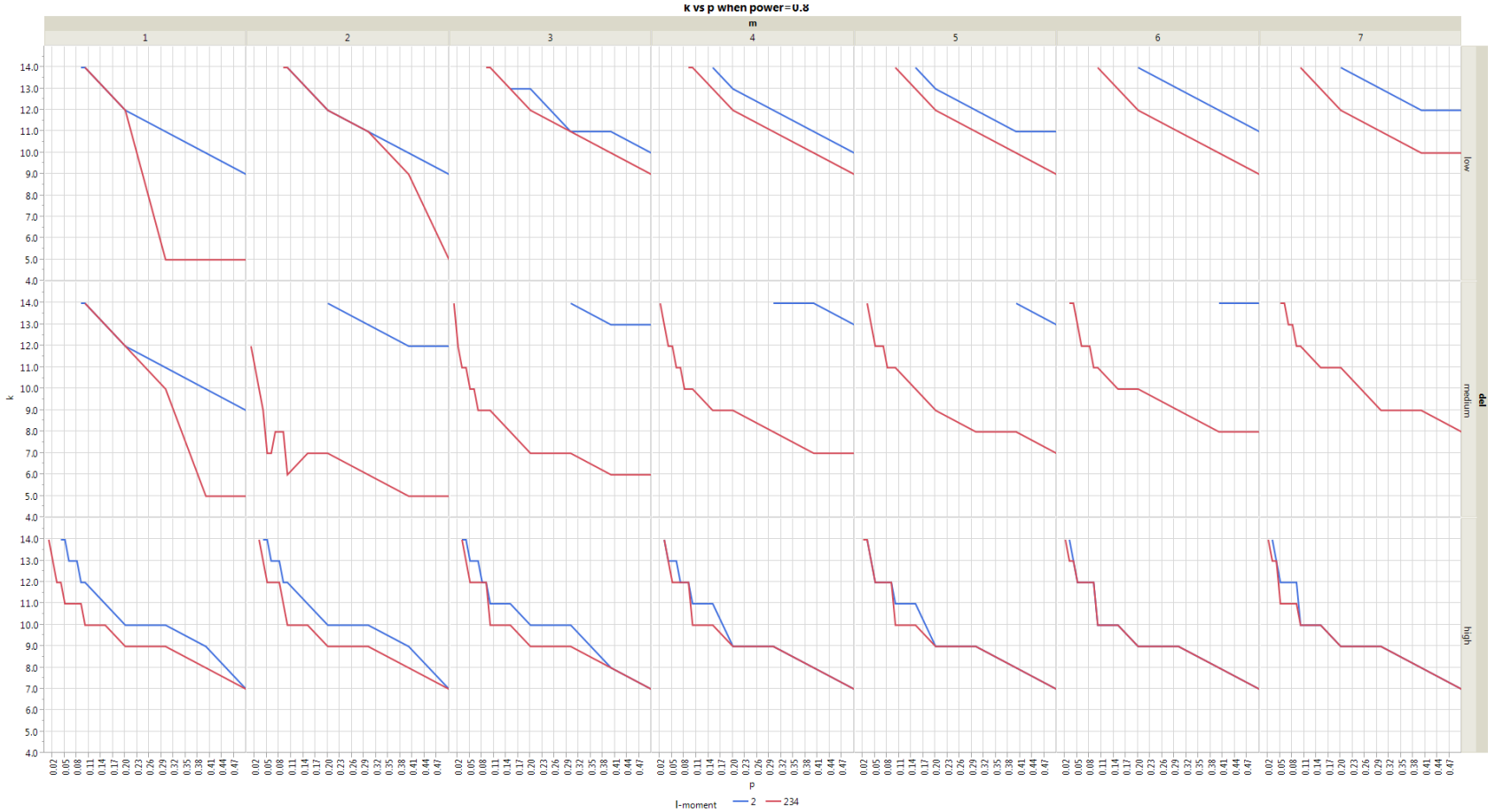


Figure C.14. k **vs proportion of node deletion** p **to achieve** $power \geq 0.8$

Appendix D. Goodness of fit

Table D.8. Goodness of fit via -loglikelihood for all m and n

-loglikelihood																
m=2	K	beta	Lower 95%	Median	Upper 95%	m=4	K	beta	Lower 95%	Median	Upper 95%	m=6	K	beta	Lower 95%	Median	Upper 95%
		1.2	48.12	51.18	54.33			1.2	74.20	76.07	77.85			1.2	88.89	90.20	91.37
		1.5	45.45	48.93	52.51			1.5	72.21	74.34	76.36			1.5	87.30	88.79	90.11
	5	1.8	44.09	47.99	52.00		5	1.8	71.54	73.92	76.18		5	1.8	87.01	88.69	90.17
		2	43.70	47.88	52.17			2	71.60	74.16	76.58			2	87.34	89.13	90.72
		2.1	43.63	47.95	52.39			2.1	71.76	74.40	76.90			2.1	87.63	89.48	91.13
		1.2	96.04	101.32	105.78			1.2	147.31	150.36	153.58			1.2	176.37	178.78	181.13
		1.5	90.53	96.54	101.61			1.5	142.90	146.36	150.02			1.5	172.53	175.27	177.93
	6	1.8	87.64	94.37	100.04		6	1.8	141.09	144.97	149.07		6	1.8	171.29	174.36	177.34
		2	86.75	93.96	100.04			2	140.92	145.08	149.47			2	171.51	174.79	177.99
		2.1	86.55	94.00	100.28			2.1	141.09	145.38	149.92			2.1	171.86	175.26	178.56
		1.2	192.40	200.34	206.86			1.2	291.82	297.26	302.46			1.2	349.32	353.34	357.20
		1.5	181.29	190.32	197.72			1.5	282.32	288.50	294.40			1.5	340.71	345.29	349.67
	7	1.8	175.41	185.52	193.82		7	1.8	278.04	284.96	291.57		7	1.8	337.34	342.46	347.37
		2	173.57	184.39	193.28			2	277.26	284.67	291.75			2	337.16	342.65	347.90
		2.1	173.14	184.33	193.51			2.1	277.36	285.02	292.35			2.1	337.57	343.24	348.67
		1.2	384.97	396.47	407.13			1.2	580.60	588.82	596.19			1.2	693.51	699.41	705.82
		1.5	362.65	375.71	387.83			1.5	560.90	570.24	578.61			1.5	675.19	681.89	689.17
	8	1.8	350.77	365.41	378.97		8	1.8	551.64	562.10	571.48		8	1.8	667.31	674.82	682.97
		2	347.00	362.68	377.22			2	549.61	560.82	570.87			2	666.20	674.25	682.99
		2.1	346.11	362.31	377.33			2.1	549.60	561.18	571.56			2.1	666.65	674.96	683.99
		1.2	770.53	790.14	805.22			1.2	1157.74	1169.97	1182.12			1.2	1379.22	1388.78	1398.20
		1.5	725.83	748.11	765.25			1.5	1117.58	1131.48	1145.28			1.5	1341.09	1351.95	1362.65
	9	1.8	702.03	726.98	746.17		9	1.8	1098.32	1113.89	1129.34		9	1.8	1323.86	1336.02	1348.01
		2	694.44	721.18	741.74			2	1093.77	1110.45	1127.01			2	1320.66	1333.69	1346.54
		2.1	692.65	720.27	741.52			2.1	1093.48	1110.72	1127.83			2.1	1321.05	1334.52	1347.79
		1.2	1547.24	1573.42	1595.59			1.2	2312.14	2329.59	2347.11			1.2	2751.44	2765.31	2780.41
		1.5	1458.53	1488.28	1513.48			1.5	2231.08	2250.92	2270.82			1.5	2673.81	2689.57	2706.73
	10	1.8	1411.63	1444.95	1473.17		10	1.8	2191.83	2214.04	2236.34		10	1.8	2637.98	2655.64	2674.86
		2	1396.94	1432.64	1462.87			2	2182.24	2206.04	2229.93			2	2630.67	2649.59	2670.18
		2.1	1393.57	1430.46	1461.70			2.1	2181.43	2206.02	2230.71			2.1	2631.00	2650.55	2671.83
		1.2	3108.80	3142.78	3172.47			1.2	4624.23	4648.14	4673.84			1.2	5490.17	5513.06	5533.05
		1.5	2933.21	2971.82	3005.56			1.5	4461.84	4489.01	4518.22			1.5	5332.77	5358.78	5381.50
	11	1.8	2841.22	2884.46	2922.25		11	1.8	4383.06	4413.49	4446.20		11	1.8	5258.97	5288.10	5313.55
		2	2813.04	2859.38	2899.86			2	4363.69	4396.30	4431.35			2	5242.92	5274.13	5301.40
		2.1	2806.92	2854.80	2896.63			2.1	4361.97	4395.66	4431.88			2.1	5242.87	5275.12	5303.29
		1.2	6229.36	6275.29	6322.34			1.2	9246.12	9280.90	9318.72			1.2	10977.71	11007.09	11037.12
		1.5	5879.64	5931.84	5985.30			1.5	8920.76	8960.28	9003.25			1.5	10662.14	10695.53	10729.65
	12	1.8	5697.12	5755.59	5815.46		12	1.8	8762.61	8806.87	8855.00		12	1.8	10513.78	10551.17	10589.39
		2	5641.75	5704.39	5768.54			2	8723.47	8770.90	8822.46			2	10481.17	10521.23	10562.18
		2.1	5630.00	5694.72	5761.02			2.1	8719.84	8768.84	8822.13			2.1	10480.80	10522.20	10564.51
		1.2	12480.05	12542.29	12611.74			1.2	18493.75	18546.19	18602.05			1.2	21948.60	21989.81	22033.14
		1.5	11783.38	11854.10	11933.03			1.5	17842.96	17902.55	17966.03			1.5	21316.13	21362.95	21412.18
	13	1.8	11421.13	11500.33	11588.73		13	1.8	17526.58	17593.32	17664.42		13	1.8	21018.06	21070.51	21125.65
		2	11312.23	11397.09	11491.81			2	17448.27	17519.78	17595.96			2	20951.95	21008.15	21067.23
		2.1	11289.65	11377.34	11475.21			2.1	17440.98	17514.87	17593.59			2.1	20950.77	21008.84	21069.88
		1.2	24990.03	25078.96	25176.31			1.2	37000.50	37073.98	37141.52			1.2	43891.00	43950.76	44007.09
		1.5	23600.64	23701.70	23812.32			1.5	35700.42	35783.91	35860.67			1.5	42624.78	42692.69	42756.71
	14	1.8	22880.07	22993.26	23117.15		14	1.8	35069.16	35162.68	35248.64		14	1.8	42027.39	42103.45	42175.15
		2	22664.91	22786.18	22918.92			2	34913.54	35013.73	35105.84			2	41894.35	41975.84	42052.66
		2.1	22621.06	22746.37	22883.54			2.1	34899.46	35002.99	35098.17			2.1	41891.56	41975.76	42055.15
		1.2	50018.59	50146.06	50285.65			1.2	74024.62	74128.23	74228.31			1.2	87779.84	87874.56	87963.13
		1.5	47244.92	47389.77	47548.39			1.5	71427.40	71545.14	71658.87			1.5	85246.71	85354.34	85454.99
	15	1.8	45808.90	45971.14	46148.80		15	1.8	70167.84	70299.71	70427.08		15	1.8	84051.23	84171.78	84284.50
		2	45381.98	45555.81	45746.15			2	69858.55	69999.84	70136.31			2	83784.66	83913.83	84034.60
		2.1	45295.99	45475.61	45672.30			2.1	69831.38	69977.38	70118.40			2.1	83778.85	83912.32	84037.12

Shadings indicate groupings of overlapping CIs.

Table D.9. Goodness of fit via MSE for all m and n

MSE																
m=2	K	beta	Lower 95%	Median	Upper 95%	m=4	K	beta	Lower 95%	Median	Upper 95%	m=7	K	beta	Lower 95%	Median	Upper 95%
		1.2	0.00227	0.00944	0.02551			1.2	0.00051	0.00228	0.00830			1.2	0.00101	0.00277	0.00806
		1.5	0.01592	0.03486	0.05255			1.5	0.00074	0.00280	0.00855			1.5	0.00076	0.00232	0.00699
	5	1.8	0.04968	0.07589	0.09198		5	1.8	0.00184	0.00575	0.01197		5	1.8	0.00089	0.00257	0.00695
		2	0.08276	0.11129	0.12460			2	0.00314	0.00880	0.01505			2	0.00115	0.00304	0.00742
		2.1	0.10234	0.13134	0.14344			2.1	0.00423	0.01061	0.01679			2.1	0.00129	0.00336	0.00763
		1.2	0.00367	0.01019	0.01995			1.2	0.00019	0.00096	0.00422			1.2	0.00039	0.00120	0.00392
		1.5	0.02523	0.03838	0.05072			1.5	0.00057	0.00191	0.00588			1.5	0.00025	0.00083	0.00292
	6	1.8	0.06787	0.08416	0.09598		6	1.8	0.00253	0.00603	0.01048		6	1.8	0.00054	0.00146	0.00356
		2	0.10743	0.12418	0.13442			2	0.00537	0.01020	0.01482			2	0.00104	0.00238	0.00449
		2.1	0.13049	0.14700	0.15553			2.1	0.00737	0.01272	0.01742			2.1	0.00139	0.00298	0.00512
		1.2	0.00500	0.01020	0.01652			1.2	0.00010	0.00049	0.00225			1.2	0.00025	0.00078	0.00234
		1.5	0.02960	0.03974	0.04838			1.5	0.00064	0.00165	0.00392			1.5	0.00012	0.00044	0.00166
	7	1.8	0.07564	0.08819	0.09682		7	1.8	0.00364	0.00627	0.00927		7	1.8	0.00057	0.00131	0.00266
		2	0.11806	0.13081	0.13848			2	0.00741	0.01105	0.01429			2	0.00133	0.00252	0.00400
		2.1	0.14275	0.15518	0.16189			2.1	0.00991	0.01396	0.01729			2.1	0.00190	0.00329	0.00480
		1.2	0.00621	0.01004	0.01447			1.2	0.00006	0.00028	0.00120			1.2	0.00023	0.00058	0.00154
		1.5	0.03279	0.03999	0.04646			1.5	0.00072	0.00158	0.00304			1.5	0.00007	0.00024	0.00081
	8	1.8	0.08097	0.08970	0.09652		8	1.8	0.00436	0.00646	0.00862		8	1.8	0.00069	0.00122	0.00209
		2	0.12491	0.13364	0.13974			2	0.00883	0.01158	0.01405			2	0.00173	0.00257	0.00362
		2.1	0.15041	0.15881	0.16426			2.1	0.01168	0.01470	0.01726			2.1	0.00245	0.00343	0.00456
		1.2	0.00700	0.00998	0.01308			1.2	0.00003	0.00016	0.00061			1.2	0.00022	0.00048	0.00103
		1.5	0.03492	0.04028	0.04496			1.5	0.00088	0.00153	0.00252			1.5	0.00005	0.00015	0.00046
	9	1.8	0.08424	0.09063	0.09569		9	1.8	0.00510	0.00658	0.00815		9	1.8	0.00079	0.00120	0.00182
		2	0.12890	0.13523	0.13979			2	0.01001	0.01189	0.01371			2	0.00199	0.00262	0.00339
		2.1	0.15476	0.16083	0.16493			2.1	0.01307	0.01510	0.01701			2.1	0.00280	0.00353	0.00438
		1.2	0.00795	0.00999	0.01215			1.2	0.00002	0.00011	0.00038			1.2	0.00025	0.00044	0.00083
		1.5	0.03692	0.04046	0.04384			1.5	0.00099	0.00149	0.00217			1.5	0.00004	0.00011	0.00029
	10	1.8	0.08699	0.09114	0.09475		10	1.8	0.00554	0.00659	0.00771		10	1.8	0.00086	0.00119	0.00160
		2	0.13202	0.13607	0.13942			2	0.01066	0.01197	0.01326			2	0.00214	0.00264	0.00317
		2.1	0.15802	0.16187	0.16491			2.1	0.01383	0.01524	0.01661			2.1	0.00301	0.00357	0.00416
		1.2	0.00851	0.00999	0.01137			1.2	0.00002	0.00007	0.00024			1.2	0.00028	0.00042	0.00069
		1.5	0.03800	0.04053	0.04271			1.5	0.00112	0.00149	0.00192			1.5	0.00004	0.00008	0.00018
	11	1.8	0.08843	0.09138	0.09380		11	1.8	0.00586	0.00664	0.00740		11	1.8	0.00093	0.00118	0.00144
		2	0.13363	0.13649	0.13873			2	0.01110	0.01208	0.01298			2	0.00228	0.00266	0.00301
		2.1	0.15970	0.16240	0.16448			2.1	0.01433	0.01538	0.01634			2.1	0.00318	0.00361	0.00400
		1.2	0.00898	0.00996	0.01097			1.2	0.00002	0.00006	0.00016			1.2	0.00029	0.00041	0.00054
		1.5	0.03883	0.04052	0.04214			1.5	0.00120	0.00147	0.00178			1.5	0.00004	0.00007	0.00013
	12	1.8	0.08949	0.09144	0.09326		12	1.8	0.00609	0.00663	0.00719		12	1.8	0.00102	0.00118	0.00137
		2	0.13479	0.13662	0.13835			2	0.01141	0.01207	0.01274			2	0.00243	0.00267	0.00293
		2.1	0.16087	0.16258	0.16419			2.1	0.01468	0.01539	0.01609			2.1	0.00336	0.00363	0.00392
		1.2	0.00923	0.00995	0.01075			1.2	0.00002	0.00005	0.00012			1.2	0.00032	0.00040	0.00051
		1.5	0.03931	0.04053	0.04180			1.5	0.00127	0.00148	0.00170			1.5	0.00005	0.00007	0.00011
	13	1.8	0.09009	0.09148	0.09290		13	1.8	0.00624	0.00666	0.00707		13	1.8	0.00105	0.00118	0.00131
		2	0.13538	0.13672	0.13805			2	0.01161	0.01212	0.01262			2	0.00249	0.00267	0.00285
		2.1	0.16146	0.16271	0.16394			2.1	0.01489	0.01544	0.01597			2.1	0.00343	0.00364	0.00384
		1.2	0.00947	0.00996	0.01046			1.2	0.00003	0.00005	0.00009			1.2	0.00034	0.00040	0.00046
		1.5	0.03974	0.04055	0.04136			1.5	0.00134	0.00147	0.00162			1.5	0.00005	0.00007	0.00009
	14	1.8	0.09060	0.09153	0.09244		14	1.8	0.00639	0.00666	0.00692		14	1.8	0.00109	0.00118	0.00127
		2	0.13590	0.13678	0.13763			2	0.01180	0.01212	0.01244			2	0.00255	0.00267	0.00280
		2.1	0.16195	0.16278	0.16358			2.1	0.01511	0.01545	0.01578			2.1	0.00350	0.00364	0.00378
		1.2	0.00960	0.00995	0.01036			1.2	0.00003	0.00005	0.00008			1.2	0.00036	0.00040	0.00046
		1.5	0.03997	0.04055	0.04120			1.5	0.00137	0.00147	0.00157			1.5	0.00005	0.00006	0.00008
	15	1.8	0.09087	0.09153	0.09228		15	1.8	0.00645	0.00665	0.00684		15	1.8	0.00110	0.00118	0.00124
		2	0.13617	0.13680	0.13750			2	0.01188	0.01212	0.01235			2	0.00257	0.00267	0.00276
		2.1	0.16222	0.16281	0.16345			2.1	0.01519	0.01544	0.01569			2.1	0.00352	0.00364	0.00373

Shadings indicate groupings of overlapping CIs.

Bibliography

1. "The PubChem Project", 2016. URL `http://pubchem.ncbi.nlm.nih.gov/`. Retrieved on 7 May 2016.

2. "RCSB Protein Data Bank", 2016. URL `http://www.rcsb.org/pdb/home/home.do`. Retrieved on 7 May 2016.

3. Adamic, Lada A. and Natalie Glance. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog". Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05, 36–43. ACM, New York, NY, USA, 2005.

4. Albert, Réka and Albert-László Barabási. "Statistical mechanics of complex networks". Reviews of Modern Physics, 74:47–97, Jan 2002.

5. Albert, Réka, Hawoong Jeong, and Albert-László Barabási. "Internet: Diameter of the world-wide web". Nature, 401(6749):130–131, 1999.

6. Barabási, Albert-László and Réka Albert. "Emergence of Scaling in Random Networks". Science, 286(5439):509–512, 1999.

7. Barabási, Albert-László and Réka Albert. "Mean-field theory for scale-free random networks". Physica A: Statistical Mechanics and its Applications, 272(12):173–187, 1999.

8. Bilgin, Cemal Cagatay, Peter Bullough, George E. Plopper, and Bülent Yener. "ECM-aware cell-graph mining for bone tissue modeling and classification". Data Mining and Knowledge Discovery, 20(3):416–438, 2010.

9. Bilgin, Cemal Cagatay, Cigdem Demir, Chandandeep Nagi, and Bulent Yener. "Cell-Graph Mining for Breast Tissue Modeling and Classification". Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 5311–5314. Aug 2007.

10. Blaha, Leslie M., Dustin L. Arendt, and Fairul Mohd-Zaid. "More Bang for Your Research Buck: Toward Recommender Systems for Visual Analytics". Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, BELIV '14, 126–133. ACM, New York, NY, USA, 2014.

11. Blasio, Birgitte Freiesleben de, Taral Guldahl Seierstad, and Odd O. Aalen. "Frailty effects in networks: comparison and identification of individual heterogeneity versus preferential attachment in evolving networks". Journal of the Royal Statistical Society: Series C (Applied Statistics), 60(2):239–259, 2011.

12. Bolland, John M. "Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks". Social Networks, 10(3):233 – 253, 1988.

13. Bonnici, Vincenzo, Rosalba Giugno, Alfredo Pulvirenti, Dennis Shasha, and Alfredo Ferro. "RI A subgraph isomorphism algorithm.", 2014. URL `http://ferrolab.dmi.unict.it/ri/datasets.html`. Retrieved on 7 May 2016.

14. Casella, George and Roger L. Berger. Statistical Inference. Thomson Learning, 2002.

15. Chan, Lionel K. "On a characterization of distributions by expected values of extreme order statistics". American Mathematical Monthly, 74:950–951, 1967.

16. Clauset, Aaron, Cosma Rohilla Shalizi, and Mark E.J. Newman. "Power-Law Distributions in Empirical Data". SIAM Review, 51(4):661–703, 2009.

17. Csardi, Gabor and Tamas Nepusz. "The igraph software package for complex network research". InterJournal, Complex Systems:1695, 2006.

18. D'Agostino, Ralph B. and Albert Belanger. "A Suggestion for Using Powerful and Informative Tests of Normality". The American Statistician, 44(4):316–321, 1990.

19. Demir, Cigdem, S. Humayun Gultekin, and Bulent Yener. "Learning the topological properties of brain tumors". Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2(3):262–270, July 2005.

20. Downton, Frank. "Linear estimates with polynomial coefficients". Biometrika, 53(1-2):129–141, 1966.

21. Elamir, Elsayed A. H. and Allan H. Seheult. "Trimmed L-moments". Computational Statistics & Data Analysis, 43(3):299–314, July 2003.

22. Elamir, Elsayed A.H. and Allan H. Seheult. "Exact variance structure of sample L-moments". Journal of Statistical Planning and Inference, 124(2):337 – 359, 2004.

23. Erdös, Paul and Alfréd Rényi. "On Random Graphs I". Publicationes Mathematicae Debrecen, 6:290–297, 1959.

24. Fienberg, Stephen E. "A Brief History of Statistical Models for Network Analysis and Open Challenges". Journal of Computational and Graphical Statistics, 21(4):825–839, 2012.

25. Freeman, Linton C. "Centrality in social networks conceptual clarification". Social Networks, 1(3):215 – 239, 1978.

26. Galvao, Antonio F., Gabriel Montes-Rojas, Walter Sosa-Escudero, and Liang Wang. "Tests for skewness and kurtosis in the one-way error component model". Journal of Multivariate Analysis, 122(0):35 – 52, 2013.

27. Gehrke, Johannes, Paul Ginsparg, and Jon M. Kleinberg. "Overview of the 2003 KDD Cup". SIGKDD Explorations, 5(2):149–151, 2003.

28. Gibert, Jaume, Ernest Valveny, and Horst Bunke. "Graph embedding in vector spaces by node attribute statistics". Pattern Recognition, 45(9):3072 – 3083, 2012.

29. Gillespie, Colin S. Fitting heavy tailed distributions: the poweRlaw package, 2013. R package version 0.20.1.

30. Gini, Corrado. "Variability and mutability, contribution to the study of statistical distributions and relations". Studi Economico-Giuridici della R. Università de Cagliari, 3:3–159, 1912.

31. Goel, N.K., Donald H. Burn, Mahesh D. Pandey, and Ying An. "Wind quantile estimation using a pooled frequency analysis approach". Journal of Wind Engineering and Industrial Aerodynamics, 92(6):509 – 528, 2004.

32. Gonzalez, Jesus A., Lawrence B. Holder, and Diane J. Cook. "Graph-Based Relational Concept Learning". Proceedings of the Nineteenth International Conference on Machine Learning, ICML '02, 219–226. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

33. Greenwood, J. Arthur, J. Maciunas Landwehr, Nicolas C. Matalas, and James R. Wallis. "Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form". Water Resources Research, 15(5):1049–1054, 1979.

34. Guzman, Joshua D., Richard F. Deckro, Matthew J. Robbins, James F. Morris, and Nicholas A. Ballester. "An Analytical Comparison of Social Network Measures". IEEE Transactions on Computational Social Systems, 1(1):35–45, March 2014.

35. Harri, Ardian and Keith H. Coble. "Normality testing: two new tests using L-moments". Journal of Applied Statistics, 38(7):1369–1379, 2011.

36. Havig, Paul R., John P. McIntire, Eric Geiselman, and Fairul Mohd-Zaid. "Why social network analysis is important to Air Force applications". Proceedings of SPIE, volume 8389, 83891E–83891E–9. 2012.

37. Helma, C., R.D. King, S. Kramer, and A. Srinivasan. "The Predictive Toxicology Challenge (PTC) for 2000-2001", 2002. URL `http://www.predictive-toxicology.org/ptc/`. Retrieved on 7 May 2016.

38. Holme, Petter and Beom Jun Kim. "Growing scale-free networks with tunable clustering". Physical Review E, 65:026107, Jan 2002.

39. Hosking, Jonathan R.M. The theory of probability weighted moments. Technical Report Research Report RC12210, IBM Research, 1986.

40. Hosking, Jonathan R.M. "L-moments: analysis and estimation of distributions using linear combinations of order statistics". Journal of the Royal Statistical Society. Series B (Methodological), 105–124, 1990.

41. Hosking, Jonathan R.M. "Distributions with maximum entropy subject to constraints on their L-moments or expected order statistics". Journal of Statistical Planning and Inference, 137(9):2870 – 2891, 2007.

42. Hosking, Jonathan R.M. and James R. Wallis. Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, 2005.

43. Inokuchi, Akihiro, Takashi Washio, and Hiroshi Motoda. "An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data". Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD '00, 13–23. Springer-Verlag, London, UK, UK, 2000.

44. Jarque, Carlos M. and Anil K. Bera. "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". Economics Letters, 6(3):255 – 259, 1980.

45. Jin, Ning, Calvin Young, and Wei Wang. "Graph Classification Based on Pattern Co-occurrence". Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, 573–582. ACM, New York, NY, USA, 2009.

46. Ketkar, Nikhil S., Lawrence B. Holder, and Diane J. Cook. "Empirical comparison of graph classification algorithms." CIDM, 259–266. IEEE, 2009.

47. Kjeldsen, Thomas Rødding, J.C. Smithers, and R.E. Schulze. "Regional flood frequency analysis in the KwaZulu-Natal province, South Africa, using the index-flood method". Journal of Hydrology, 255(1-4):194–211, 2002.

48. Knuth, Donald E. The Stanford GraphBase: A Platform for Combinatorial Computing. pub-ACM, 1993.

49. Konheim, A.G. "A note on order statistics". American Mathematical Monthly, 78:524–524, 1971.

50. Körner, János. "Coding of an information source having ambiguous alphabet and the entropy of graphs". Proceedings of the 6th Prague Conference on Information Theory, 411–425. Prague, Czech Republic, 1973.

51. Kroll, Charles N. and Richard M. Vogel. "Probability Distribution of Low Streamflow Series in the United States". Journal of Hydrologic Engineering, 7(2):137–146, 2002.

52. Kvam, Paul H. and Brani Vidakovic. Nonparametric Statistics with Applications to Science and Engineering (Wiley Series in Probability and Statistics). Wiley-Interscience, 2007.

53. Lagraa, Sofiane, Hamida Seba, Riadh Khennoufa, Abir MBaya, and Hamamache Kheddouci. "A distance measure for large graphs based on prime graphs". Pattern Recognition, 47(9):2993 – 3005, 2014.

54. Landwehr, J. Maciunas, N. C. Matalas, and J. R. Wallis. "Probability weighted moments compared with some traditional techniques in estimating Gumbel Parameters and quantiles". Water Resources Research, 15(5):1055–1064, 1979.

55. Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations". Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, 177–187. ACM, New York, NY, USA, 2005.

56. Leskovec, Jure and Andrej Krevl. "SNAP Datasets: Stanford Large Network Dataset Collection". `http://snap.stanford.edu/data`, June 2014. Retrieved 9 December 2015.

57. Leskovec, Jure, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters". Internet Mathematics, 6(1):29–123, 2009.

58. Li, Geng, Murat Semerci, Bülent Yener, and Mohammed J. Zaki. "Effective Graph Classification Based on Topological and Label Attributes". Statistical Analysis and Data Mining, 5(4):265–283, August 2012.

59. Li, Ping, Jie Zhang, and Michael Small. "Emergence of scaling and assortative mixing through altruism". Physica A: Statistical Mechanics and its Application, 390(11):2192–2197, 2011.

60. Lusseau, David, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson. "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations". Behavioral Ecology and Sociobiology, 54(4):396–405, 2003.

61. Macindoe, Owen and Whitman Richards. "Graph Comparison Using Fine Structure Analysis". Social Computing (SocialCom), 2010 IEEE Second International Conference on, 193–200. Aug 2010.

62. Mallows, C.L. "Bounds on Distribution Functions in Terms of Expectations of Order- Statistics". The Annals of Probability, 1(2):297–303, 04 1973.

63. McAuley, Julian J. and Jure Leskovec. "Discovering Social Circles in Ego Networks". CoRR, abs/1210.8182, 2012.

64. Mohd-Zaid, Fairul and Christine M. Schubert Kabban. "Using Moments and L-Moments to Characterize Graphical Networks". JSM Proceedings, Section on Statistical Learning and Data Mining, ENAR, 2387–2399. American Statistical Association, Alexandria, VA, 2015.

65. Moonesinghe, H.D.K., Hamed Valizadegan, Samah Fodeh, and Pang-Ning Tan. "A Probabilistic Substructure-Based Approach for Graph Classification". Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on, volume 1, 346–349. Oct 2007.

66. Morris, James F., Jerome W. ONeal, and Richard F. Deckro. "A random graph generation algorithm for the analysis of social networks". The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 11(3):265–276, 2014.

67. Mowshowitz, Abbe and Matthias Dehmer. "Entropy and the Complexity of Graphs Revisited". Entropy, 14(3):559–570, 2012.

68. Newman, Mark E.J. "Clustering and preferential attachment in growing networks". Physical Review E, 64:025102, Jul 2001.

69. Newman, Mark E.J. "The structure of scientific collaboration networks". Proceedings of the National Academy of Sciences, 98(2):404–409, 2001.

70. Newman, Mark EJ. "Power laws, Pareto distributions and Zipf's law". Contemporary Physics, 46(5):323–351, 2005.

71. Newman, Mark E.J. "Network data", 2013. URL `http://www-personal.umich.edu/~mejn/netdata/`. Retrieved on 7 May 2016.

72. Nguyen, Phu Chien, Kouzou Ohara, Akira Mogi, Hiroshi Motoda, and Takashi Washio. "Constructing Decision Trees for Graph-structured Data by Chunkingless Graph-based Induction". Proceedings of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD'06, 390–399. Springer-Verlag, Berlin, Heidelberg, 2006.

73. Pareto, Vilfredo and Giovanni Busino. Ecrits sur la courbe de la repartition de la richesse: reunis et presentes par G.Busino (Originally published in 1896). Travaux de driot, d'économie, de sociologie et de sciences politiques. Droz, 1965.

74. Pearson, Egon S. "Some problems arising in approximating to probability distributions, using moments". Biometrika, 50(1-2):95–112, 1963.

75. Pearson, Egon S. "Tables of Percentage Points of $\sqrt{b_1}$ and b_2 in Normal Samples; a Rounding Off". Biometrika, 52(1/2):pp. 282–285, 1965.

76. Pearson, Egon S., Ralph B. D'Agostino, and K. O. Bowman. "Tests for departure from normality: Comparison of powers". Biometrika, 64(2):231–246, 1977.

77. Richards, Whitman and Nicholas Wormald. "Representing Small Group Evolution". Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04, CSE '09, 159–165. IEEE Computer Society, Washington, DC, USA, 2009.

78. Riesen, Kaspar. "IAM Graph Database Repository", 2016. URL `http://www.fki.inf.unibe.ch/databases/iam-graph-database/`. Retrieved on 7 May 2016.

79. Royen, Thomas. "Exact distribution of the sample variance from a gamma parent distribution". arXiv preprint arXiv:0704.1415, 2007.

80. Royston, J. P. "Some Techniques for Assessing Multivarate Normality Based on the Shapiro-Wilk W". Journal of the Royal Statistical Society. Series C (Applied Statistics), 32(2):121–133, 1983.

81. Seier, Edith. "Comparison of tests for univariate normality". InterStat, 1:1–17, 2002.

82. Serfling, Robert and Peng Xiao. "A contribution to multivariate L-moments: L-comoment matrices". Journal of Multivariate Analysis, 98(9):1765 – 1781, 2007.

83. Sillitoe, I., T. Lewis, D. Lee, J. Lees, and C. Orengo. "CATH: Protein Structure Classification Database at UCL", 2016. URL `http://www.cathdb.info/`. Retrieved on 7 May 2016.

84. Sillitto, George P. "Interrelations between certain linear systematic statistics of samples from any continuous population". Biometrika, 56:377–382, 1951.

85. Sillitto, George P. "Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample". Biometrika, 56(3):641–650, 1969.

86. Silverman, Bernard W. Density Estimation for Statistics and Data Analysis. Chapman and Hall, 1986.

87. Simon, Herbert A. "On a class of skew distribution functions". Biometrika, 42(3-4):425–440, 1955.

88. Small, Michael, Kevin Judd, and Linjun Zhang. "How is that complex network complex?" Circuits and Systems (ISCAS), 2014 IEEE International Symposium on, 1263–1266. June 2014.

89. Snijders, Tom A.B. "The degree variance: An index of graph heterogeneity". Social Networks, 3(3):163 – 174, 1981.

90. de Solla Price, Derek J. "Networks of Scientific Papers". Science, 149(3683):510–515, 1965.

91. Ugander, Johan, Lars Backstrom, and Jon Kleinberg. "Subgraph Frequencies: Mapping the Empirical and Extremal Geography of Large Graph Collections". Proceedings of the 22nd International Conference on World Wide Web, WWW '13, 1307–1318. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2013.

92. Vaserstein, Leonid Nisonovich. "Markov processes over denumerable products of spaces, describing large systems of automata". Problems of Information Transmission, 5(3):64–72, 1969.

93. Wand, Matt P. and M. Chris Jones. Density Estimation for Statistics and Data Analysis. Chapman and Hall, 1986.

94. Wang, Q.J. "Direct Sample Estimators of L-Moments". Water Resources Research, 32(12):3617–3619, 1996.

95. Wasserman, Stanley and Katherine Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

96. Watts, Duncan J. and Steven H. Strogatz. "Collective dynamics of 'small-world' networks". Nature, 393(6684):440–442, 1998.

97. West, Douglas B. Introduction to Graph Theory. Prentice Hall, 2001.

98. Yan, Xifeng, Hong Cheng, Jiawei Han, and Philip S. Yu. "Mining Significant Graph Patterns by Leap Search". Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, 433–444. ACM, New York, NY, USA, 2008.

99. Yanou, Ghislain. Extension of random matrix theory to the L-moments for robust portfolio allocation. Documents de travail du centre d'economie de la sorbonne, Universit Panthon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne, 2008.

100. Yue, Sheng and Paul Pilon. "Probability distribution type of Canadian annual minimum streamflow". Hydrological Sciences Journal, 50(3):427–438, 2005.

101. Yule, G. Udny. "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S." Philosophical Transactions of the Royal Society of London B: Biological Sciences, 213(402-410):21–87, 1925.

102. Zachary, Wayne W. "An information flow model for conflict and fission in small groups". Journal of Anthropological Research, 33:452–473, 1977.

103. Zaharevitz, Daniel. "AIDS Antiviral Screen Data", 2015. URL `https://wiki.nci.nih.gov/display/NCIDTPdata/`. Retrieved on 7 May 2016.

104. Zaidman, Maxine D, Virginie Keller, Andrew R Young, and Daniel Cadman. "Flow-duration-frequency behaviour of British rivers based on annual minima data". Journal of Hydrology, 277(34):195 – 213, 2003.

105. Zeng, Zhiping, Anthony K. H. Tung, Jianyong Wang, Jianhua Feng, and Lizhu Zhou. "Comparing Stars: On Approximating Graph Edit Distance". Proceedings of the VLDB Endowment, 2(1):25–36, August 2009.

106. Zhang, Linjun, Michael Small, and Kevin Judd. "Exactly scale-free scale-free networks". Physica A: Statistical Mechanics and its Application, 433:182–197, 2015.

107. Zhao, Zhi-Dan, Zimo Yang, Zike Zhang, Tao Zhou, Zi-Gang Huang, and Ying-Cheng Lai. "Emergence of scaling in human-interest dynamics". Scientific Reports, 3:3472, December 2013.

108. Zipf, George K. "Human behavior and the principle of least effort". Journal of Clinical Psychology, 6(306):573, 1949.

Vita

Fairul Mohd-Zaid graduated from Norcross High School in Norcross, Georgia, in May 2004. Upon graduating, he attended the Southern Polytechnic State University in Marietta, Georgia, where he graduated with a Bachelor of Science in Mathematics in May 2008.

He then spent a year as an intern to the Chief Scientist of the Air Mobility Command (AMC/ST) at Scott AFB. There he assisted the Chief Scientist in evaluating the possibilities of adopting new technologies into the operations of AMC, and he also coordinated numerous ST hosted visits by other Department of Defense (DoD) agencies and industries.

Following his employment, he enrolled at the Air Force Institute of Technology (AFIT) in Wright-Patterson AFB with funding from the DoD's *Science, Mathematics, And Research for Transformation* (SMART) scholarship program where he graduated with a Master of Science in Operations Research in June 2011. From there, he spent two years as an Associate Mathematician at the 711^{th} Human Performance Wing, Battlespace Visualization Branch (711HPW/RHCV) conducting research on image fusion and network visualization as well as providing data analysis consultation within the branch.

In October 2013, he returned to AFIT to pursue the degree of Doctor of Philosophy in Applied Mathematics, again, through the SMART scholarship. After completion of the program, he will return to 711HPW/RHCV to continue his research endeavors.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
07–31–2016	Doctoral Dissertation	Sept 2013 — Jul 2016

4. TITLE AND SUBTITLE

A STATISTICAL APPROACH TO CHARACTERIZE AND DETECT DEGRADATION WITHIN THE BARABÁSI-ALBERT NETWORK

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Mohd Fairul Mohd-Zaid

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

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

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENC-DS-16-S-003

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

711th Human Performance Wing, Battlespace Visualization Branch
2255 H Street
Bldg 248
WPAFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)

711HPW/RHCV

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

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Social Network Analysis (SNA) is widely used by the intelligence community when analyzing the relationships between individuals within groups of interest. Hence, any tools that can be quantitatively shown to help improve the analyses are advantageous for the intelligence community. To date, there have been no methods developed to characterize a real world network as a Barabási-Albert network which is a type of network with properties contained in many real-world networks. In this research, two newly developed statistical tests using the degree distribution and the L-moments of the degree distribution are proposed with application to classifying networks and detecting degradation within a network. The feasibility of these tests is shown by using the degree distribution for network and sub-network characterization of a selected scale-free real world networks. Further, sensitivity to the level of network degradation, via edge or node deletion, is examined with recommendation made as to the detectable size of degradation achievable by the statistical tests. Finally, the degree distribution of simulated Barabási-Albert networks is investigated and results demonstrate that the theoretical distribution derived previously in the literature is not applicable to all network sizes. These results provide a foundation on which a statistically driven approach for network characterization can be built for network classification and monitoring.

15. SUBJECT TERMS

Hypothesis Testing, Network Science, Social Network Analysis, Statistics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Christine Schubert Kabban, AFIT/ENC
U	U	U	U	175	19b. TELEPHONE NUMBER *(include area code)* (937)255-3636 x4549; christine.schubertkabban@afit.edu