DETECTION OF THE DISTRIBUTION AND PARAMETER ESTIMATION FOR THE DEPARTING CONNECTIVITY IN BIOLOGICAL NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

OMOLOLA DORCAS ODUNSI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2014

Approval of the thesis:

DETECTION OF THE DISTRIBUTION AND PARAMETER ESTIMATION FOR THE DEPARTING CONNECTIVITY IN BIOLOGICAL NETWORKS

submitted by **OMOLOLA D ODUNSI** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İnci Batmaz _____
Head of Department, **Statistics**

Assoc. Prof. Dr. Vilda Purutçuoğlu _____
Supervisor, **Statistics Department, METU**

Examining Committee Members:

Assis. Prof. Dr. Bala Gür-Dedeoğlu _____
Institute of Biotechnology, Ankara University

Assoc. Prof. Dr. Vilda Purutçuoğlu _____
Statistics Department, METU

Assis. Prof. Dr. Ceylan Yozgatlıgil. _____
Statistics Department, METU

Assis. Prof. Dr. Yeşim Aydın-Son _____
Informatics Institute, METU

Dr. Tamay Şeker _____
Central Laboratory, METU

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: OMOLOLA D, ODUNSI

Signature :

ABSTRACT

DETECTION OF THE DISTRIBUTION AND PARAMETER ESTIMATION FOR THE DEPARTING CONNECTIVITY IN BIOLOGICAL NETWORKS

Odunsi, Omolola Dorcas

M.S., Department of Statistics

Supervisor: Assoc. Prof. Dr .Vilda Purutçuoğlu

September 2014, 90 pages

The connectivity density is one of the characteristics features in the topology of the network. This density describes the total number of the in-degree and out-degree of a node in a system.

In a network, the in-degree or arriving connectivity represents the number of links coming to a target gene and the out-degree or departing connectivity stands for the number of links departing from the target gene.

For biological networks, the density of the in-degree is represented by the exponential distribution and the distribution of the out-degree is generally referred by the power-law density. But the truncated power-law, generalized pareto law, stretched exponential, geometric and combination of these densities can be also strong alternatives for the out-degree densities which satisfy the centrality and small-world properties without the scale-free feature of the biological networks.

In this study we investigate the out-degree of the biological network within the Pearson curves. For the detection, we use both real and simulated datasets and compute the moments of the data for the plausible classification of the density. Moreover we investigate the application of the three-moment chi-square and four-moment F approximations for the out-degree distributions.

Keywords: Out-degree distribution, Biological network, Topology of the network, Pearson system, Moment estimation, Three-moment chi-square approximation, Four-moment F approximation.

ÖZ

BİYOLOJİK AĞLARDA AYRILMA BAĞLATISININ DAĞILIMINI BULMA VE PARAMETRE TAHMİNİ

Odunsi, Omolola Dorcas

Yüksek Lisans, İstatistik Bölümü

Tez Yöneticisi: Doç. Dr. Vilda Purutçuoğlu

Eylül 2014, 90 sayfa

Bağlantı dağılımı, ağ topolojisinde, karakteristik özelliklerden biridir. Bu dağılım, sistemdeki düğümün dış-derece ve iç-derece toplam sayısını ifade eder. Ağ için, iç-derece veya gelen bağlantı, hedef gene gelen bağların sayısını ve dış-derece veya ayrılma bağlantısı, hedef genden çıkan bağların sayısını gösterir. Biyolojik ağlarda, iç-derece dağılımı üstel dağılımla belirtilir ve dış-derece dağılımı güç-yasası dağılımı ile ifade edilir. Fakat kesilmiş güç-yasası, genelleştirilmiş pareto-yasası, gergin üstel, geometrik ve bu dağılımların kombinasyonları, merkezcilik, öldürücülük ve küçük-dünya özelliklerini serbest-ölçek özelliği olmaksızın sağlayan, biyolojik ağlar için kuvvetli dış-derece alternatif dağılımlardandır.

Bu çalışmada, biyolojik ağda dış-dereceyi Pearson eğrileri içinde araştıracağız. Bu incelemede, ağ analizlerinde hem gerçek hem de simülasyon veri setlerini kullanacağız ve dağılımın muhtemel sınıflandırması için verilerin momentlerini hesaplayacağız. Ayrıca dış-derece dağılımlarında üç-moment ki-kare ve dört-moment F yaklaşımlarının uygulamalarının inceleyeceğiz.

Anahtar Kelimeler: Dış-derece dağılımı, Biyolojik ağ, Ağ topolojisi, Pearson sistemi, Moment tahmini, Üç-moment ki-kare yakınsaması, Dört-moment F yakınsaması.

To my husband: Samod Atanda Yusuff

The pillar behind the successful completion this programme.

ACKNOWLEDGEMENT

Thank you God for carrying me through, thank you for bringing this to pass and thank you for the uncommon favour that defiled human principles. Praise unto your holy name!!!

First and foremost my sincere APPRECIATION and THANKS goes to my amiable, kind, humble and intelligent supervisor **Assoc. Prof. Dr. VİLDA PURUTÇUOĞLU**. The simple word is that you are the BEST! You made this work a great success.

Special thanks to all the lectures of the department. I respect and salute you all. I will not fail to mention a peculiar person Assist. Prof. Dr. Ceylan Talu Yozgatligil, ma, I emulate your simplicity and kindness you are the first person I met in the department and I can never forget your kind of warm reception and the assistance rendered.

To my colleagues in department (including the senior) you are ALL such a wonderful and people to associate with. Gül İnan; I so much like your humility. Ezgi and Duygu, thanks for always being there, you guys made my stay in METU. All the research assistants, I appreciate each and every one of you even those I can't mention here. And to a dear friend Ahmed, thank you.

I will not fail to mention the great contribution of Assist. Prof. Dr. Yeşim Aydın Son head of Health Informatics Department, who in her capacity helped discovered how to get real datasets. Thank you ma, I am grateful. And also to Remzi Celebi who actually made the real datasets available, thanks for your time and great contribution.

My sincere appreciation goes to The Federal Government of Nigeria and Turkey government for the Bilateral Educatıon Award .

Above all, thanks to my parent, siblings and in-laws, for your prayers, support and encouragement from beginning till the end.

TABLE OF CONTENTS

ABSTRACT v
ÖZ vii
ACKNOWLEDGMENT ix
TABLE OF CONTENTS x
LİST OF TABLES xi
LİST OF FİGURE xii
CHAPTERS
1. INTRODUCTION.....1
1.1 Aim of study.....2
1.2 Motivation 2
2. LITERATURE REVIEW.....5
2.1 Network 5
2.2.1 Homogenous (Erdös-Renyi) network…… …….....6
2.2.2 Non homogenous network……..7
2.2.2.1 Scale free (Barabasi – Albert) network……. …….. ……...…...…8
2.2.2.2 Hierarchical network……8
2.2.2.3 Modular network…….……..9
2.3 Topology of network ….……..…….....……...…..10
2.3.1 Degree distribution ……..……..11
2.3.2 Clustering coefficient ….……..……...…… ……...14
2.3.3 Characteristics path length and diameter….. …….. ……...….15
2.3.4 Presence of hubs.....17
2.4 Characteristics of different networks…….……..…..……..….....18

2.4.1 Random (Erdös Renyi) network …….....18
2.4.2 Scale-free (Barabasi-Albert) network ..…….…..……19
2.4.3Hierarchical and modular networks….. ……...……..21
3. METHODOLOGY.....25
3.1 Pearson system ……...…….……25
3.2 Classification and selection of distribution under Pearson system …….…….28
3.3 Alternative degree distributions and parameter estimation.....28
3.3.1Generalized Pareto distribution….. …….……...28
3.3.2 Geometric distribution.....32
3.3.3 Strechted exponential distribution.....34
3.3.4 Truncated power law.....36
3.4 Three moment chi-square and Four moment F approximation.....37
3.4.1 Three- moment chi-square approximation.....38
3.4.2 Four moment F approximation.....38
3.5 Stimulation study.....39
3.6 Measures of goodness of fit 40
4. RESULT OF ANALYSES.....41
4.1 Description of the real datasets and their network graphs ……....…42
4.2 Results of the real data analyses 47
4.3 Results of the stimulated data analyses 55
4.4 Three-moment chi-square and four-Moment F approximations results……...57
4.5 Test of alternative distributions.....59
5. DISCUSSION AND CONCLUSION .....……..…69
5.1 Discussion 69
5.2 Conclusion 70
REFERENCES ... ……75
APPENDICES
A DESCRIPTION FOR THE PEARSON CURVES
A.1 Areas of the Skewness and kurtosis for the Pearson curve……..…....79
A.2 Distributions under the Pearson curve……... …….…….....80
A.3 Pearson table of distribution function.....81

B

CODES OF THE REAL DATA SETS……...83

B.1 Codes for the network analysis.....83

B.2 Codes for the stimulation study.....89

LIST OF TABLES

TABLLES

Table 1: Summary of the network analysis for the Paget disease …….47

Table 2: Summary of the network analysis for the Menkes disease ……..……...48

Table 3: Summary of the network analysis for the Inflammatory bowel disease.....48

Table 4: Summary of the network analysis for the Glycogen storage disease.....48

Table 5: Summary of the network analysis for the Muscle disease …….....49

Table 6: Summary of the network analysis for the Lafora disease ……...49

Table 7: Summary of the network analysis for the HIV disease with 1469 gene interactions …….49

Table 8: Summary of the analysis for the HIV 1152 diseases´ interaction analysis ……...…..50

Table 9: Summary of the network analysis for the HIV disease with 722 gene interactions. ……50

Table 10: Summary of the network analysis for the HIV disease with 306 gene interactions.……....50

Table 11: Summary of the 1000 Monte Carlo iterations…….....56

Table 12: Result of the three-moment chi-square approximation which is detected by the inequality given in the third column…… ……..…….....58

Table 13: Result of the four-moment F approximation which is detected by the inequalities given in the third and fourth column…….....……..……..….58

Table 14: Summary of chi-square test for the three distributions (weibull, pareto and geometric) at 5% level of significance 60

LIST OF FIGURES

FIGURES

Figure 1: Example of the view of biological network……..……5

Figure 2: A figure of random network ……7

Figure 3: A figure of scale-free network.……8

Figure 4 : Example of the view of a hierarchical network ….……9

Figure 5: A figure of modular network……..10

Figure 6: Figures of degree distribution……...12

Figure 7: Graphical representation of the Paget disease taken from the OMIM database. ……..……...42

Figure 8 : Graphical representation of the Menkes disease taken from the OMIM database. ……....43

Figure 9: Graphical representation of the inflammatory bowel disease taken from the OMIM database. ……..……..…….43

Figure 10: The graphical representation of the glycogen storage disease taken from the OMIM database. ……..……....44

Figure 11: Graphical representation of the Muscle disease taken from the OMIM database ……..44

Figure 12: Graphical representation of the Lafora disease taken from the OMIM database ……..45

Figure 13: Three dimensional illustration of the HIV ..……...45

Figure 14: Graphical representation of the HIV disease's interactions with 306 genes. ……46

Figure 15: Graphical representation of the HIV disease's interactions with 1152 genes ……..46

Figure 16: Graphical representation of the HIV disease's interaction with 722 genes. ……46

Figure 17: The graphical representation of the HIV disease's interactions with 1469 genes. …….47
Figure 18: Graph of the beta distribution for the network of the Paget disease. ……51
Figure 19: Graph of the beta distribution for the network analysis of the Menkes disease. ……..52
Figure 20: Graph of the beta distribution for the network analysis of the Muscle disease ……...52
Figure 21: Graph of the beta distribution for the network analysis of the Lafora disease ……...…….52
Figure 22: Graph of the beta distribution for the network analysis of the HIV disease with 306 gene interactions. ……...…….53
Figure 23: Graph of the beta distribution for the network analysis of the HIV disease with 1152 gene interactions ……...……53
Figure 24: Graph of the beta distribution for the network analysis of the HIV disease with 722 gene interactions. ……...…….53
Figure 25: Graph of the beta distribution for the network analysis of the HIV disease with 1469 gene interactions ……..……..……..…54
Figure 26: Graph of the beta distribution for the network analysis of the inflammatory bowel disease. ……...……..54
Figure 27: Graph of the beta distribution for the network analysis of the Glycogen storage disease. ……..……55
Figure 28:The Q-Q plots of the pareto distribution against the theoretical distribution……..61
Figure 29:The Q-Q plots of the Weibull distribution against the theoretical distribution……..62
Figure 30:The Q-Q plots of the geometric distribution against the theoretical distribution……..63
Figure 31:Histogram and density plot of the original data sets……...…..64
Figure 32:Band graph of the Skewness under Pearson Type I family……...65
Figure 33:Band graph of the Skewness under Pearson Type VI family……....65
Figure 34:Band graph of the kurtosis under Pearson Type I family…….....66

Figure 35:Band graph of the kurtosis under Pearson Type VI family……...66
Figure 36:Location of each of the dataset on the Pearson plane……..……..67
Figure 37:Pearson curve which indicates the relation between the Skewness and the kurtosis ……...79
Figure 38:Distribution cover a wide region in the Skewness-kurtosis plane ……....80

CHAPTER 1

INTRODUCTION

In the past decades, the term network is one of the most popular terms used in different disciplines including sciences from engineering to biology. In all these fields the concepts and ideas vary, thereby, the theories and properties have distinct assumptions. In simple word, it is a structure that devices a mechanism specialized on various functions. For example, the brain is a network of nerve cells joined by axons. Here the cell is a network of molecules connected by biochemical reactions.

The networks can be extended to societies, families, friendship and also professional ties. The large networks can be described on food webs, ecosystem, internet, power grids and so on. We can also describe the language that we are using here as networks in such a way that every though is a network which is made up of connected words.

Despite the usefulness and importance of networks, scientists had little knowledge of understanding their structures and properties. For instance how some systems (networks) function after the important nodes in the system have failed. Recently, the investigations from various field discovered that the most large networks (e.g. biological networks) has relatively small number of nodes that are connected while few has numerous connected nodes. These highly connected nodes are also named as hubs. The networks featured with such important nodes and hubs are referred to as the scale-free. Various biological networks, namely, gene interactions (Tong et al., 2004), gene expression networks by various scholars (Featherstone and Broadie, 2002; Agrawal 2002; Bergmann et al., 2004; Van Noort et al., 2004), protein- protein networks (Jeong et al., 2004), yeast transcriptional regulatory network (Nabil Guezli et al., 2002) and many others have been described to have a scale-free nature.

Many say biological networks are scale-free; it is the degree (i.e the number of connections per node) distribution that has this scale-free nature and not the network itself. Hereby, a particular importance is the degree distribution of nodes in a network and it is one of the important measures of the network topology.

1.1 Aim of study

The goal of this study is to establish the degree distribution of the departing connectivity in the directed biological networks. In this study, we investigate the out-degree of the biological networks within the Pearson curves. For the detection, we use the real and simulated datasets which are generated under various scenarios. The original datasets are from directed networks while the stimulated study contains data from undirected networks. Hence in the stimulated study, we assumed the in-degree and out-degree are the same. Then we compute the moments of these data for the plausible classification of the degree density.

In the analyses, we compare our results with other alternative distributions in the literature for the departing connectivity and consider whether the three-moment chi-square and four-moment F distribution are suitable for the detection of the distributions when the final Pearson densities have undefined forms.

Lastly, we test if the original datasets follows any of the defined alternative distributions for the departing connectivity.

1.2 Motivation

In recent years, the fact that the biological networks are scale-free in nature became very clear. It is also clear that the biological networks have a modular topology with a high clustering coefficient. The presence of these topology features restates the role of the degree in biological networks. Past works, such as the study of the hierarchical organization of modularity in metabolic networks (Ravasz et al., 2002), the

topological and the casual structure of the yeast transcriptional regulatory network (Guelzim et al., 2002), the network biology for understanding the cell´s functional organization (Barabasi and Oltvai, 2004) and the effects of the sampling on the predictions of the topology for the protein-protein interaction networks (Jing-Dong et al., 2005) and many others show that the out-degree distribution follows a power-law distribution with hubs at the tail of the distribution which plays a crucial role in scale-free networks.

However, the study of Khanin and Wit (2006) contradicts the assumed distribution. They studied 10 published different biological datasets and found out that the degree (i.e. the number of connections) distribution significantly differs from the power-law distribution and more so not scale-free. In their work, they suggested four alternative distributions, namely, the truncated power-law, the generalized pareto-law, the stretched exponential and the geometric distribution which may best describe the out-degree distribution.

Hence our interest is on these listed distributions. This analysis will help us to understand the structure of a biological network which helps for discovery the hidden features and more topological properties.

Accordingly the remaining chapters of this study are structured as follows: Chapter 2 presents the literature review on networks. It entails the type of networks, its classification based on different approaches and also the topological features which include the degree distribution, clustering coefficient, characteristic path length and diameter, presence of hubs and the network robustness. Also Chapter 3 reviews the methodology used in this study. In this chapter we discuss the Pearson system; some common approximation methods to find distributions under this system. Furthermore, we define two of the alternating densities for scale-free networks and how these alternating densities are defined in the Pearson system. We estimate their parameters by using the moments and moment generating functions.

Chapter 4 contains the results of real datasets, Monte Carlo runs, three and four moment's approximations. Moreover, we present the summary of each analytical

finding and graphical output. Finally in Chapter 5 we describe the summary of our study in the form of conclusions and give recommendations for the future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Network

A network is a series of points or *nodes* interconnected by communication paths or *genes* constructing multiple modules to perform one or more biological processes. The term node refers to a point in a network topology at which lines intersect. While genes are a set of instructions that decide what the organism is like, how it survives, and how it behaves in its environment.

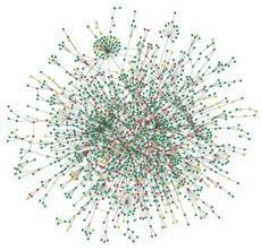


Figure 1: Example of the view of a biological network.

Networks can be classified based on their

i. Components,
ii. Links or connections and
iii. Distribution of the links.

Networks can be further simplified based on the above classification as summarized below:

Classification based on components:

a. Networks on the microscopic scale
   1. Transcription regulation networks
   2. Signal transduction networks
      i. Protein interaction networks
      ii. Metabolic networks
b. Networks on the macroscopic scale
   1. Food webs
   2. Ecological networks
   3. Phylogenetic networks

Classification based on links:

a. Directed networks
b. Undirected networks

Classification based on the distribution of links

a. Homogenous networks
   1. Random networks
b. Non homogenous networks
   1. Scale-free networks
   2. Hierarchical networks
   3. *Modular networks*

The focus of this work is on the type of networks based on the distribution of links.

2.2.1 Homogenous (Erdös-Renyi) network

This is also known as the *random* network**.** The homogenous network is a network in which each node has almost the same numbers of connections (i.e. links). The model of a random network begins with N nodes and links each pair of nodes with probability p, which gives a graph with approximately pN(N-1) randomly placed links.

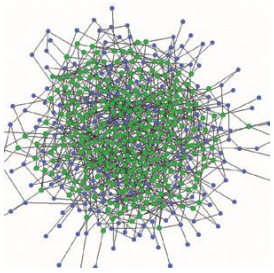


Figure 2: A figure of random network.

The random network model is a set of N_v nodes with each pair of nodes connected with an equal probability of $p \leq 1$. The number of edges N_t is the random variable and its expected value is found as $N_E = p\, N_v(N_v-1)/2$.

Moreover, the degree distribution of the model is given by the binomial distribution that becomes approximately the poisson density in the limit of large networks. (i.e. $N_v \rightarrow \infty$). The probability of a node that has a degree k is $p(k) \approx e^{-k}(k)^k/k!$. The poisson distribution shows that most nodes have approximately the same number of connections or links.

Furthermore the tail (i.e. the degree distribution P(k)) decreases exponentially. This shows significantly that the nodes rarely deviate from the average and the clustering coefficient is independent on a node´s degree. Additionally the mean path length (L) is proportional to the logarithm of the network size (N), i.e. $L \sim \log N$, which indicates the small-world property. (Wit et al. 2010)

2.2.2 Non-homogenous network

In these networks each node has different number of links. These networks are grouped into the three branches as stated below:

i. Scale-free, also called the Barabasi –Albert, networks,
ii. Hierarchical networks,
iii. Modular networks.

2.2.2.1 Scale-free (Barabasi – Albert)network

The term scale-free means a system defined by a functional form f(x) that remains unchanged within a multiplicative factor under a rescaling of the independent variable x. The scale-free networks are characterized by a power-law degree distribution. In other words, the probability that a node has k links connections or links follows $P(k) \sim k^{-y}$ where y is the degree exponent. A special property of the scale-freeness is invariance to changes in the scale. The scale invariance property is mostly interpreted as the self-similarity. Any part of the scale-free network is stochastically similar to the whole network and the parameters are assumed to be independent of the system size (Jeong et al., 2000).

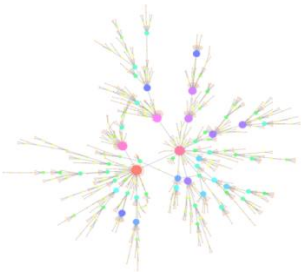


Figure 3: A figure of scale-free network.

A scale-free network has some intriguing properties.

a. There are a lot of hubs in the biological networks and a large number of nodes with few connections.
b. It belongs to the class of the small world networks (Amaral et al., 2000), which allows fast communication between different nodes.
c. It is robust to random breakdowns.

Meanwhile the properties stated above are not unique for scale-free networks, as other networks can exhibit some of the properties as well. (Barabasi and Oltvai, 2004 and Wit et al. 2010)

2.2.2.2 Hierarchical network

This is a network consists of nodes and links interconnecting the nodes. The network consists of disjoint sets of nodes, denoted as clusters. In addition, each cluster

contains at least one hub node. In hierarchical networks, the degree of clustering characterizing the different groups follows a strict scaling law, which can be used to identify the presence of a hierarchical organization in real networks.

Many real networks, such as the www(worldwideweb), actor network, the Internet at the domain level, and the semantic web obey this scaling law, indicating that the hierarchy is a fundamental characteristic of many complex systems.

If a scale-free network has clusters or modules connected to each other iteratively, giving a tree like structure, the system is referred to as *hierarchical network.* A hierarchical architecture implies that sparsely connected nodes are part of highly clustered areas with communications between different highly clustered neighborhoods being maintained by a few hubs. (Barabasi and Oltvai, 2004., Wit et al. 2010, and Davis and Barabasi, 2003).

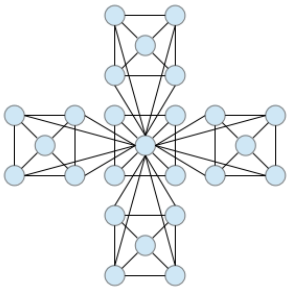


Figure 4: Example of the view of a hierarchal network.

2.2.2.3 Modular network

A module is topologically defined as a subset of highly inter-connected nodes which are relatively sparsely connected to nodes in other modules.

In the literature, often the terms hierarchy and modularity have been used almost interchangeably, although, they represent distinct properties of network. However, it is interesting to note that these two networks have been found to have similar network properties. Most of the complex systems seen in real life have associated dynamics and the structural properties of modular networks. The networks have been sought to be linked with their dynamical behaviour.

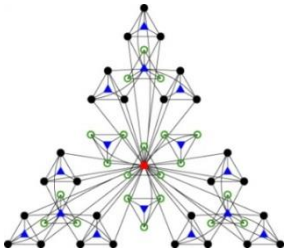


Figure 5: A figure of modular network

In hierarchical network, if the modularity seems non-hierarchical, it is called as the *modular network.* Both hierarchical and modular networks have clear modularity designs, but, they are not necessary to be scale-free. (Barabasi and Bonabeau 2003., Barabasi and Oltvai, 2004., Wit et al, 2010. And Han et al, 2004)

2.3 Topology of network

In the determination of the network topology, the configuration of its nodes and the connecting edges are relevant for the assessing network stability, dynamics and function and ultimately for being able to design and reengineer the networks of interest. Only recently has it become possible to discern the topology of large complex networks.

The biological networks are different in terms of connections and structures of nodes like their modularities and randomness. In order to differentiate them, we need to define some measures which are quantitative criteria describing the pattern of the genomic connectivity. These criteria, as listed below, are the topological features or topological measures. (Barabasi and Oltvai, 2004., Junker and Schreiber, 2008.)

1. Degree distribution,
2. Clustering coefficient,
3. Characteristic path length and diameter,
4. Presence of hubs and network robustness.

Topological features can indicate differences based on the directed and undirected networks. The degree of distribution, flux of the reaction and presence of hierarchical

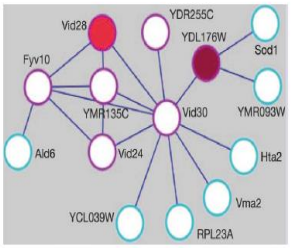
modularity can be computed through direct networks since the calculations are found by directions of connections. While other measures such as clustering coefficient, characteristics path length and diameter, presence of hubs and the network robustness can used for both directed and undirected networks. (Wit et al., 2004)

2.3.1 Degree distribution

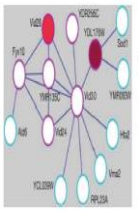
The degree distribution $\rho_{(k)}$ is one of the most prominent characteristics of network topology. The degree distribution k_i is defined as the number of arcs adjacent to the nodes. In a network without self loops, which are the arcs connecting a node to itself, and multiple links which indicate two nodes connected by more than one arc, the degree equals to the number of neighbours of the nodes.

The degree distribution $\rho_{(k)}$gives the probability that a chosen node has exactly k links. In a system, the number of connections of each node can be described by a probability distribution. $\rho_{(k)}$ is obtained by counting the number of nodes N(k) with k links (connections) and dividing by the total number of nodes N. The degree distribution helps to identify different classes of networks. For example, a peaked degree distribution shows that the system has a characteristic degree and that there are no highly connected nodes. On the other hand, a power-law degree distribution shows that a few hubs bind together to many small nodes.

If our focus is on directed networks, two types of connectivity can be observed. These are one from the number of links coming to the target gene and one from the number of links departing from the target gene. If the number of gene which regulates is observed, this is called the *incoming connectivity or arriving connectivity* or *in-degree denot*ed by $k_{in.}$ On the other hand, if the number of genes which are regulated is observed, this is called the *outgoing connectivity or departing connectivity* or *out-degree* denoted by $K_{out.}$. (Junker and Schreiber, 2008., Wit et al, 2010.)



Node	Degree
Vid30	10
Fyv10	5
YMR135C	4
Vid24	3
Vid28	3
YDL176W	3
Ald6	1
YDR255C	1
Sod1	1
YMR093W	1
Hta2	1
Vma2	1
RPL23A	1
YCL039W	1



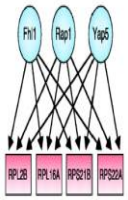
YDL176W

Degree: 3

dTMP

In degree: 3

Out degree: 2



Fhl1

Out degree: 4

In degree: 0

RPL2B

Out degree: 0

In degree: 3

Figure 6: Figures of degree distribution.

In biological networks, the distribution of K_{out} is generally referred by the power-law density denoted by

$$\rho_{k=C_0 K^{-\gamma}}, \quad (1)$$

where Υ is referred to the power-law exponent, and K represents the average distance between any two nodes in a system.

In every directed networks, the calculation of the degree distribution has different form, we can write k in an undirected network as $k = K_{in} + K_{out}$.

Otherwise, other strong alternative density such as the *truncated power-law distribution* can be considered as shown in Equation (2).

$$\rho_{(k)}=\frac{e^{-k/k_c K^{-\gamma}}}{c_1(\gamma,k_c)} \quad (2)$$

In Equation (2), C_1 (Υ, k_c) is the normalizing factor, Υ represents the power-law exponent while k_c is the cut-off parameter. (Khanin and Wit, 2006)

The studies on analyses of benchmark network datasets revealed that besides power-law and truncated power-law distributions, the stretched exponential, the generalized pareto law distribution and combination of these densities satisfy some characteristics of biological networks, that is, the centrality and lethality (small world properties) without the scale-free features.

On the other hand, the information on the number of connectivity can also be computed by the average number of degrees for both directed and undirected systems. This topological measure is known as the *average degree* or *connectivity* of the network that is computed by μ in Equation (3).

$$\mu = \sum_{i=1}^{N} k_i / N \quad (3)$$

where N shows the total number of nodes and k_i presents the number of links associated to the ith node or gene.

Also, if our interest is on the undirected networks in which the connection doesn't specify the targets of transcriptional factors, the probability distribution of links for a node can be found by giving to k other nodes. This conditional probability is called the *connectivity distribution* denoted by p_k and the total number of links attached to the ith gene or node (i= 1, ...,N), this is named as the *degree* or *connectivity* of the node i shown by k.

The degree or the average degree distribution of a system enables us to distinguish different networks via their connectivities by using the following approaches.

1. The first approach is to draw a Q-Q plot between the connectivity relative distributions $p_{k.}$ after which the number of the nodes k is observed. We check the density of the graph by observing the graph on a log-log scale; if it's a straight line then we conclude as the power-law density. Another way to use the graphical test is by observing the correlation coefficient R^2 between the numbers of connection (k) on a log scale. If the value of R^2 obtained is high (close to 1), then we conclude that there is evidence of the power- law density.
2. The second approach is by estimating the power-law exponent Υ in $p_k \alpha k^{-\Upsilon}$ where $k \geq 1$ for the dataset. Υ is derived by the maximum likelihood method under the independence assumption of the connectivity for the node i (i= 1,..,N) in a large network.

$$L(\frac{\Upsilon}{k}) = \Pi_{i=1} K_i^{-\Upsilon} / (\zeta/\Upsilon).$$

Here, N-1 is the maximum number of connectivity in an N-dimensional system, C_0 in Equation (1) is also called the Riemann zeta function (ζ/Υ).

Υ in the power-law distribution cannot be easily computed, hence an iterative techniques such as Newton-Raphson is used to compute the estimate of Υ, i.e $\widehat{\Upsilon}$. Once the estimate $\widehat{\Upsilon}$ is computed, the scale-freeness property is checked by the chi-square goodness of fit test under the power-law with exponent Υ as shown in Equation (4). Here, the estimated connectivity E_k and the observed connectivity O_k can be computed from the data

$$X^{*2}= \sum_{k=1}(O_k - E_k)^2/k \sim X^2_{\alpha,j^*-2} \quad \text{where } j^* = j\geq 5. \tag{4}$$

In Equation (4), $\chi^2_{\alpha,\ j^*-2}$ gives the chi-square critical value with j^{*-2} degree of freedom for a given significance level α. (Wit et al, 2010., Stefano et al, 2009.)

2.3.2 Clustering coefficient

Another basic measure of the network topology is the *clustering coefficient* C_i. The clustering coefficient relates to the local cohesiveness of a network and measures the probability that two nodes with a common neighborhood are connected (Junker and Schreiber, 2007).

More precisely, the clustering coefficient of a node is the ratio of existing links connecting a node's neighbors to each other to the maximum possible number of such links. The clustering coefficient for the entire network is the average of the clustering coefficients of all the nodes. For the node i, it is formulated as

$$C_{i=}\frac{2e_i}{k_i(k_i-1)},$$

where k_i is the number of neighbours of the i^{th} node and e_i is the number of connections between these neighbours. Accordingly, the maximum possible number of connections between neighbors is found as

$$\binom{k}{2} = \frac{k(k-1)}{2}.$$

In this expression, C_i lies between 0 and 1. If $C_i = 0$, it means that the nodes are not connected or totally dispensed, while $C_i = 1$ shows that the nodes are totally connected. Moreover, the high clustering coefficient for a network is an indication of a small world property of the system. (Wit et al, 2010)

On the other side, the clustering coefficient depends on the total number of links which is based on network types, i.e. directed or undirected. For instance, if the network is directed but no self-loops, Ci can be computed as

$$C_i = \frac{e_i}{k_i(k_i-1)} .$$

As we compute the clustering coefficient for individual nodes, a unique value for whole system can be computed by averaging these coefficients. Hereby, the average clustering coefficient is characterized by the overall tendency of nodes to form clusters of groups and is denoted by μ_c. For example, if the network is directed and has no self loops, μ_c is represented as

$$\mu_c = \sum_i \frac{2e_i}{k_i(k_i-1)/N} .$$

Strictly speaking, the clustering coefficient C_i is not a property of node n_i itself, but rather a property of its neighbours. The global or mean clustering coefficient $C = C_i$ of the network is the average cluster coefficient of all vertices. Many empirical networks exhibit a rather high clustering coefficient, indicating a local cohesiveness and a tendency of nodes to form clusters or groups. (Junker and Schreiber, 2008., Wit et al., 2010.)

2.3.3 Characteristics path length and diameter

The distance in networks is measured with the path length, which tells us how many links we need to pass through to travel between two nodes. The *characteristic path length* or *shortest path length*, denoted by L, represents the shortest distance between any two nodes. As there are many alternative paths between two nodes, the shortest path, that is the path with the smallest number of links between the selected nodes, has a special role. In a graphical representation, the characteristic path length as

described in Equation (5) refers to the minimum number of links or edges to move from one node to other node.

$$L = 2\sum_i \sum_j \frac{l_{ij}}{N(N-1)}, \quad (5)$$

where l_{ij} is the shortest distance between the two nodes i and j.

In a direct network, the distance from node A to node B (l_{AB}) can be different from the distance between node B and A (l_{BA}). So the shortest distance (L) between the nodes is the min { $l_{AB,}$ l_{BA} }. In the undirected network, l_{AB} and l_{BA} can be the same, i.e, l_{AB}= l_{BA}, because the destination is not considered.

Accordingly, the mean or the average path length represents the average over the shortest paths between all pairs of nodes and offers a measure of a network´s overall navigability.

The average path length is calculated by finding the shortest path between all pairs of nodes, adding them up, and then dividing by the total number of pairs. This shows us, on average, the number of steps it takes to get from one member of the network to another.

On the other hand, if our concern is on the longest distance, rather than the shortest path, this measure is known as the *diameter* D. It can be computed by D= max { l_{ij}}. In other words, we can define the diameter as the longest of all the calculated shortest paths in a network. That is, once the shortest path length from every node to all other nodes is calculated, the diameter is the longest of all the calculated path lengths. The diameter is representative of the linear size of a network. Biologically, for both L and D, the small values present the fast actions and the big ones refer to slower actions within intermediate stages of the system. (Wit et al, 2010). The shortest path grows proportionally in logarithm to the number of nodes in the network, i.e L α logN. (Bing Zhang, 2009). Specifically there is no mathematical definition for the L and D in networks.

Moreover, both L and D can be used in the evaluation of the flux of interactions that is for the interpretation of the speed of communication between nodes. If a system has very small D and L, with large clustering coefficient and the power exponent of the out-degree distribution ϒ satisfies ϒ > 3, we can say that the system satisfies the

small-world property. But this is not the case in the biological networks. In biological networks, l is always shorter results in ϒ lying between 2 and 3. This feature is called as the ultra small-world property. Both the small-world and the ultra small-world properties stand for the modular structure in a system. This is another common feature of biological networks. (Junker and Schreiber, 2008., Barabsi and Oltvai, 2004.)

2.3.4 Presence of hubs

The number of connections for nodes shows a heterogeneous structure in biological networks. That is, most of the genes have very few links with other nodes and few of these genes possess many links with others. The highly connected nodes are called the *hubs* or *global regulators.* If the shortest path L is small, then the validity of this feature in a system is sure.

In addition, the network robustness can be observed if there exists hubs in the network. In networks, if the hubs are kept, they help to maintain the modules, i.e. the major functional groups. Otherwise, the network can be divided into the isolated node clusters which may bring about a lethal disability in some functions.

Therefore, the existence of hubs in a network controls the actual connectivity of the pathway and this feature can be remarked as the *centrality principle.* Because its presence in network has the abilities to direct the overall system, which can also be remarked by the *lethality principle.*

On the other hand, the network robustness can be detected through different approaches. For this purpose, one can compute the characteristic path length since it shows the connectivity in a system. Thereby, if the system maintains the same path length after the removal of random nodes, this is an evidence of robustness. Also the entropy measure can be implemented to observe the change in the system after the random attacks. (Han et al. 2005., Guelzim et al, 2002., Wit et al, 2010.)

2.4. Characteristics of different networks

The network types are quite different in their topological features. Their differences can be described and contrasted based on the distribution of their links. Major topological properties of different network types are discussed below.

2.4.1 Random (Erdös Renyi) network

This network belongs to the homogenous type of networks. Here, each node in the system has a similar number of interactions K. The network starts with *N* nodes and connects every pair of nodes with probability p. This creates a graph with approximately pN(N-1)/2 randomly distributed edges. Accordingly, the distribution follows a poisson distribution and consequently, the average degree *k* of the network describes the properties of a typical node. The *poisson distribution* has mean μ_k with the following distribution function.

$$\rho_k = \lambda^k expe^{-\lambda/k!}. \quad (6)$$

In Equation (6), λ is the mean number of connections per node.

On the other hand, for totally N number of nodes in a system (i=1,…,N), the random networks have clustering coefficients C_i's, which are invariant to the degree of the node and most of C_i's is approximately close to each other. This makes random networks different from most of the non-homogenous networks.

Moreover in random networks, there are no hubs and clusters due to the fact that the nodes are not so connected. Also the number of connections k, in the system is not related to the average clustering coefficient μ_k. In other words, they are independent. Hence the plot of $C_{(k)}$ against k shows a horizontal straight line when it is drawn on the original scale. This reveals that there is an essential modularity in the construction of the network.

Furthermore, in random networks, there is a proportional relationship between the mean path length μ_L and the logarithm of the total number of nodes N, i.e. $\mu_L \alpha$ logN. The proportionality relation indicates that the random networks do not exhibit the

small world property. This may be as a result of the links that are poisson distributed. Finally of metabolic reactions in random networks have a linear pathway when there are stable concentrations in all metabolites and we cannot observe any traces of the power- law property in the distributions of the fluxes. (Junker and Schreiber, 2008., Wit et al, 2010.)

2.4.2 Scale-free (Barabasi-Albert) network

A scale-free network was discovered by Albert-Laszlo Barabasi, and two of his students. They mapped the topology of a small part of the World Wide Web and they discovered that some group of nodes (hubs) had many more connections than others. This structure did not map the model of random networks. Thus, they concluded that the network had a power-law distribution of the number of links connecting to a node. (Almaas et al, 2007)

We can simply say that in scale-free networks the nodes are not randomly nor evenly connected. But it includes many "very connected" nodes, i.e. hub and these connectivity's shape the operation of the network. On the other hand, the ratio of connected nodes to the number of nodes in the rest of the network remains fixed even as the network expands. As a result, the scale-free networks cannot be easily degraded as random nodes fail. Because they are very highly connected and a lot of random failure can be realized before the hubs disappear. Whereas, if the hubs are chosen, then the major regulation of the network can fail. Thus, the connections in the scale-free networks are maintained under random conditions.

Additionally, the scale-free networks have two main ingredients. These are the *growth* and *preferential attachment.*(Barabasi et al, 2002.). The growth means that the number of nodes increases in the network by time while the preferential attachment refers to the assumption that new nodes will connect with nodes with large degrees. In summary, the scale-free networks are made up of many nodes, but with only a few connections and the network is held together by a few highly

connected hubs. Hereby, mathematically, we assume that the probability P that a new node will be connected to the node i depends on the degree k_i of the node i via

$$P \sim \frac{k_i}{\sum_i k_i} . \quad (7)$$

The numerical simulations and analytic results indicate that this sort of networks evolves into a scale-invariant state with the probability that a node has k edges following a power-law with an exponent $\gamma=3$ (Hidalgo and Barabasi, 2008).

Furthermore, another important topology of scale-free networks can be seen in the distribution of their clustering coefficients, which decreases as the node degree increases. This distribution follows a power-law density. Accordingly, it shows that the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through hubs. In directed networks, the degree distribution of nodes explained by the different number of nodes connected to each node via an exponential function. This function explains the highly connected and sparsely connected nodes with heavy tail densities. (Wit et al, 2010.)

For the in-degree, i.e. incoming connectivity

$$\rho_k = N_o \exp^{-\alpha k}. \quad (8)$$

In Equation (8), N_o is the normalization factor and α refers to the exponential exponent and also describes the number of the regulating genes which arrive at the same gene or node. The higher α means a higher number of genes that is directed or linked to a target gene. This reduces the number of target genes and increases the number of regulators in the network.

Also for the out-degree density, it is described by the power-law distribution. As discussed earlier, the power-law density is not unique for all biological networks. Previous works suggest the geometric, stretched exponential, truncated power-law, generalized pareto distribution and combinations of these distributions as alternatives for the power-law distribution. (khanin and Wit, 2006.). Because they maintain the characteristics of biological networks such as centrality and lethality except for the scale-free networks. In this sense, maybe it can be assumed that the scale –free data

are not scale-free indeed, but only satisfies the exponential probability function in relation to the number of connections per mode.

With the assumption that the scale-free networks have the power-law distribution, the degree of each node is proportionally formulated by $k^{-\Upsilon}$, where Υ is the degree exponent of the power-law and k is the number of links. The possible range of Υ can be found in biological networks although Υ can take any value from 2 to ∞. When it lies between 2 and 3 ($2< \Upsilon<3$), it means that the system has the ultra-small world characteristics and most of the biological network satisfies this condition. Υ can also be 3. In this case the nodes are relatively less densely connected and the shortest path length becomes proportional to log /log (logN). When $\Upsilon > 3$, the nodes are moderately less connected and the shortest path length is proportional to log N.
Furthermore, due to the dense connections in scale-free networks, we cannot estimate the average number of links per node, μ_k, unlike the random networks. This also implies that the presence of hubs in such structures. As there are different numbers of connections for each node and they are also invariant in changes of scale, a linear decreasing function on the logarithm scale is found showing that even though most of nodes have few links, very dense connections belong to only a small amount of nodes. Finally in this type of networks, the clustering coefficient for every node Ci is equal to the total number of connections k of the system. (Wit et al, 2010., Tolba, 2010.)

2.4.3 Hierarchical and modular networks

The hierarchical and modular networks are the non-homogenous type of networks. These networks are grouped based on the distribution of their links. When the system possess iterative connections of clusters or modules linked to each other resulting in a tree structure, this type of network can generate hierarchical systems without the scale-freeness. Meanwhile, if the nodes are connected to each other iteratively in absence of the hierarchy as well as the scale-freeness, a modular network is observed.

Moreover, the hierarchical network has a network topology in which a central "root" node is connected to one or more other nodes that are one level lower in the hierarchy with a point-to-point link between each of the second level nodes and the top level central "root" node. In a simpler word, it is a network that consists of small groups of nodes organized in a hierarchical manner into increasingly large groups, while maintaining the scale-free topology. Furthermore the network is significantly differs from the other similar models, i.e. Erdos Renyi and Barabási–Albert, in the distribution of the nodes and clustering coefficients. Other networks predict a constant clustering coefficient as the function of the degree of the node, while in hierarchical models nodes with more links are predicted to have a lower clustering coefficient.

More so, the Barabasi-Albert model predicts that the average clustering coefficient decreases as the number of nodes increases. While in hierarchical models there is no relation between the average clustering coefficient and the size of the network. Additionally, unlike other scale-free models, the clustering coefficient does not depend on the degree of a specific node. In hierarchical models, the clustering coefficient is a function of the degree and can be expressed as $C(K) \sim K^{-\beta}$. Accordingly, the degree distribution follows the power-law meaning that a randomly chosen node in the network has k edges with a probability

$$P(K) \sim ck^{-\Upsilon} . \quad (9)$$

In Equation (9), c is a constant and Υ is the degree exponent such that $\Upsilon \in [2, 3]$.

Many real networks are expected to be fundamentally modular, meaning that the network can be seamlessly partitioned into a collection of modules where each module performs an identifiable task, separable from the function(s) of other modules Therefore, the scale-free property must be reconciled with potential modularity. In order to account for the modularity as reflected in the power-law behavior of $C(k)$ and a simultaneous scale-free degree distribution, there is an assumption that the clusters are combined in an iterative manner, generating a hierarchical network. Such a network emerges from a repeated duplication and an integration process of the clustered nodes, which can be repeated indefinitely.

On the other hand, the hierarchical and modular networks are different from scale-free networks in the sense that the average clustering coefficients μ_k are linearly proportional to the total number of connections K with a ratio 1/k when they are plotted on the logarithmic scale. This relation shows that the most clustered areas have nodes that are sparsely connected. Also, the graph of C_k against K shows a straight line with a slope of -1. In scale-free networks the case is the reverse due to the fact that the values of C_i are non-homogenous. (Wit et al, 2010.)

In the next chapter, we deal with only the scale-free networks and its degree distributions which are taken as the power-law density. Also other alternatives densities like the generalized pareto law, the geometric and stretched the exponential distribution will be discussed.

CHAPTER 3

METHODOLOGY

In this chapter, we defined a unit of distribution families for several networks by using the Pearson system. Here, we explain this system, alternative densities for scale-free networks and how these alternative densities are defined in this system.

Also, three and four-moment approximation of chi-square and F distributions are discussed for cases where there is no unique Pearson family.

Moreover, we discuss stimulation study. The study which we will conduct, aims to compare the results of real datasets so as to arrive at reasonable conclusion.

Finally, we test the original datasets for the alternative distributions of departing connectivity of biological networks as suggested by the literature.

3.1 Pearson system

The Pearson system is a parametric family of distributions. The system was introduced by Karl Pearson in 1985. He worked out a set of four-parameter probability density functions as a solution to its differential equation (Andreew et al., 2005) as shown in Equation (10).

$$\frac{f\prime(x)}{f(x)} = \frac{P(x)}{Q(x)} = \frac{x-a}{b_0+b_1x+b_2\,x^2} \quad , \tag{10}$$

where f is a density function and a, b_o, b_1 and b_2 indicate the parameters of the distributions in the Pearson´s four-parameter system as the direct correspondence of the parameters and the central moments (μ_1, μ_2, μ_3, μ_4) of the distribution (Stuart and

Ord, 1994). The explicit forms of these parameters are presented in the following expressions.

$$b_1 = a = \frac{-\mu_3(\mu_4 + \mu_2^2)}{A} = \frac{-\mu_2^{1/2}\beta_1(\beta_2 + 3)}{A'}, \quad (11)$$

$$b_0 = \frac{-\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{A} = \frac{-\mu_2(4\beta_2 - 3\beta_1^2)}{A'}, \quad (12)$$

$$b_2 = \frac{-(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)}{A} = \frac{-(2\beta_2 - 3\beta_1^2 - 6)}{A'}, \quad (13)$$

$$\beta_1^2 = \frac{\mu_3^2}{\mu_2^3}, \quad (14)$$

$$\beta^2 = \frac{\mu_4}{\mu_2^2}. \quad (15)$$

In Equation (14) and (15), β_1^2 and β_2 present the skewness and the kurtosis, respectively. In these expressions, the scaling parameters A and A' are found as below.

$$A = 10\mu_4\mu_2 - 18\mu_2^3 - 12\mu_3^2, \quad (16)$$

$$A' = 10\beta_2 - 18 - 12\beta_1^2. \quad (17)$$

As alternative to the basic four-parameter system, various extensions have been proposed using higher order polynomials. The typical extension modifies Equation (10) by setting $P(x) = a_0 + a_{1x}$. Thus, we have

$$\frac{f(x)}{f'(x)} = \frac{P(x)}{Q(x)} = \frac{ao + a1x}{b_0 + b_1x + b_2x^2}. \quad (18)$$

This parameterization has the advantage that a_1 can be zero and the values of the parameter are bound when the fourth cumulant exists (Karvanen, 2002). One of the ways to classify the distribution generated by the roots in Equation (10) – (17) is by using the Pearson´s table. Pearson provides a solution to the equations with a table of 12 classes which are identified in Table I in Appendix.

Another alternative approach is to use two statistics as shown in Equation (19) and (20) that are the functions of the four-parameter.

$$D = b_o b_2 - b_1^2 = \alpha\beta - (\alpha + \beta)^2 \quad , \tag{19}$$

$$\lambda = b_1^2 / b_o b_2 = (\alpha + \beta)^2 / \alpha\beta \quad . \tag{20}$$

Here Equation (18) can be rewritten as

$$\frac{f(x)}{f'(x)} = \frac{P(x)}{Q(x)} = \frac{ao + a1x}{b_0 + b_1 x + b_2 x^2}$$

$$= \frac{ao + a1x}{b_2(x-\alpha)(x-\beta)} \quad \text{if } b_2; = 1$$

$$= \frac{ao + a1x}{(x-\alpha)(x-\beta)}$$

$$= \frac{m}{x-\alpha} + \frac{n}{x-\beta},$$

where

$$m = -\frac{ao - a1x}{\beta - \alpha} \tag{21}$$

and

$$n = \frac{ao - a1\beta}{\beta - \alpha} \quad , \tag{22}$$

in which the signs of D and λ are obtained for different supports of x as given below.

i. If $x \in [\alpha, \beta]$, $\alpha < 0 < \beta$ then $\alpha\beta < 0$ leading to $\lambda < 0$ and $D < 0$.

ii. If $x \in [-\infty, \alpha]$, $\alpha < \beta < 0$ or $x \in [\beta, \infty]$, $0 < \alpha < \beta$, then $0 < \alpha\beta < (\alpha + \beta)^2$ causing $\lambda > 0$ and $D < 0$.

This approach is more useful because of its simplicity and it can be easily implemented in order to

1. Estimate the moments from the data,
2. Calculate the Pearson parameters (a_0, b_1, and b_2),
3. Use the parameter to compute the selection criteria (D and λ) and
4. Select an appropriate distribution from the Pearson table based on the criteria given in Table II in Appendix.

Hence, our aim is to investigate whether the alternative distributions for the degree distribution of the protein- protein interaction network belongs to one of the Pearson system's families.

3.2 Classification and selection of distributions under the Pearson system

There are basically two ways to classify the distribution generated by the roots of the polynomial in Equation (11) - (17). One of these ways is to classify the distribution generated by the roots of the equation using the Pearson table. Pearson proposes solution to this set of equations by identifying 12 classes of distribution which are the variant of three major distributions. Because actually the Pearson family of distributions consists of seven distributions, called as Type I to Type VII. Figure (38) and (39) in the appendix show this classification (Lahcene, 2013). On the other hand the alternative approach, as stated beforehand, is to identify the distribution via the two statistics which are the functions of the four Pearson parameters as given in Table 14 in the appendix.

3.3 Alternative degree distributions and parameter estimation

As declared previously, the scale-free feature can be described under the following list of distributions. The derivations of necessary moments in order to investigate the true Pearson family are also presented under each density separately.

3.3.1 Generalized Pareto distribution

The probability and the cumulative distribution functions of the generalized Pareto distribution are given as below.

$$f(x) = \frac{ak^a}{x^{(a+1)}}$$

for $k > 0$, $a > 0$ and $x \geq k$. Hereby,

$$F(x) = 1 - \left(\frac{k}{x}\right)^{a},$$

where a and k are the scale and shape parameter, respectively.

On the other hand, the estimation of its parameters via the moment estimation method is shown as follows (Quandt, 1966).

$$k^* = \frac{(a*n-1)x_r'}{a*n} \quad \text{and} \quad a^* = \frac{n\bar{x}-x_r'}{n(\bar{x}-x_r')}.$$

If the parameters are inferred by the method of the maximum likelihood, then the following loglikelihood function and associated estimators can be obtained.

$$L = \prod_{i=1}^{n} \frac{ak^a}{x_i^{(a+1)}} \quad \text{for } 0 < k \leq \min\{x_i\}\ a > 0,$$

where

$$\hat{a} = \frac{n}{\left[\sum_{i=1}^{n} \log\left(\frac{x_i}{\hat{k}}\right)\right]},$$

$$\hat{k} = \min\{x_i\} = x_i.$$

Therefore, the moment estimators for the parameter a and k can be written via their maximum likelihood estimators by the following form.

$$a^* = (1 - 2n^{-1})\,\hat{a} \quad \text{and} \quad k^* = [1 - (n-1)^{-1})\hat{a} - 1]\hat{k}.$$

On the other side, if the parameter estimation is implemented by using the Pearson condition, the rth moment under the pareto distribution can be derived as below.

$$U_r' = \frac{ak^r}{a-r} \quad \text{where } r < a.$$

Accordingly,

$$U_1' = \frac{ak^1}{a-1},$$

$$U_2' = \frac{ak^2}{a-2},$$

$U_3' = \frac{ak^3}{a-3}$ and

$U_4' = \frac{ak^4}{a-4}$.

Hereby, the associated expectation and the variance are found as follows.

$E(x) = U_1' = \frac{ak^1}{a-1}$ where $a > 1$.

$V(x) = U_2' - (U_1')^2$

$= \frac{ak^2}{a-2} - \left(\frac{ak^1}{a-1}\right)^2 = \frac{(a-1)ak^2-(a-2)a^2k^2}{(a-1)^2(a-2)} = \frac{ak^2}{(a-2)(a-1)^2}$.

So the following equation can be solved by

$U_3 = U_2' - 3U_2' U_1' + 2(U_1')^3$

$= \frac{ak^3}{a-2} - 3\left[\frac{ak^2}{(a-2)} \times \frac{ak}{(a-1)}\right] + 2\left[\left(\frac{ak}{a-1}\right)^3\right]$.

Considering only the numerator of the above expression, we can get

$= \frac{ak^3(a-1)^3(a-2)}{a} - \frac{3a^2k^3(a-1)^2(a-3)}{b} + \frac{2a^3k^3(a-2)(a-3)}{c}$

$a = (a^3 - 3a^2 + 3a - 1)(a^2k^3 - 2ak^3)$

$= (a^5k^3 - 5a^4k^3 + 9a^3k^3 - 7a^2k^3 + 2ak^3)$.

$b = (-[3a^2k^3(a-3)(a-1)^2])$

$= (-[3a^2k^3 - 9a^2k^3(a^2 - 2a + 1)])$

$= (-3a^2k^3 + 15a^4k^3 - 21a^3k^3 + 9a^2k^3)$.

$c = \left(2a^2k^3(a-2)(a-3)\right)$

$= 2a^2k^3(a^2 - 5a + 6)$

$= 2a^5k^3 - 10a^4k^3 + 12a^3k^3$.

Then by combining a, b and c in $2a(a+1)k^3$, we can obtain the following formulas.

$$U_3 = \frac{2a(a+1)k^3}{(a-1)^3(a-2)a-3)} .$$

$$U_4 = U_4' - 4U_3'U_1' + 6U_2'(U_1')^2 - 3(U_1')^4$$

$$= \frac{ak^4}{a-4} - 4\left[\frac{ak^3}{(a-3)} \times \frac{ak}{(a-1)}\right] + 6\left[\frac{ak^2}{(a-2)} \times \frac{(ak)^2}{(a-1)^2}\right] - 3\left[\left(\frac{ak}{a-1}\right)^4\right]$$

$$= \frac{ak^4}{a-4} - 4\left[\frac{a^2k^4}{(a-1)(a-3)}\right] + 6\left[\frac{a^3k^4}{(a-1)^2(a-2)}\right] - 3\left[\frac{a^4k^4}{(a-1)^4}\right]$$

$$= \frac{ak^4\left(\begin{matrix}(a-1)^4(a-1)(a-3)-4[(a-1)^3(a-2)(a-4)a^2k^4]+\\ 6[(a-1)^2(a-3)(a-4)a^3k^4]-3[(a-2)(a-3)(a-4)a^4k^4]\end{matrix}\right)}{(a-1)^4(a-2)(a-3)(a-4)}.$$

Later, if we consider merely the numerator, we can find the expressions below.

a = $ak^4(a-1)^4(a-1)(a-3)$,

b = $-4[(a-1)^3(a-2)(a-4)a^2k^4]$,

c = $6[(a-1)^2(a-3)(a-4)a^3k^4]$ and

d = $-3[(a-2)(a-3)(a-4)a^4k^4]$.

Finally, we combine a,b,c and d as follows $(9a^3+3a^2+6a)k^4$ and get

$$U_4 = \frac{(9a^3+3a^2+6a)k^4}{(a-1)^4(a-2)(a-3)(a-4)} = \frac{3a(3a^2+a+2)k^4}{(a-1)^4(a-2)(a-3)(a-4)} .$$

Then, we obtain the skewness

$$\beta_1 = \frac{(U_3)^2}{(U_2)^3} = \left[\frac{2a(a+1)k^3}{(a-1)^3(a-2)(a-3)}\right]^2 \div \left[\frac{ak^2}{(a-2)(a-3)}\right]^3$$

$$= \frac{2(a+1)}{a-3} \cdot \sqrt{\frac{a-2}{a}} \quad \text{for a>3}$$

and the kurtosis

$$\beta_2 = \frac{(U_3)}{(U_2)^2} = \left[\frac{3a(3a^2+a+2)k^4}{(a-2)(a-3)(a-4)}\right]^3 \div \left[\frac{ak^2}{(a-1)^2(a-2)}\right]^2$$

$= \frac{3(3a^2+a+2)}{(a-3)(a-4)} . \frac{(a-2)}{a} = \frac{3(a-2)(3a^2+a+2)}{a(a-3)(a-4)}$ for a>4.

Here, as a → ∞, $\beta_1 \rightarrow 2$ and $\beta_2 \rightarrow 9$.

3.3.2 Geometric distribution

There are two cases for the geometric distribution which are as below.

1. The probability that the kth trial (out of k trials) is the first success is found as below.
   Pr(x=k) = $(1- P)^{k-1}P$ for k=1, 2,3..... and
   F(x)= $\sum(1-P)^{K-1}P$.
2. The number of failures until the first success
   Pr(Y=k) = $(1- P)^{k}P$ for k=1, 2,3.....

Hereby, considering the first case, the parameters estimation using the moment generating function is obtained via

G (t) = $\sum_{k=1}^{\infty} e^{tk}(1-p)^{k-1}\,p$,

$= \sum_{k=1}^{\infty} e^{tk}q^{k-1}$,

$= Pe^{t}\sum_{k=1}^{\infty}(e^{t}q)^{k}$,

$= \frac{Pe^{t}}{1-e^{t}q}$,

and its derivatives is

$$G'(t) = \frac{pe^{t}}{(1-qe^{t})^{2}}.$$

Thereby, the mean E(.) and the variance V(.) of the random variable *x* are calculated as below.

E(x)= $G'(0) = \frac{1}{p}$.

V(x)= $G''(0) - (G'(0))^2$ when $G''(0) = \frac{2-p}{(p)^2}$

$= \frac{2-p}{(p)^2} - \frac{1}{(p)^2}$.

Here, $G'(0)$ and $G''(0)$ show the first and second derivative of the given function respectively.

On the other hand, the second case can be described as follows.

P(k) = $p(1-p)^k = pq^k$ for 0<p<1 and q= 1-p.

F(x) = $\sum_{k=0}^{\infty} p(k) = 1 - q^{k+1}$.

By using the moments to estimate the parameters,

$U_k' = \sum_{x=0}^{\infty} p(x)x^k = \sum_{x=0}^{\infty} p(1-p)^x x^k$.

Therefore, the first four moments are found by

$U_1' = \sum_{x=0}^{\infty} p(1-p)^x x = \frac{1-p}{p}$,

$U_2' = \sum_{x=0}^{\infty} p(1-p)^x x^2 = \frac{(2-p)(1-p)}{(p)^2}$,

$U_3' = \sum_{x=0}^{\infty} p(1-p)^x x^3 = (1-p)(6 + (p-6)6)$

and

$U_4' = \sum_{x=0}^{\infty} p(1-p)^x x^4 = (1-p)(2-p)(12 + (p-12)p)$ as

$U_k' = \sum_{x=0}^{\infty} p(x)(x - \frac{1-p}{p})^k$.

In these derivations, the kurtosis and skewness of both cases are the same. Hence, we choose the second case in order to derive the skewness and kurtosis when the mean is $\frac{1-p}{p}$ and the variance is $\frac{1-p}{p^2} = \frac{q}{p^2}$. Thus, for $U_3 = \frac{(p-1)(p-2)}{p^3}$ and $U_4 = \frac{(p-1)(-p^2+9p-9)}{p^4}$, the skewness and the kurtosis are found as

$\beta_1 = \frac{(U_3)^2}{(U_2)^3} = \left[\frac{(p-1)(p-2)}{(1-p)^3}\right]^2$,

$\sqrt{\beta 1} = \frac{(2-p)}{\sqrt{1-p}}$,

$$\beta_2 = \frac{(U_4)}{(U_2)^2} = \frac{(p-1)(-p^2+9p-9)}{p^4} \cdot \frac{(p^2)^2}{(1-p)^2},$$

respectively.

3.3.3 Stretched exponential distribution

The stretched exponential function which has the following probability density function (pdf)

$$f_\beta(k) = e^{-k^\beta}$$

is derived by replacing a functional power-law into an exponential function with the assumption that k lies within the range of [0, +∞], (Kohlrausch, R,1854). In this pdf ,when β=1, the result gives the usual exponential function with a stretching exponent β between 0 and 1. When the graph of the logarithm of f is plotted against k, it is characteristically stretched that gives the name of the function.

On the other hand, the case of $\beta > 1$ has little practical usefulness with the exception of $\beta = 2$ which gives the normal distribution.

More so, the stretched exponential is also known as the complementary cumulative Weibull distribution, (Berberan-Santos et.al, 2005**)**. The higher moments of this function is given as

$$\langle T^n \rangle = \int_0^\infty dt t^{n-1} e^{-(k/T_k)^\beta} = T_k^n \Gamma(\frac{n}{\beta}).$$

where Γ is the gamma function. For the exponential decay $\langle T \rangle = T_k$ is recovered in which T shows the shape parameter and β indicates the scale parameter. But in the practical purpose, most of the researchers refers to the stretched exponential to be same as the Weibull distribution, (Mohammad A. Al-Fawzan,2000). The following expression is the probability density function of a Weibull random variable.

$$f(x, \lambda, \mathrm{k}) = \begin{cases} \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{-(\frac{x}{\lambda})^k} & , \quad x > 0 \text{ and} \quad \text{when } x < 0 . \end{cases}$$

In this equation, λ >0 presents the scale parameter and k> 0 denotes the shape parameter of the distribution. Accordingly, the moment generating function of the Weibull distributed random variable $X=x$ in the logarithmic scale is computed as

$E[e^{tlogx}]= \lambda\Gamma(\frac{t}{k}+1)$,

where Γ represents the gamma function as used beforehand. Hence, the mean E(x), variance V(x), the skewness β_1 and the kurtosis β_2 of this function are listed as below, in order.

$E(x) =\mu= \lambda\Gamma(1+\frac{1}{k})$.

$V(x) =\sigma^2 = E(x^2) - (E(x))^2$

$=\lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - (\Gamma(1+\frac{1}{k})^2\right]$.

$\beta_1 = \frac{(U_3)^2}{(U_2)^3}$.

where $U_2= V(x)$ and $U_3 = U_2' - 3U_2' U_1' + 2(U_1')^3$. So,

$$(U_3)^2 = \left[\lambda^2\Gamma\left(1+\frac{2}{k}\right) - 3\lambda^2\Gamma\left(1+\frac{2}{k}\right)\lambda\Gamma(1+\frac{1}{k}) + 2(\lambda\Gamma(1+\frac{1}{k}))^3\right]^3$$

and $(U_2)^3 = \left[\lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - (\Gamma(1+\frac{1}{k})^2\right]\right]^3$

Hence,

$\beta_1 = \frac{(U_3)^2}{(U_2)^3} = \frac{\Gamma\left(1+\frac{3}{k}\right)\lambda^3 - 3\mu\sigma^2 - \mu^3}{\sigma^3}$.

Furthermore, the kurtosis $\beta_2= \frac{(U_4)}{(U_2)^2}$ while

$U_4 = U_4' - 4U_3'U_1' + 6U_2'(U_1')^2 - 3(U_1')^4$

$U_4 = \lambda^4\Gamma\left(1+\frac{4}{k}\right) - 4\lambda^3\Gamma\left(1+\frac{3}{k}\right)\lambda\Gamma\left(1+\frac{1}{k}\right) + 6\lambda^2\Gamma\left(1+\frac{2}{k}\right)\Gamma(1+\frac{1}{k})^2 - 3\Gamma(1+\frac{1}{k})^4$.

$U_2{}^2 = \left[\lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - (\Gamma(1+\frac{1}{k})^2\right]\right]^2$.

$\beta_2 = \frac{(U_4)}{(U_2)^2} = \frac{6\Gamma_1^4+12\Gamma_1^2\ \Gamma_2 - 3\Gamma_2^2 - 4\Gamma_1\Gamma_3+\Gamma_4}{[\Gamma_2 - \Gamma_1^2\ ]^2}$.

In the expressions,

$\Gamma_i = (1+\frac{i}{k})$ for i=1,..,k.

3.3.4 **Truncated Power-law**

A truncated distribution simply means a conditional distribution. The truncation emerges from restricting the domain of the probability distribution. In practical sense, it results from cases where the ability to get full detail is limited or limited to values which should be between a given range. (Dodge, Y. 2003)

The power-law is a kind of distribution that has special probability distribution. It has many ways of defining it mathematically. It is mostly used in Geoscience. In the past the empirical evidence of the power-law has not been strong.
Hereby, the probability density function or the mass function for a power-law can be written as

$$f(x) \propto \frac{1}{x^a}$$

given that $x \geq a$ and $a > 0$ with the normalization factor which depends on the nature of x (discrete or continuous). In whatever form (discrete or continuous), the normalization gives $a > 1$. Sometimes the power-law distributions are referred to as Pareto distributions (Evans et al., 2000; Johnson et al., 1994) also referred it to be Riemann zeta distributions when it is in a discrete for, (Johnson et al., 2005).

In the recent work of Clauset et al (2011), a method is suggested to find the range at which certain distributions can be power-law. But this method failed as the stimulated power-law data are not recognized. There is a noticeable and non-statistical properties of this distribution in the sense it has the divergence of moments and also the scale invariance.

The first one implies that only the first moment of the power-law distribution exists and all of the rest are infinite (Clauset. (2011). Hence, when $1 < \alpha < 2$, the mean and other moments are finite. When $2 < \alpha < 3$, the mean but other moments are infinite. This is in contrast with other pdf. On the other hand, the second property is that when the function is defined between 0 and ∞, it has no characteristic scale.

Then because of the problematic nature of the power-law when $2 < \alpha < 3$, the simple solutions are the ***truncation*** of the tail, (Aban et al., 2006; Burroughs and Tebbens, 2001; Johnson et al., 1994).

As a result, this gives the birth of the truncated power-law distribution which is defined as

$f(x) \propto \frac{1}{x^a}$ under $a \leq x \leq b$

where b is the normalizing factor. The existence of a finite upper cut-off b gives a well-behaved moment.

However, as in the biological literature, the percentage of the truncatıon of the data has not been defined so that the remaining datasets fits the power-law, we do not use the truncated power law as one of the alternative degree distribution in our analyses.

3.4 **Three moment Chi-Square and Four moment F approximation**

In the analysis, we further check for the chi-square and F-distributions assumption under the third and the four moment approximations. In the Pearson's family of distribution, there are zones with no defined distribution family, therefore if any of the result falls in this zone we can investigate it further under the third and the fourth -moment approximations.

3.4.1 **Three-Moment Chi-Square Approximation**

Given the first four moments, $\mu_1^1, \mu_2, \mu_3, \mu_4$ which are the mean, the variance, the third moment and the fourth central moment as well as the Pearson coefficients β_1 and β_2, that indicate the skewness and kurtosis, the following inequality can be satisfied

$$\left| \beta_2\text{-}(3+1.5\beta_1) \right| \leq 0.5 \qquad (23)$$

with

$\chi^2 = \frac{y+i}{j}$..

Here *y* is the random variable and *v* shows the degree of freedom while the values of *i, j* and *v* are obtained by equating the first three moments on both sides of the expression in Equation (23) via

$v = \frac{8}{\beta_1}$, $b = \sqrt{\frac{\mu_2}{2v}}$ and $a = bv - \mu_1'$.

Then, we can conclude that the distribution is a central chi-square (Moti L Tiku, 1998) with degree of freedom *v*. On a Pearson curve, the chi-square distribution is defined on the line

$\beta_2 = (3+1.5\beta_1)$, (24)

which is defined as the Type III distribution (Pearson and Tiku, 1970). The three-moment chi –square provides a good approximation given that the difference in β_2 and $3+1.5\beta_1$ does not exceed 0.5

3.4.2 **Four-moment F approximation**

Also the approximation of the four-moment F distribution approximation is defined by the first four moments, μ_1^1, μ_2, μ_3, μ_4, the procedure is similar to the three-moment chi-square approximation. By equating the first four moments on both sides of Equation (23) (Tiku and Yip, 1978),

$$F = \frac{Y+m}{n} \quad , \tag{25}$$

$$m = \frac{v_2}{v_2-2}h - \mu_1^{'} \, ,$$

$$n = \sqrt{\left\{\frac{v_1(v_2-2)^2(v_2-4)}{2v_2^2(v_1+v_2-2)}\mu_2\right\}} \quad ,$$

$$v_1 = \frac{1}{2}(v_2-2)\left(-1+\sqrt{1+\frac{32(v_2-4)/(v_2-6)^2}{\beta_1-32(v_2-4)/(v_2-6)^2}}\right),$$

$$v_2 = 2\left[3+\frac{\beta_2+3}{\beta_2-(3+1.5\beta_1)}\right].$$

For valid Equation (25), the $\beta_1$and β_2 of Y must satisfy Equation (26) as given below.

$$\beta_1 > \frac{32(v_2-4)}{(v_2-6)^2} \qquad \text{and} \qquad \beta_2 > 3+1.5\beta_1. \tag{26}$$

The inequalities in Equation (26) determine the F-region in the Pearson (β_1, β_2) plane in such a way that it is bounded by a chi-square line and the reciprocal of the chi-square line (Pearson and Tiku, 1970).

Tiku and Yip (1978), further explains that whenever $\beta_1$and β_2 point the random variable *Y* lying within the F-region, the four-moment F approximation gives an accurate approximation for the probability integral and the percentage points of *Y*.

3.5 Stimulation study

A Monte Carlo runs stimulation study is performed in order to compare result with the results of real datasets. The aim of this study is to check the out-degree of the directed networks. But due to the default of the available package for constructing the networks and analyzing the degree distribution (the packages focus only on undirected networks), we can only work on undirected networks under the stimulation study. Here we assumed the in-degree and out-degree to be the same in the calculation. The *huge* package in R is used.

The *huge* package provides a general framework for the high-dimensional undirected graph estimation, (Zhao et al, 2012.). In the package, *huge.generator* is used to generate the network under different nodes sizes namely 500, 1000, 5000, 10000 and 15000. Here the nodes are assumed to be genes. Furthermore, the package does not make provision for feed forward loop (FFL). (Barabasi and Oltvai, 2004., Wit et al, 2010.). Therefore, we assume that each gene has a single function. The dimension of the systems is also referred to as the number of genes. Moreover, in our analysis, we check three graphical structures which are assumed to be biological networks namely, scale-free, clusters and hubs networks. We used the default settings of the package to calculate clusters in the network. The cluster value is about p/20 if $p \geq 40$ and if p < default value is 1, while p is the variable number, i.e the number of genes. Hence, the limitation of this package is that, it focuses only on undirected networks. Thus the in-degree and out-degree are the same.

After generating the networks, we further used another package in R called PearsonDS. This is a package that fit the probability distribution and also Pearson family via moments and the maximum likelihood of the data. This package is used to identify suitable Pearson family. (Becker et al, 2014.)

The codes and further explanation are given in the Appendix. B.2

3.6 Measures of goodness of fit

In other to test for the alternative degree distributions of the departing connectivity of the real dataset, we perform goodness of fit test**.**

The Goodness of fit test indicates whether or not it is reasonable to assume that a random sample comes from a specific distribution. They are a form of hypothesis testing where the null and alternative hypotheses are. This can be described as below.

H_0: The data comes from the specified distribution

H_1: The null hypothesis is not true.

The chi-square test is the oldest goodness of fit test dating back to Karl Pearson (1900). In this work, we apply chi-square test. Hereby, we reject the null hypothesis when the specified significant level α value is greater than the calculated p value.

CHAPTER 4

RESULTS OF ANALYSES

All the datasets used are from directed networks. The data are taken from the Biod2rdf project which is one of the largest databases in bioinformatics. Bio2RDF is an open-source project that uses Semantic Web technologies to build and provide the largest network of Linked Data for the Life Sciences. Hereby, this project aims to transforms silos of life science data into a globally distributed network of linked data for the discovery of the biological knowledge (Michel DumontieR, 2014).

Bio2RDF defines a set of simple conventions to create RDF(S) compatible Linked Data from a diverse set of heterogeneously formatted sources which are obtained from providers of the multiple data, (Michel Dumontier, 2014).

Hence, in order to investigate the degree distribution of the biological networks, we control both real systems and stimulated systems under different topologies. For the real systems, we check 10 realistically complex biological pathways which are described via directed networks. Among them, the first six datasets are the gene-gene interaction networks which are extracted from the gene-disease relations and are taken from the OMIM (Online Mendelian Inheritance in Man**)** database. OMIM is a comprehensive, authoritative and timely compendium for the human genes and genetic phenotypes (McKusick-Nathans, et al, 2014) and the selected networks are constructed by connecting the genes having relation with the same disease. On the other hand, the last four real datasets are taken from the HIV-1 human Protein Interaction database (Fu et al, 2009). This database contains the HIV human protein reactions that are created to catalog all interactions between HIV-1 genes.

Finally, Cytoscape is used to visualize and to analyze the code which is presented in Appendix B.1.

4.1 Description of the real datasets and their network graphs

In this study, we apply 6 different dimensional real datasets from human diseases interactions. Data 1 to 6 consist of 724, 223, 1008, 987 643 and 188 genes interactions, respectively. Data 7-10 are the HIV diseases interactions with different number of genes. Data 7 are composed of 1469 genes. Data 8 and 9 have 1152 and 722 genes, in order. Finally, Data 10 consist of 306 genes. The graphical representation of each network is shown in the following figures.

Data 1

The Paget disease is a metabolic bone disease known by the abnormalities of the focal which increases the bone turnover and thereby, affecting one or more sites throughout the skeleton, majorly, the axial skeleton. The network of this disease is described via 724 genes whose graphical representation is shown in Figure 6.

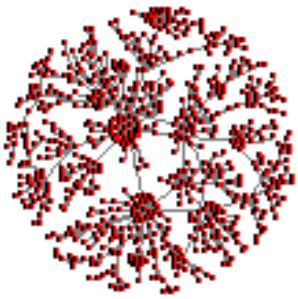


Figure 7: Graphical representation of the Paget disease taken from the OMIM database.

Data 2

Menkes disease is an X-linked recessive malfunctioning caused by the copper deficiency. The associated network is presented via 223 genes whose graph is given in Figure 7.

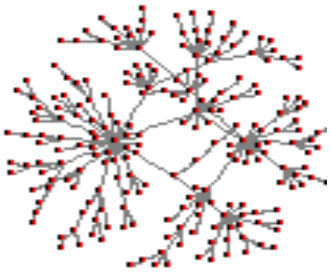


Figure 8: Graphical representation of the Menkes disease taken from the OMIM database.

Data 3

The Inflammatory bowel disease is caused by a severe degenerating intestinal infection and it is displayed by 1008 genes as visualized in Figure 8.

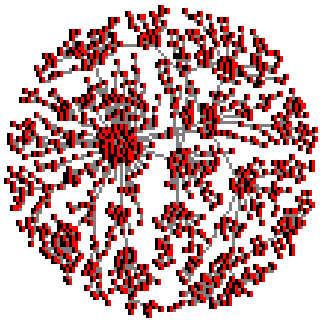


Figure 9: Graphical representation of the inflammatory bowel disease taken from the OMIM database.

Data 4

The Glycogen storage disease is related with Glycogen the deficiency and is described via 987 genes in the database as seen in Figure 9.

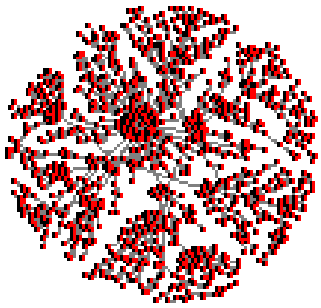


Figure 10: Graphical representation of glycogen storage disease taken from the OMIM database.

Data 5

The data belong to a particular muscle disease, called the hereditary rippling muscle disease. This illness is an autosomal dominant disorder characterized by mechanically triggered contractions of skeletal muscle. The network of this disease is described via 643 genes in Figure 10. Similar to previous datasets, these data are also taken from the OMIM database.

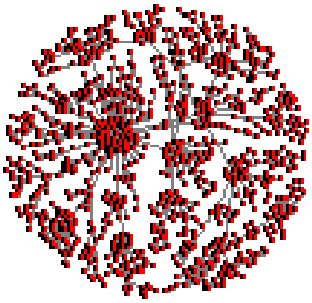


Figure 11: Graphical representation of Muscle disease taken from the OMIM database.

Data 6

The dataset 6 is from the Lafora type of the progressive myoclonic epilepsy. This Lafora disease is an autosomal recessive disorder characterized by the insidious onset of progressive neurodegeneration between 8 and 18 years of age. The network of this disease is presented via 188 gene interactions in the OMIM database.

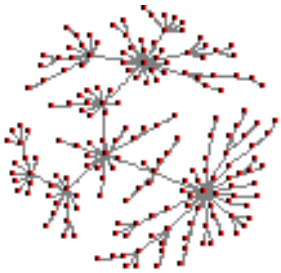


Figure 12: Graphical representation of Lafora disease taken from the OMIM database.

Data 7-10

The data from 7 to 10 belong to the human immunodeficiency virus or shortly HIV disease. HIV, as simply shown in Figure 12, is a virus that attacks the human immune system. The immune system protects human from germs that cause infections. But, if HIV is in the system, overtime, it reduces the immune cells (CD4). People get infected with HIV through bodily fluids such as blood, semen, breast milk and vaginal fluids. People do not get HIV through insect bites, casual contact such as hugging, shaking hands, or living with someone who has HIV.

Hereby, HIV can damage the immune system to such a degree that infections may begin to occur as a result of a weakened immune system. Eventually, one may acquire various illnesses due to the damage done by the virus. The networks in Figures 13 -16 show the related illnesses and diseases acquired as a result of HIV.

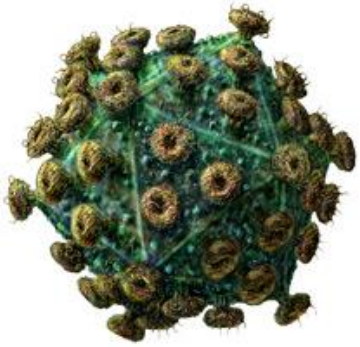


Figure 13: Three dimensional illustration of HIV (Country Awareness Network Victoria Inc; www.can.org.au/Pages/About.aspx)

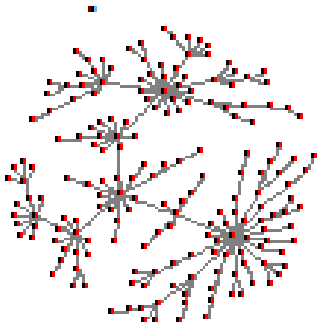


Figure 14: Graphical representation of the HIV disease's interactions with 306 genes.

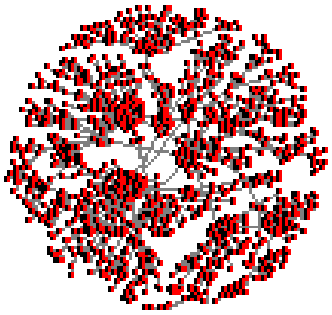


Figure 15: Graphical representation of the HIV disease's interactions with 1152 genes.

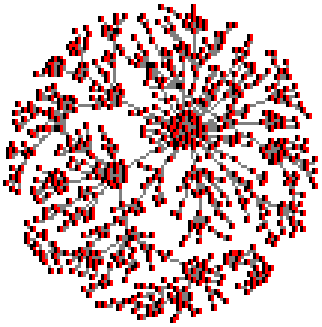


Figure 16: Graphical representation of the HIV disease's interaction with 722 genes.

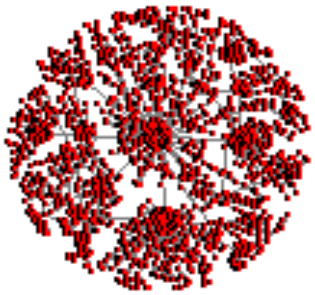


Figure 17: Graphical representation of the HIV disease's interactions with 1469 genes.

4.2 Results of the real datasets analyses

The results of analyses give a four- parameter distribution in the sense that a and b in this distribution present the shape parameters for the left and right sides, respectively. On the other hand, for the remaining two parameters, it uses the location and the scale parameters. Here, the former is the minimum and is shown via *l* and the latter indicates the difference in the maximum and the minimum of the range and is displayed by *s*.

Thereby, in the following tables, the estimation of these parameters, their means, variances, skewness and kurtosis values are reported and finally, the associated Pearson families are declared regarding their first four moments.

Table 1: Summary of the network analysis for the Paget disease.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson Family	Parameters
Paget disease	724	16.58	300.37	2.50	4.97	I	a: 1.07 b: 39.54 l: -0.83 s: 630.03

Table 2: Summary of the network analysis for the Menkes disease.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
Menkes disease	223	11.62	134.18	5.60	7.58	I	a: 0.60 b: 4.21 l: 1.03 s: 84.26

Table 3: Summary of the network analysis for the Inflammatory bowel disease.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
Inflam-matory bowel disease	1008	2.05	11.33	1.03	4.78	VI	a: 5.72 b: 44.24 l: 0.51 s: 19.37

Table 4: Summary of the network analysis for the Glycogen storage disease.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson Family	Parameters
Glycogen storage disease	987	21.02	464.33	12.26	23.79	VI	a: 1.14 b: 12.80 l: 0.02 s: 217.87

Table 5: Summary of the network analysis for the Muscle disease.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
Muscle disease	643	104.82	996.51	1.57	4.84	I	a: 0.81 b: 7.08 l: 4.23 s: 981.16

Table 6: Summary of the network analysis for the Lafora disease.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
Lafora disease	188	48.47	1867.86	1.34	4.31	I	a: 0.57 b: 2.71 l: 7.27 s: 235.41

Table 7: Summary of the network analysis for the HIV disease with 1469 gene interactions.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
HIV interaction	1469	83.28	590.19	0.90	2.91	I	a: 0.66 b: 1.87 l: -2.67 s:328.56

Table 8: Summary of HIV 1152 disease interaction analysis

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
HIV interaction	1152	11.51	140.80	3.84	7.37	I	a: 0.78 b: 11.86 l: 0.28 s: 182.26

Table 9: Summary of the network analysis for the HIV disease with 722 gene interactions.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
HIV interaction	722	7.56	67.83	7.6	13.04	I	a: 0.88 b: 13.68 l: 0.70 s: 136.46

Table 10: Summary of the network analysis for the HIV disease with 306 gene interactions.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Parameters
HIV interaction	306	3.53	709.38	2.95	5.9	I	a: 0.001 b: 1.09 l: 1.96 s: 947.72

From the results in Tables 1-10, it is shown that the majority (8 out of 10) of the data falls into the Pearson Type I distribution which presents the beta family. On the other side, two datasets follow the Pearson Type VI distribution.

The beta family belongs to a family of the continuous probability distributions parameterized by two positive shape parameters (*a* and *b*), location (*l*) and scale (*s)*. In beta family, the location parameter controls the position of the distribution on the x-axis and the scale parameter controls the spread of the distribution on the x-axis. One of the most common applications of the beta distribution is to model the uncertainty about the probability of successes in an experiment.

Accordingly, from the tabulated results of analyses, it is observed that the second shape parameter (*b*) is greater than the first shape parameter (*a*) in all analyses. Thus, the graphs, as represented in Figures 12-19, are found to be right-skewed. Also the values of the variance values affect the scale values.

Figure 18: Graph of the beta distribution for the network analysis of the Paget disease.

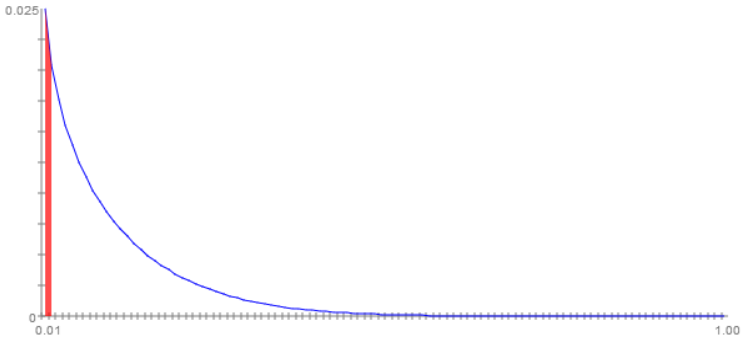


Figure 19: Graph of the beta distribution for the network analysis of the Menkes disease.

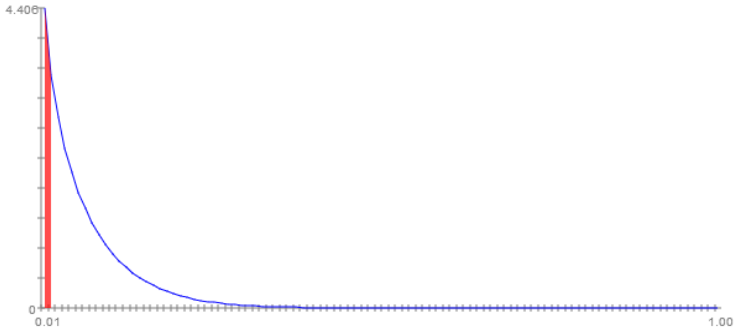


Figure 20: Graph of the beta distribution for the network analysis of the Muscle disease.

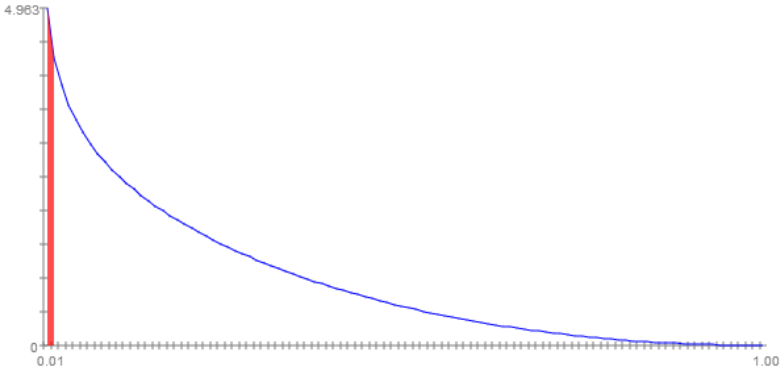


Figure 21: Graph of the beta distribution for the network analysis of the Lafora disease.

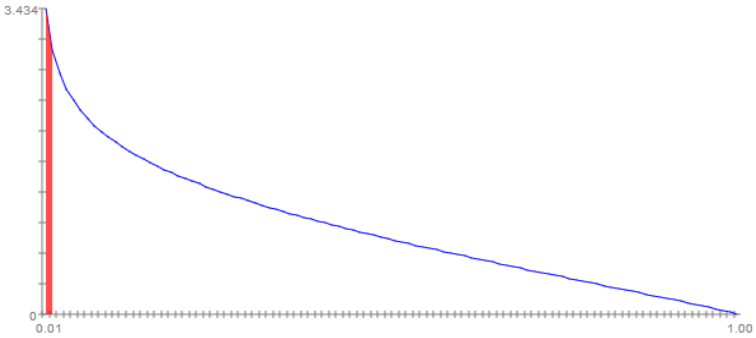


Figure 22: Graph of the beta distribution for the network analysis of the HIV disease with 306 gene interactions.

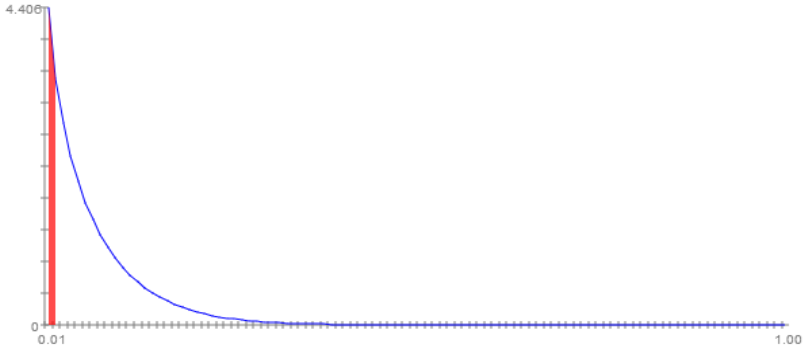


Figure 23: Graph of the beta distribution for the network analysis of the HIV disease with 1152 gene interactions.

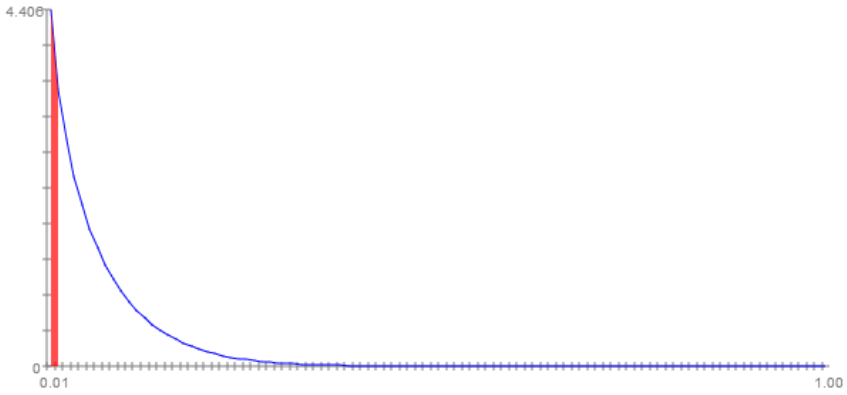


Figure 24: Graph of the beta distribution for the network analysis of the HIV disease with 722 gene interactions.

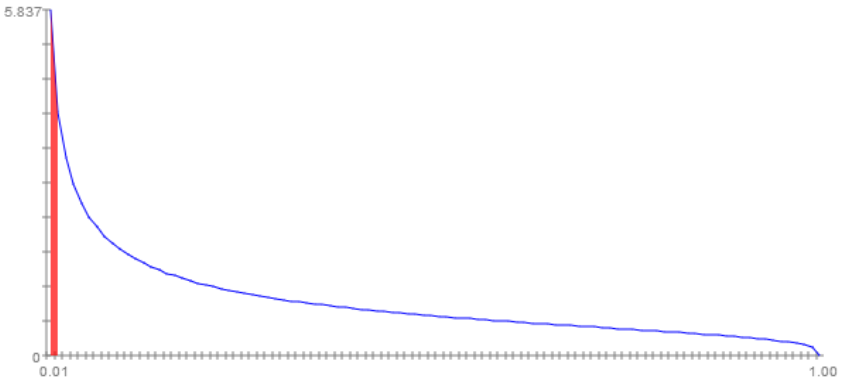


Figure 25: Graph of the beta distribution for the network analysis of the HIV disease with 1469 gene interactions.

On the other hand, as stated previously, the network of the Inflammatory bowel disease and the network of the Glycogen storage disease fall in the Pearson Type VI family of distribution. This type of distribution indicates an area defined as the region between the gamma and the Pearson Type V family. The major cases in this family can be the beta distribution of the second kind and the Fisher F distribution (Lahcene, 2013). From our results based on four parameters, we detect these two networks indicate the beta distribution of the second kind.

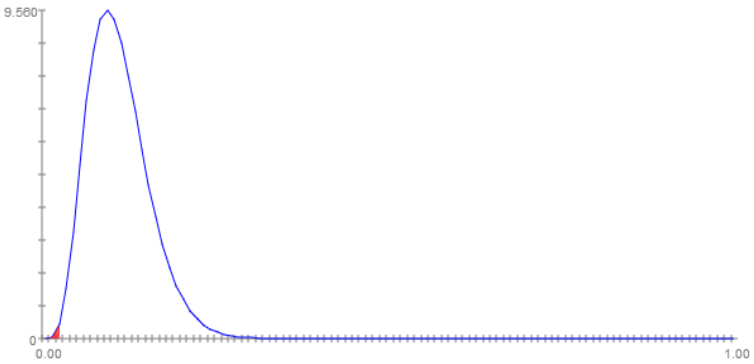


Figure 26: Graph of the beta distribution for the network analysis of the inflammatory bowel disease.

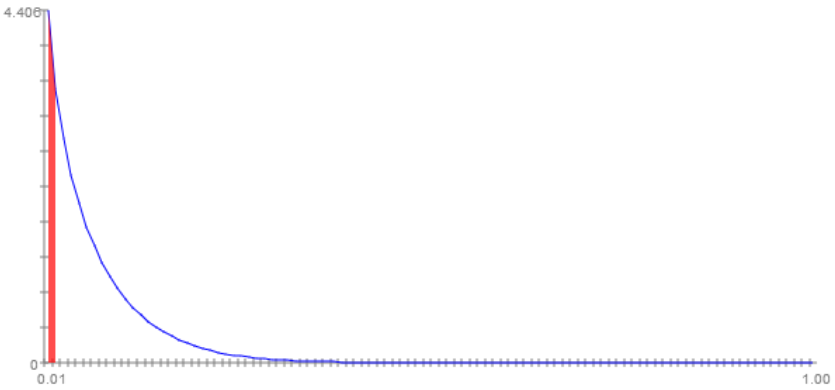


Figure 27: Graph of the beta distribution for the network analysis of the Glycogen storage disease.

4.3 Results of the stimulated data analysis

In order to verify and compare the results from the real datasets, we conduct different stimulation studies by using Monte Carlo runs. In each run, we compute 1000 iterations (which is the count). In the stimulation, genes are assumed to be nodes and also, the dimension of the systems is the same as total number of genes, which are chosen as 500, 1000, 5000, 10000 and 15000, respectively. Once the networks are generated, the detection for the suitable Pearson family is found.

Below is the summary of the results and also attached in Appendix B.2 is the code used in the stimulation studies.

Note: in the table number of genes indicate number of nodes and count refers to the iteration times in Monte carlo runs

Table 11: Summary of the 1000 Monte Carlo iterations.

Name of network	No of genes	Mean	Variance	Skewness	Kurtosis	Pearson family	Count
Scale-	500	2.8000	1.7333	6.6594	57.9493	I	1000
free	1000	2.0123	5.6718	10.8285	167.5472	I	1000
	5000	2.1245	12.2815	9.9095	121.2844	I	994
						VI	6
	10000	2.9996	13.9644	9.1401	107.3029	I	1000
	15000	2.2344	12.4761	9.9998	124.6722	I	1000
Cluster	500	6.7000	4.6881	0.3545	0.1733	No family	1000
	1000	6.5200	3.2861	0.1537	-0.2046	No Family	1000
	5000	7.000	4.5859	1.8343	1.2371	I	12
						No family	988
	10000	6.2400	3.0433	0.1127	-0.6946	No family	1000
	15000	6.5600	3.4211	0.1631	0.0035	No family	1000
Hub	500	2.86	15.5848	4.0986	14.8726	No family	1000
	1000	2.90	15.5455	4.1083	14.9325	No Family	1000
	5000	2.88	12.3935	4.1171	14.9805	No family	1000
	10000	2.90	15.6509	4.0267	14.4559	No family	1000
	15000	2.23	15.4443	4.0023	14.6785	No family	1000

From the results in Table 11, it is shown that for the scale-free networks, almost all the observed networks belong to the Pearson's Type I family while just 6 out of 1000 runs under the 5000 dimensional networks belong to the Pearson Type VI family.

Furthermore, from the outputs of the cluster and hub networks, the results of stimulations present that there is no Pearson's family for the observed gene numbers, apart from the clusters network under 5000 dimensional networks. Under this condition, we detect 12 out of the 1000 systems belonging to the Pearson Type I family. In the study of Bachioua Lahcene (2013), it is presented that there is no Pearson family if the kurtosis values are very small. Moreover, Pearson does not

define any family for the values of the kurtosis less than 1. Accordingly, in our stimulation studies, all the generated kurtosis values for the cluster type for networks are computed as less than one.

Finally, for the hub networks, the kurtosis falls in the region of the limit of the Pearson's distribution with no define family.

4.4 Three-moment chi-square and four-moment F approximations' results

In order to verify and make better conclusion as regard tothe result of our analyses, we further investigate the networks with no Pearson family under the three-moment chi-square and four-moment F approximations. Table 12 and 13 give the summary of analysis. β_1 and β_2 show the skewness and kurtosis, respectively, as used previously, while $\boldsymbol{\partial_2}$ in Table 13 presents the second degree of the freedom in the F-distribution. The explicit form of this term is given in Equation (22) and (23).

In the analyses, again, we consider three types of networks, which are scale-free, hubs and clusters, i.e. modular, due to the fact that the biological networks can satisfy all these cases, as used beforehand. . Then, we detect the practical application of the approximations under distinct dimensional systems. For the calculations, we use the systems with 500, 1000 and 5000 genes, respectively, as still applied in previous analyses and then, check the validity of the inequalities for the three and four - moment approximations.

Table 12: Results of the three-moment chi-square approximations which are detected by the inequality given in the third column

Name of Network	**No of genes**	$\lvert \boldsymbol{\beta_2}$-(3+1.5$\boldsymbol{\beta_1}$) $\rvert \leq \mathbf{0.5}$
Scale-free	500	44.960281 > 0.5
	1000	148.30440 > 0.5
	5000	103.42015 > 0.5
Cluster	500	3.358483 > 0.5
	1000	3.435096 > 0.5
	5000	3.378627 > 0.5
Hub	500	5.724734 > 0.5
	1000	5.769201 > 0.5
	5000	5.804839 > 0.5

Table 13: Results of the four-moment F approximation which is detected by the inequalities given in the third and the fourth column

Name of network	**No of genes**	$\boldsymbol{\beta_1} > \frac{\mathbf{32}(\boldsymbol{v_2}-\mathbf{4})}{(\boldsymbol{v_2}-\mathbf{6})^2}$	$\boldsymbol{\beta_2}$ > 3+1.5$\boldsymbol{\beta_1}$
Scale –free	500	6.659373 < 8.711253	**57.94934 >12.98906**
	1000	**10.82853 > 8.299961**	**167.5472 > 19.24279**
	5000	**9.90950 > 8.403485**	**121.2844 >17.86425**
Cluster	500	0.354513 < 0.988266	0.173287 < 3.531770
	1000	0.153672 < 4.498817	-0.204588 < 3.230508
	5000	0.094327 < 4.360671	-0.237136 < 3.141491
Hub	500	**4.098551 > 1.759712**	**14.87256 > 9.147827**
	1000	4.108853 < 6.803525	**14.93248 > 9.163280**
	5000	4.117101 < 6.833076	**14.98049 > 9.175652**

From Table 12 and 13, the results indicate that the scale-free networks satisfy the F distribution inequalities and hub networks under 500 genes. This finding shows that

the distribution is most likely to be in the F-region in the Pearson table and the area is bounded by the chi –square line, i.e.β_2 = 3+1.5β_1, and the reciprocal of the chi-square line (Pearson and Tiku 1970).

4.5 **Test for alternative distributions**

As stated earlier in the literature that the degree of the departing connectivity in biological networks follows the generalized pareto, the geometric and the Weibull distribution (alternative to the stretched exponential). We test the original 10 datasets under each of these distributions. A chi-square goodness of fit test is used and below is the summary. Accordingly the associate hypotheses are constructed as below.

Hypothesis:

H_0: The data come from the (geometric, pareto, or Weibull distribution).

H_1: The null hypothesis is not true.

Table 14: Summary of the chi-square test for the three distributions (Weibull, pareto and geometric) at 5% level of significance.

Data	P value for Chi-square Test	Conclusion
1	4.444e-12	Reject all
2	0.02034	Reject all
3	0.9933	Do not reject the null hypothesis (Pareto)
4	1.384e-10	Reject all
5	2.32e-16	Reject all
6	1.2e-16	Reject all
7	2.64e-16	Reject all
8	0.005662	Reject all
9	4.765e-05	Reject all
10	1.77e-16	Reject all

The results of chi-square goodness of fit test shows that none of the original datasets follows any of the tested alternative distribution, except in the case of the inflammatory disease network, where the pareto distribution is significant and we accept the null hypothesis that the data follow a pareto distribution.

We further draw the Q-Q plots of the empirical and theoretical distributions, which are presented below

.

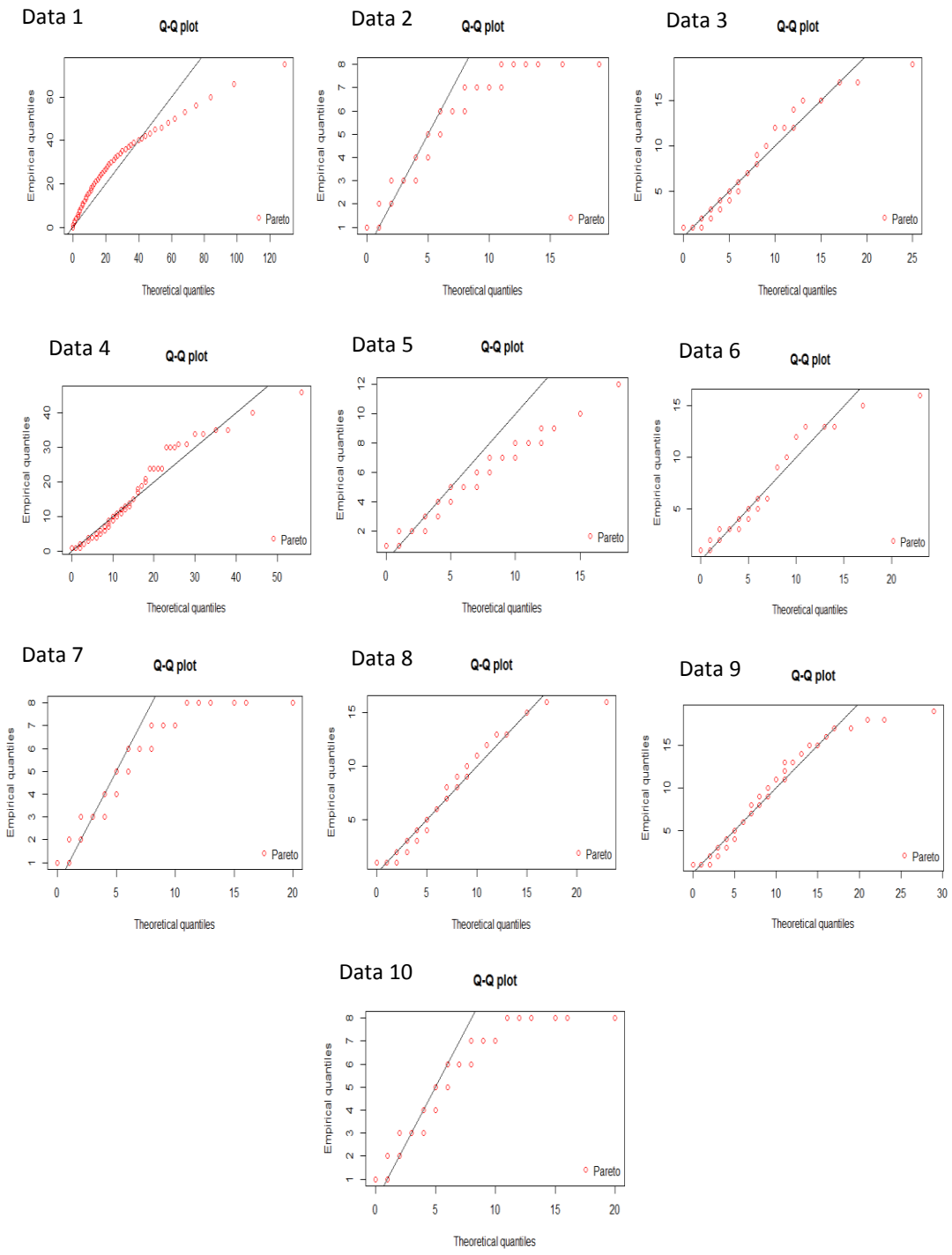


Figure 28: The Q-Q plot of the pareto distribution against the theoretical distribution.

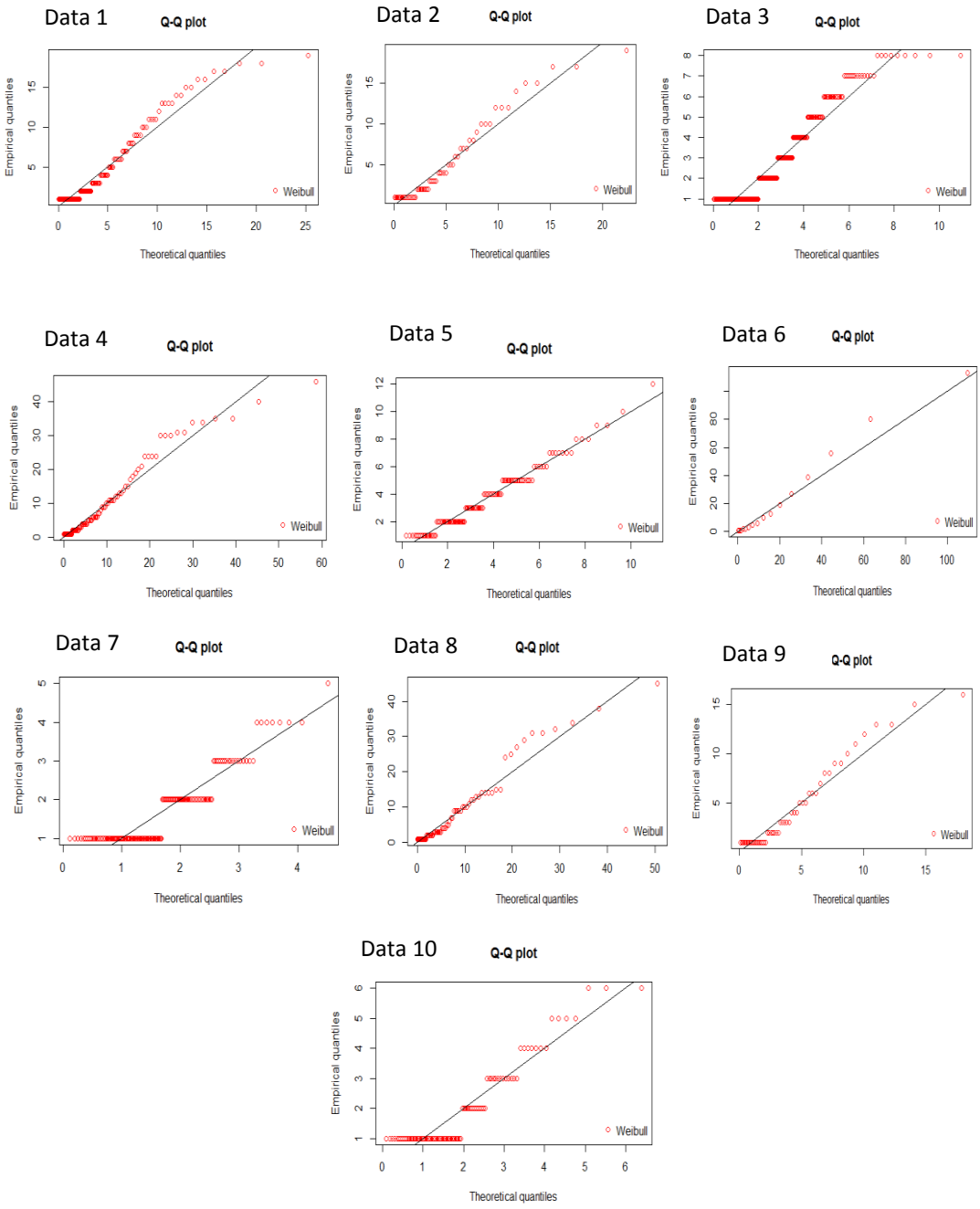


Figure 29: The Q-Q plot of the weibull distribution against the theoretical distribution

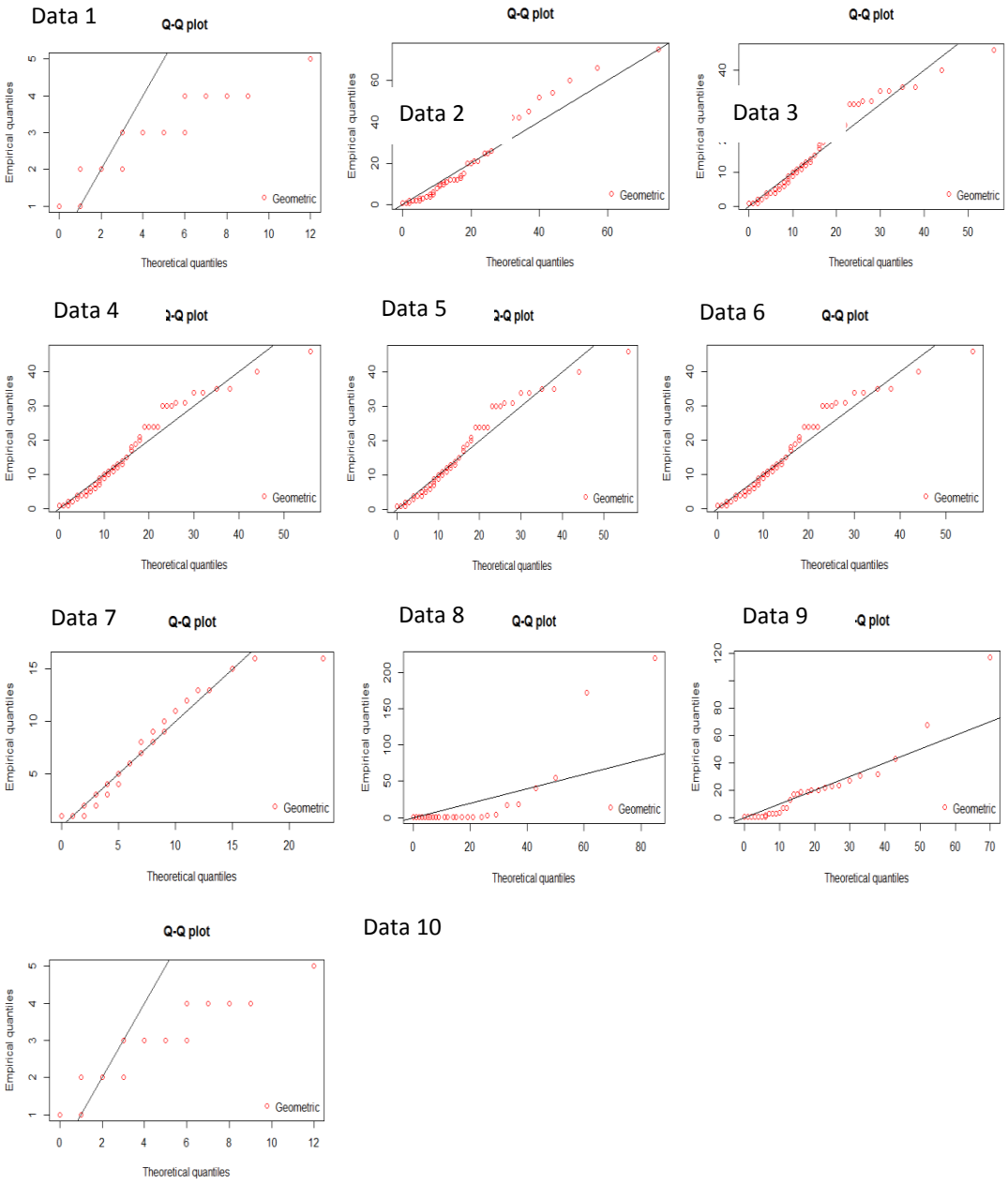


Figure 30: The Q-Q plot of the geometric distribution against the theoretical distribution.

Furthermore we draw the histogram and the theoretical graphs of each of the datasets.

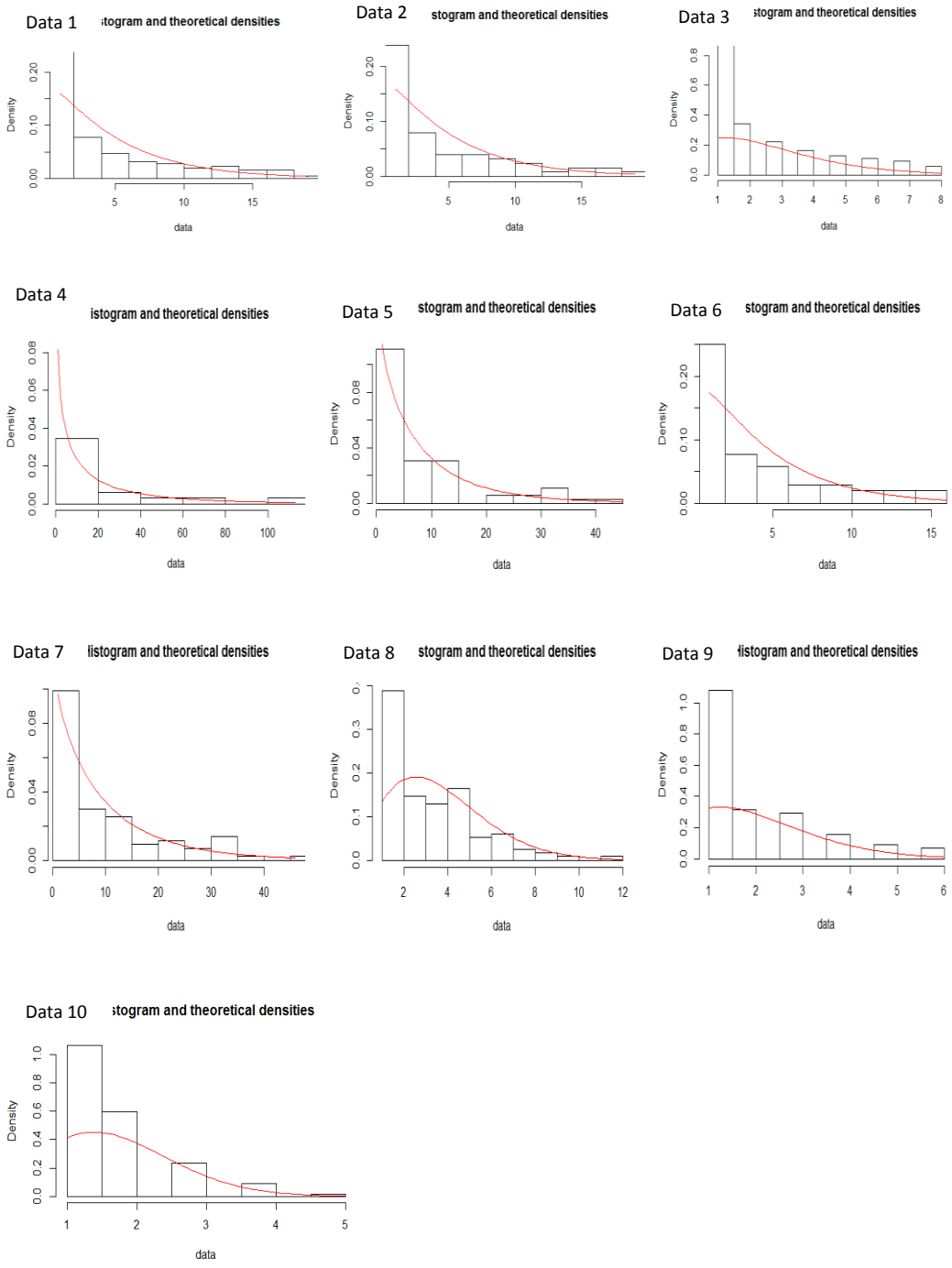
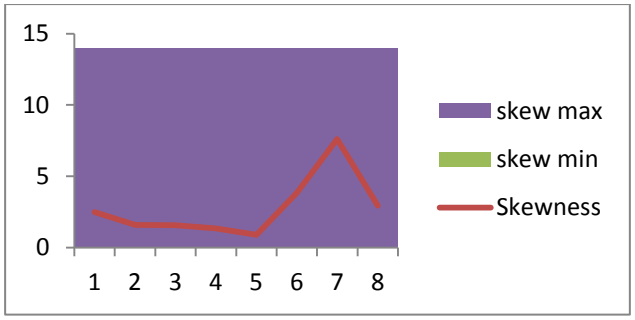


Figure 31: Histrogram and density plot of the original data sets**.**

The Q-Q plots graphs shows that none of the dataset is likely to come from the observed datasets. Also, the histograms shows that the all the datasets are rightly

skewed,an obervation that is the same as the graph of the beta family is observed in figure 18-27.

Finally, under the data analyses we observe the graphs of the band of the skewness and the kurtosis under the Pearson Type I and VI. We observe the datasets under the limit defined for the skewness and the kurtosis in the Pearson family. On the other hand, the relation of the mean and variance under the Pearson family cannot be established just because the relations between skewness-kurtosis and the mean-variance pairs are non-linear. Thus, we cannot define a unique inequality for the borders of the mean and variance separately.



Note: Here as the minimum value of the y axis is zero, the line cannot be visible

Figure 32: Band graph of the skewness under Pearson Type I family.

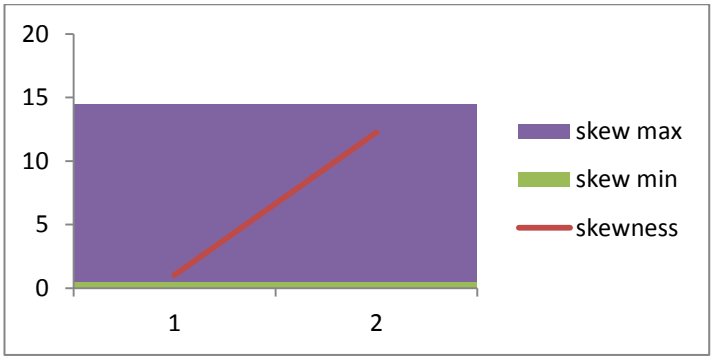


Figure 33: Band graph of the skewness under Pearson Type VI family.

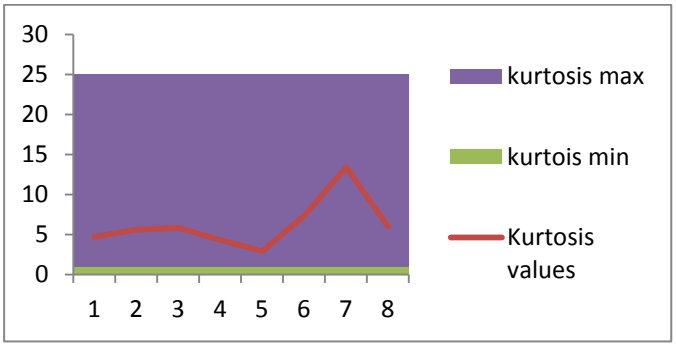


Figure 34: Band graph of the kurtosis under Pearson Type I family.

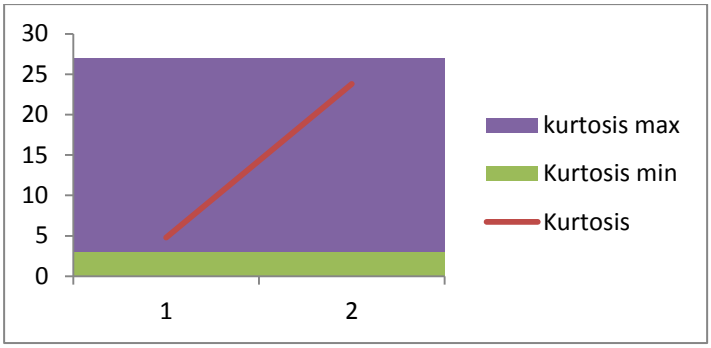
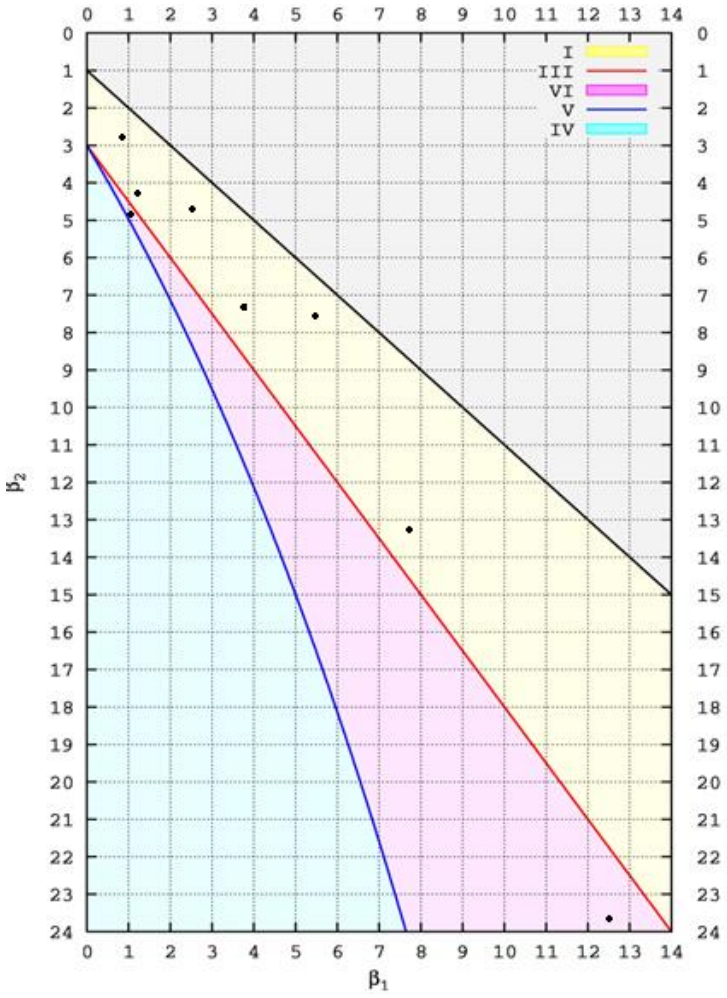


Figure 35: Band graph of kurtosis under Pearson Type VI family.



Data	Skewness	Kurtosis
1	2.5	4.7
2	5.6	7.58
3	1.03	4.78
4	12.26	23.79
5	1.57	4.84
6	1.34	4.31
7	0.9	2.91
8	3.84	7.37
9	7.60	13.4
10	2.95	5.98

Figure 36: Location of each of the dataset on the Pearson plane.

Figures 32-36 show that the results of β_1 and β_2 fit well in the defined Pearson boundary for Type I and Type VI family. Also, figure 30 shows the location of each of the dataset on the Pearson type of family plane.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Discussion

The goal of this study is to establish the degree distribution of the departing connectivity in the directed biological networks. In this study, we have investigated the out-degree of the biological networks within the Pearson curves by studying 10 real datasets. Furthermore, we have checked the real datasets for the alternative degree distribution of departing connectivity has suggested in literature. Also, we have stimulated data under various scenarios to be able to compare and arrive at cogent conclusions.

For the stimulation study, we have arrived at the following discussion:

- The Monte Carlo runs are based on the undirected network and not directed network as we have aimed to investigate. This is the result of the default of the huge.generator package used in the calculation.
- Also, one gene is assumed to be one node without multiple functions, therefore, the feed forward loop (FFL) function is ignored.
- The number of clusters in the network is calculated based on the default settings of the program. Also in the literature there is no evidence of minimum or maximum number of clusters in biological networks.

For future studies and more clarification, the above observation is subjected to improvement.

5.2 Conclusion

In this thesis, shortly, we have aimed to investigate the degree distribution of the biological networks. In order to find a suitable distribution in the Pearson family by using the moment's estimators, we have done analyses for 10 real datasets from different biological networks and also performed stimulation studies for the distinct sorts and dimensional networks.

With more details, in the calculations, the networks have been represented through edges in a graph, where the nodes have represented genes or diseases. The collections of the nodes interaction under directed networks have been indicated as the networks studied. The structure of the chosen real datasets, that is their topologies, have been assumed as scale-free while in the stimulated datasets, hubs and cluster networks are also examined along with the scale-free types of networks.

In the calculations, the real datasets have been checked for their robustness and then conformity before settled for. Because in the networks, there are various nodes with less number of connections and just few nodes with numerous connections. By this way, we can assume that the networks satisfy the small world properties, that is, there is a short path (d) between nodes, and the centrality as well as lethality properties via the presence of hubs.

Hereby, the goal of this study is to establish the degree distribution of the departing connectivity for the directed biological networks by observing the data under the Pearson system. The results of real datasets have showed that the degree distributions fall under the Pearson Type I family and only few of them are under the Pearson Type VI. The Pearson Type I refer to the beta family and the Type VI family indicates the region between Gamma and Type V family. Thereby, we have concluded that the major cases in this family are observed as the beta distribution of the second kind and the Fisher F distribution.

Biologically the musculosketal disease (paget and ripping muscle diseases), nervous diseases (menkes and lafora disease) and HIV virus disease all belong to Pearson

type I family. While digestive (inflammatory disease) and congenital disease (glycogen storage diseases) both belong to type VI under Pearson system.

Also from the Monte Carlo runs, we have showed that the scale-free networks belong to the Pearson Type I family when very few ones belonging to the Pearson Type VI family. But we have observed other types of biological networks under the stimulation studies as well. For instance, 12 out of 1000 runs in the cluster networks and with 5000 nodes, we have found also the Pearson Type I family while the rest of the results shows no particular family. On the other hand, the hubs networks are have been observed under the stimulation studies. From the outputs it has been seen that the entire results give no specific Pearson family.

Later, we have examined the three types of networks under the three and four-moment approximations. It is interesting to note that the scale-free networks satisfy the four moment inequalities. The findings also show that for large hub networks, Monte Carlo runs satisfy the inequalities of the four-moment approximations. Here we have detected that the distribution is most likely to be in the F-region in the Pearson table and the area has been bounded by the chi –square line ($\beta_2 = 3+1.5\beta_1$) and the reciprocal of the chi-square line (Pearson and Tiku 1970).

Furthermore, we have observed that the results of the four-moment F approximation and the result of real datasets are similar. Both outputs for the scale-free and hub networks show that the degree distribution can also lie in the F-region of the Pearson curve. On the other side, for the non-Pearson family, the kurtosis values are less than 1. But the Pearson system does not define any family for values of the kurtosis less than one and for those with values above 1, the kurtosis falls in the region of "Limit of Pearson´s distribution with no define family" as stated in the study of Lahcene (2013).

Finally, we have checked our datasets under three alternative distributions as suggested in the literature. Therefore, the pareto, geometric, and Weibull (alternative to stretched exponential) distributions. The result of the chi-square goodness of fit tests have revealed that none of these dataset can be assumed to come from any of the alternative distribution except in the case of inflammatory disease(Data 3). That

result, have shown that the out-degree is likely to follow the pareto distribution. The Q-Q plots have also shown the spread of the dataset on the empirical and the theoretical distribution. We can see that all the datasets are not well fitted into the empirical distributions. We have accepted the result of the chi-square to be superior over the Q-Q plot and conclude that the out-degree of the inflammatory network is likely to follow a pareto distribution, the distribution suggested by literature for the departing connectivity.

Moreover, in this work, we have found that the studied datasets share some important properties as listed below.

1. Some nodes have large number of connections to other nodes, whereas, most nodes have few. The hubs have hundred (or thousand in bigger network) of links. With this, we can assume that the networks have no scale resulting in the validity of the scale-free feature. Accordingly, these networks are robust to random attacks. However, they can be easily destroyed with coordinated attacks as stated in the literature about the biological networks.
2. Fitting a plausible distribution for the degree distributions of the biological networks can be very helpful in order to estimate the structure of the networks by different approaches such as the estimation of the interactions between genes via the Bayesian approach. In the Bayesian framework, the calculation of the model parameters can be applicable if we define a suitable prior distribution. Hereby, the distribution family which we detect from different network analyses can be a successful choice for the prior density, rather than using a non-informative prior for the computation.

On the other hand, as the progress can be feasible only if analytical and numerical works are combined with proper empirical studies, we suggest a further study on the topology of biological networks in order to reveal more unexpected window for future studies and cogent conclusions. Accordingly, for the extension of these studies we consider to detect other types of biological networks such as metabolic networks, gene regulatory and cell signaling pathways by focusing on their topological features, in particular, their degree distributions. By this way, we can fundamentally

find a concluding distribution or alternating distribution of the scale-free networks, rather than the assumed power-law distribution. F distribution is also suggested as an alternating distribution for large biological networks. Finally, as shortly described above, we propose to use these distributions in the estimation of the networks via the Bayesian approach where the prior distribution for the departing connectivity becomes essential for the further calculations.

REFERENCES

- Al-Fawzan, M. A. (2000). *Methods for Estimating the Parameters of the Weibull Distribution.* King Abdulaziz City for Science and Technology.

- Almaas, E., Vázquez, A. and Barabási, A.L. (2007). *Scale-free networks in biology.* Center for Network Research and Department of Physics, university of Notre Dame, Notre Dame, IN 46556,USA .

- Andree, A., Kanto, V.A. and Malo, P. (2005). *Simple approach for distribution selection in the Pearson system. Helsinki school of Economics. Working papers W-388)*

- Azrulhisham, E.A., Zakaria, K.P., Samizee, A. and M.B.M. (2008) *Pearson System Distribution Approximation in Wind Energy Potential Analysis.* Juhari Malaysia France Institute, Universiti Kuala Lumpur, Malaysia.

- Junker, B. H and Schreiber, N. (2008). *Analysis of Biologıcal Networks*. .Wiley & Sons.

- Barabasi, A. L., Jeong, J., Neda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002). *Evolution of the social network of scientific collaborations.* Physica A, 312, PP. 590-614

- Barabasi, A.L. and Bonabeau, E. (2003). *Scale-free networks.* Scientific American.

- Barabasi, A.L. and Oltvai, Z.N. (2004). *Network Biology:Understanding the Cells functional organization.* Nature Review, Vol 5. Pages 101 -113

- Clauset, A. (2011). *Inference, Models and Simulation for Complex Systems*. Lecture 2, CSCI 70001003.

• Davis, C. S. (1993). *A new approximation to the distribution of Pearson´s square*. University of Iowa.

• Ravasz, E., Somera, A.L., Ongru, D.A. and Oltvai, Z.N. (2002). *Hierarchical organization of Modularity in Metabolic Networks.* Vol. 297 no. 5586 pp. 1551-1555

• Guelzim, N., Bottani, S., Bourgine, P and Kepes, F. (2002). *Topological and causal structure of the yeast transcriptional regulatory network.* Nature Genetics. Vol. 31, page 60-63

• Han, J. D., Bertin., N., Hao, T., Goldberg, D. S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P and Vida, M. (2004). *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Letters to Nature, doi:10.1038/nature02555.

• Han, J.D., Duppy, D., Bertin, N., Cusick, M.E and Vidal, M.(2005). *Effect of sampling on topology predictions of protein-protein interaction networks.* Nature BioTechnology. Vol 23, pages 839-844.

• Stefano, B., Vito, L. and Yamir M. (2009). *Handbook on Biological Networks*. World Scientific Lecture notes in complex System- Vol 10, pg 21-22.

• Hanrahan, G. (2001). *Artificial neural networks ın biological and environmental analysıs*. CRC Press.

• Hidalgo, C.A. and Barabasi, A.L. (2008). *Scale-free networks*. Scholarpedia, vol 3(1)

• Ravasz, E., and Barabasi, A.L. (2003). *Hierarchical organization in complex networks* (Phys. Rev. E 67, 026112 2003)

• Khanin, R and Wit, E. (2006). *How Scale-free are Biological networks.* Journal of Computational Biology. Vol 13, pg 810-818.

• Lee, T. I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger,

J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young R.A.(2002). *Transcriptional Regulatory Networks in Saccharomyce cerevisiae.* Science AAAAS. Vol. 298 no. 5594 pp. 799-804.

- Matlis, J. (2002). *Scale-Free Networks.* Computerworld

- Mossa, S., Barthelme, M., Stanley, H. E. and Amaral, L. A. *Truncation of Power-law Behavior in "Scale-Free.* (2002). Physical Review Letters. DOI: 10.1103/PhysRevLett.88.138701.

- Pan, R. K. and Sınha, S.(2010). *Modular networks with hierarchical organization "The dynamical implications of complex structure*. The Institute of Mathematical Sciences, C.I.T. Campus, Chennai 600 113, India.

- Petersen, J. L. *Estimating the Parameters of a Pareto Distribution; Introducing a Quantile Regression Method.*

- Ron, K. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence vol **2** (12): 1137–1143

- Schwikowki, B., Uetz, P., and Fields, S. (2000). *A network of protein-protein interactions in yeast.* Nature Biotechnology Vol 18(12): 1257-1261

- Sheridan, P., Kamimura, T. and Shimodaira, H. (2010). *A Scale-Free Structure Prior for Graphical Models with Applications in Functional Genomics.* Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, Japan.

- Thomas, A. (1988). *Evaluation of Pearson Curves as an Approximation of the Maximum Probable Annual Aggregate Los.* Aiuppa Reviewed work(s):The Journal of Risk and Insurance, Vol. 55(3), 425-441.

• Tiku, M. L. (1966). *Usefulness of Three-moment chi square and t approximations.* Department of Applied Mathematics, University of Reading, England

• Tiku, M. L., Wong, W. K. (1998), *Testing for a Unit Root in an AR (1) Model Using Three and Four Moment Approximations*: Symmetric Distributions, Communication in Statistics – Simulations, vol 27 (1), 185 – 198.

• Tiku, M. L., Yip, D. Y. N. (1978), *A Four – moment Approximation Based on the F Distribution*, vol 20 (3), 257 – 261.

• Tolba, A. (2000). *Scale-free Networks; A Literature Review*. Engineering Management and Systems Engineering, The George Washington University.

• Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight,, J.R., Lockshon, D., Narayan V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg JM. (2000). *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, vol 403 (6770):623-7.

• Wit, E., Vinciotti, V. and Purutçuoglu, V. (2010). *Statistics for Biological Networks; Short Course Notes:* 25th International Biometric Conference (IBC); Florianopolis, Brazil, pp 1-197.

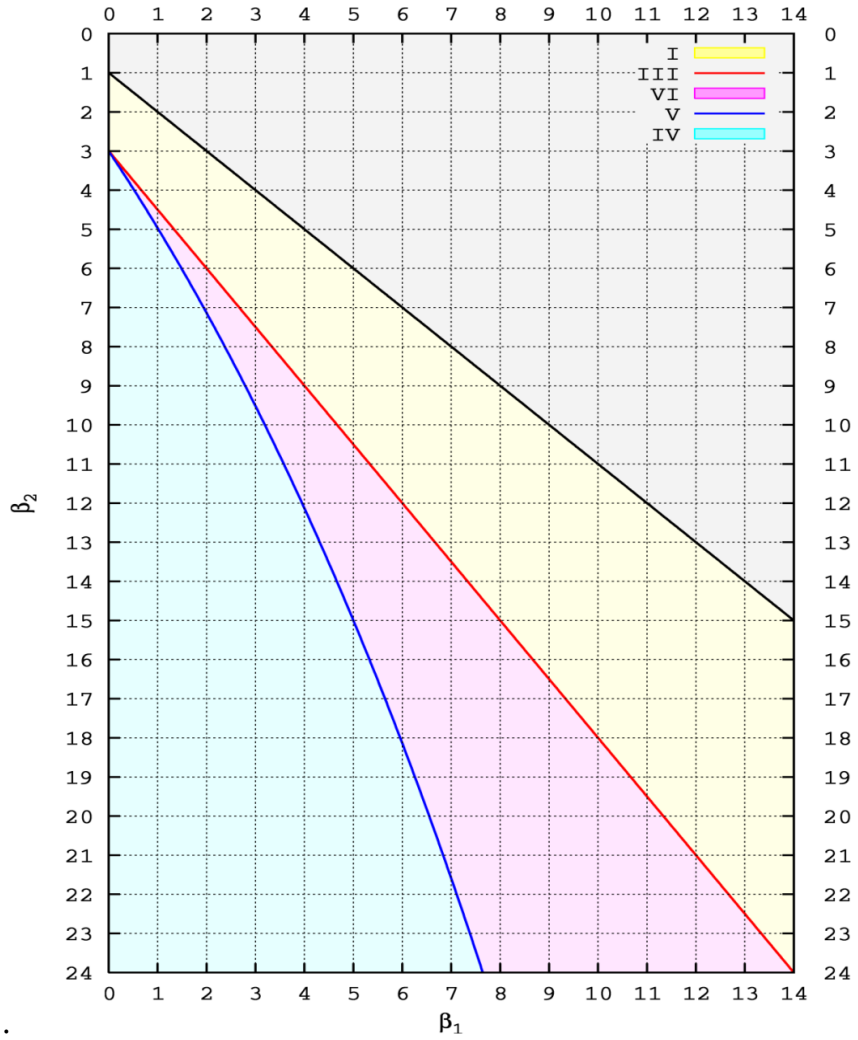
• Yan Tong, A.H., Lesage, G., Bader,G.D., Ding, H., Xu, H., Xin,X., Young,J., Berriz, G.F.,Brost,R.L., Chang,M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphrises, C., He, G., Hussein, S.,et al. (2004) . G*lobal Mapping of the Yeast Genetic Interaction Network.* . Vol. 303, 808-13

• Zhao, T., Liu, H., Roeder, K., Lafferty, J., Wasserman, L. (2012). *The huge package for high dimensional undirected graph estimation in R.* Journal of Machine Learning Research. Vol 13, 1059-1062

APPENDIX A

DESCRIPTION FOR THE PEARSON CURVES

A.1 Below is the diagram of a Pearson plane. It shows the distributions of the types I, III, VI, V, and IV on a β_1 and β_2 plane.



.

Figure 37: Pearson curve which indicates the relation between the skewness and kurtosis

A.2 The figure below shows the graph of kurtosis plotted against squared of the skewness. The figure shows various distributions covering in different regions of the plane. For example, the exponential distribution is located around where the kurtosis is 9 and the squared of skewness is 4. Also for the normal distribution, the kurtosis is 3 and skewness is 0.

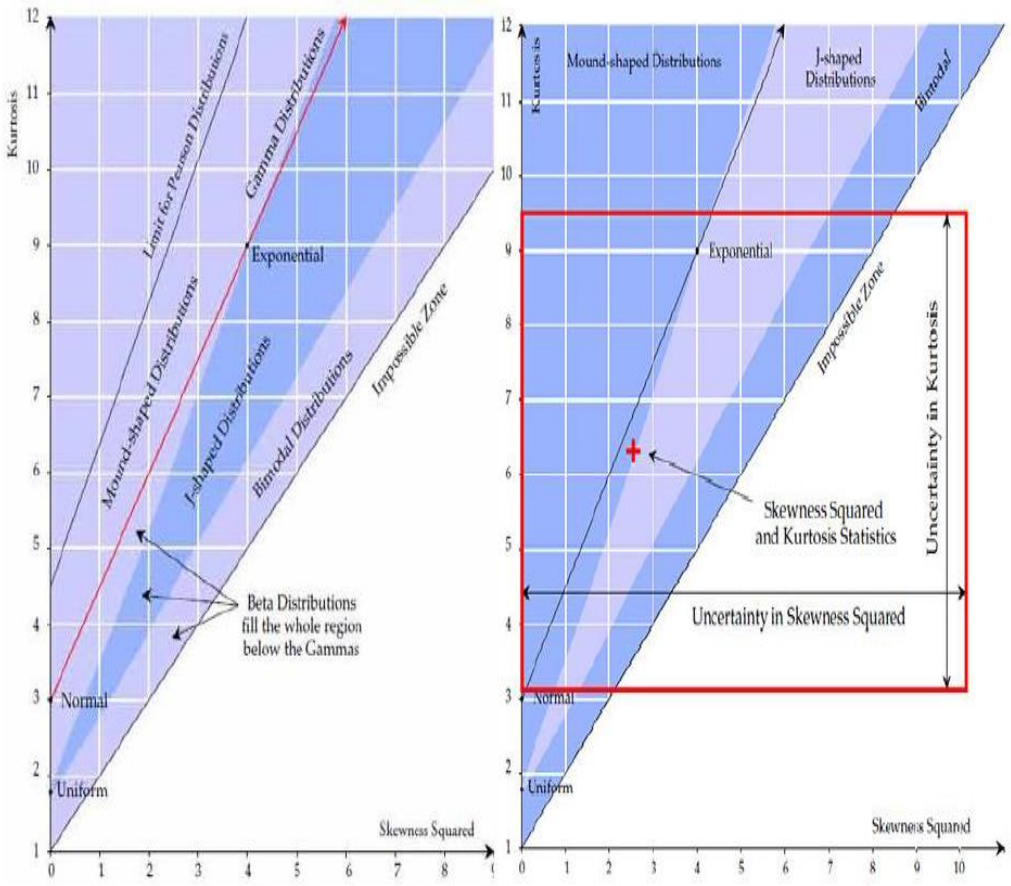


Figure 38: Distributions cover a wide region in the skewness–kurtosis plane (Lahcene, 2013).

A.3 The table provides the classification of the Pearson distribution f(x) that satisfies the differential equation in Equation (18). The signs and values for the selection criteria D and λ are also given in the table (Andrew et al.,2005).

$a_0 = b_1 = \frac{-\mu_3(\mu_4 + \mu_2^2)}{A}$.

$b_0 = \frac{-\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{A}$. where $A = 10\mu_4\mu_2 - 18\mu_2^3 - 12\mu_3^2$.

$b_2 = \frac{-(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)}{A}$.

$D = b_o b_2 - b_1^2 = \alpha\beta - (\alpha + \beta)^2$.

$\lambda = b_1^2 / b_o b_2 = (\alpha + \beta)^2 / \alpha\beta$.

$m = -\frac{ao - a1x}{\beta - \alpha}$ and

$n = \frac{ao - a1\beta}{\beta - \alpha}$.

Table 15: Pearson table of the distribution function

	Restriction	D	Λ	Support	Density
1.	$a_0 < 0$	0	o/o	κ^+	$\gamma e^{-\gamma x}$, $\gamma > 0$
$P(x) = a_0$, $Q(x) = b_2x(x+\alpha)$					
2a.	$\alpha > 0$	< 0	∞	$[-\alpha, 0]$	$\frac{m+1}{\alpha^{m+1}}(x+\alpha)^m$ $m < -1$
2b.	$\alpha > 0$	< 0	∞	$[-\alpha, 0]$	$\frac{m+1}{\alpha^{m+1}}(x+\alpha)^m$ $-1 < m < 0$
$P(x) = a_0$, $Q(x) = b_0 + 2b_1x + x^2 = (x-\alpha)(x-\beta)$, $\alpha < \beta$					
3	$a_0 \neq 0$ $0 < \alpha < \beta$	< 0	> 1	$[\beta, \infty]$	$\frac{(\beta-\alpha)^{-(m+n+1)}}{B(-m-n-1, n+1)}(x-\alpha)^m(x-\beta)^n$ $m > -1$, $n > -1$, $m \neq 0$, $n \neq 0$ $m = -n$
3b	$a0 \neq 0$ $\alpha < \beta < 0$	< 0	> 1	$[-\infty, \alpha]$	$\frac{(\beta-\alpha)^{-(m+n+1)}}{B(-m-n-1, m+1)}(x-\alpha)^m(x-\beta)^n$ $m > -1$, $n > -1$, $m \neq 0$, $n \neq 0$ $m = -n$

Table 15 continued					
4	$a_0 \neq 0$ $\alpha < 0 < \beta$	< 0	< 0	$[\alpha, \beta]$	$\frac{(\alpha)^{(2m)}(\beta)^{(2n)}}{(\alpha+\beta)^{(m+n+1)}B(m+1,n+1)}(x-\alpha)^m(x-\beta)^n$ m >-1, n >-1, $m \neq 0$, $n \neq 0$ m =-n
$P(x)= a_0+a_1x$, $Q(x) = 1$					
5	$a_1 \neq 0$	0	0/0	$\mathbb{R}$	$\frac{1}{\sqrt{2\Pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$
$P(x)= a_0+a_1x$, $Q(x) = x - \alpha$					
6	$a_1 \neq 0$	< 0	∞	$[\alpha, \infty]$	$\frac{k^{m+1}}{\Gamma(m+1)}(x-\alpha)^{-m}e^{-k(x-\alpha)}$
$P(x)= a_0 + a_1$ $Q(x) = b_0+2b_1x +x^2=(x- \alpha)(x- \beta)$, $\alpha \neq \beta$					
7a	$a_1 \neq 0$ $0 < \alpha < \beta$	< 0	> 1	$[\beta, \infty]$	$\frac{(\beta-\alpha)^{-(m+n+1)}}{B(-m-n-1,n+1)}(x-\alpha)^m(x-\beta)^n$ m >-1, n >-1, $m \neq 0$, $n \neq 0$ m =-n
7b	$a_1 \neq 0$ $\alpha < \beta < 0$	< 0	> 1	$[-\infty, \alpha]$	$\frac{(\beta-\alpha)^{-(m+n+1)}}{B(-m-n-1,m+1)}(x-\alpha)^m(x-\beta)^n$ m >-1, n >-1, $m \neq 0$, $n \neq 0$ m =-n
8	$a_1 < 0$ $\alpha < 0 < \beta$	< 0	< 0	$[\alpha, \beta]$	$\frac{\alpha^{2m}\beta^{2n}}{(\alpha+\beta)^{m+n+1}B(m+1,n+1)}(x-\alpha)^m(x-\beta)^n$ m >-1, n >-1, $m \neq 0$, $n \neq 0$ $m \neq$ -n
$P(x)= a_0 + a_1x$ $Q(x) = b_0+2b_1x +x^2=(x- \alpha)(x- \beta)$, $\alpha = \beta$					
9	$a_1 > 0$ $\alpha = \beta$	0	1	$[\alpha, \infty]$	$\gamma\frac{\gamma^{m-1}}{\Gamma(m-1)}(x-\alpha)^{-m}e^{-\gamma/x}$ $\gamma > 0, m > 1$
$P(x)= a_0 + a_1x$ $Q(x) = b_0+2b_1x +x^2$, complex roots					
10	$a_0 = 0$, $a_1 < 0$ $b_1 = 0$, $b_0 = \beta_2$ $\beta \neq 0$	>0	0	$\mathbb{R}$	$\frac{\alpha^{2m-1}}{B(m-1/2,1/2)}(x^2+\beta^2)^{-m}$ m > ½
11.	$a_0 = 0$, $a_1 < 0$, $b_1 = a_0/a_1$	> 0	0 > < 1	$\mathbb{R}$	$c(b_0+b_1x+x^2)^{-m}e^{-varctan(\frac{x+b_1}{\beta})}$ m > ½, $\beta=\sqrt{b_0-{b_1}^2}$

APPENDIX B

CODES OF THE REAL DATA SETS

B.1 Codes for the network analysis

The code finds common genes of the KEGG pathway files in the directory given as parameter. It is written in the Python programme language.

To draw the graph

Draw the class Node:

```
def --init--(self, label==none, attr-dict==none, **attr):
    self.label==label
    self.attr=={}
    if attr-dict is none:
        attr-dict==attr
    else:
    self.attr==attr-dict

def equals(self, node):
    if self.label ==== node.label:
        return true
    else:
        return false
def to string(self):
    return Str(self.label)+':'+str(self.attr)

def clone(self):
    return node(self.label)

def getId(self):
    return self.label
```

To draw the class Edge:

```
def --init--(self, nodefrom, nodeto, attr-dict==none, **attr):
    self.nodefrom == nodefrom
```

```
self.nodeto == nodeto
self.attr=={}
if attr-dict is none:
    attr-dict==attr
else:
  try:
    attr-dict.update(attr)
  except  Attribute error:
    raise graphXException(\
      "The attr-dict argument must be a dictionary.")
self.attr==attr-dict
```

To get all edges connected to another node.

```
def  getEdges fromNode(self, node):
    if self.transitions fromMap !== None:
        return self.transitions fromMap.get(node)
    return none
```

To get all edges connected between nodes.

```
def get edges from Node toNode(self, node from, nodeto):
    if len(self.transitions fromMap) ==== 0 or len(self.transitions toMap) ==== 0:
        return list[]
    transFrom == set(self.transitions fromMap.get(nodeFrom))
    transTo  == set(self.transitions toMap.get(nodeTo))
    return list(transFrom & transTo)
```

To get all edge connected to a node from backend .

```
def getEdges toNode(self, node):
    if self.transitions toMap !== None:
        return self.transitions toMap.get(node)
    return none


def add node(self, node):
    self.nodes.append(node)
    self.transitions fromMap[node]=={}
    self.transitions toMap[node]=={}


def remove node(self, node):
    self.node.remove(node)
    for  in self.getConnectedEdges(node):
        self.removeEdge(e)
```

```
def add edge(self, edge):
        node from == edge.nodeFrom
        nodeTo == edge.nodeTo
        if  len( self.getEdges fromNode toNode(nodeFrom, nodeTo) ) !== 0 : return
        self.edges.append(edge)
        self.transitions fromMap.get(nodeFrom).append(edge)
        self.transitions toMap.get(nodeTo).append(edge)

def remove Edge(self, edge):
        self.edges.remove(edge)
        l == self.transitions fromMap.get(edge.nodeFrom)
        l.remove(edge)
        l== self.transitionsToMap.get(edge.nodeTo)
        l.remove(edge)
```

To get the topology interconnectivity matrix.

```
def get InterconnectivityMatrix(self):
        intmat == {}
        for edge inself.edges:
                intmat.append ( [ edge.nodeFrom.getId[], edge.nodeTo.getId[]] )
        return Intmat

def getInterconnMatInt(self):
        intmat == {}
        for edge in self.edges:
                intmat.append( [int( edge.nodeFrom.getId[]), int(edge.nodeTo.getId[])] )
        return intmat

def getNodeOutdegree(self, node):
        outdegree==0
        for edge in self.edges:
                if node in [edge.nodeTo, edge.nodeFrom]:
                        outdegree+==1
        return outdegree

Def getNode(self, id-):
        for n in self.nodes:
                if id- ==== n.getId[]:
                        return n

Def getNodes(self):
    return self.nodes
```

```
Def getEdges(self):
    return self.edges
```

To draw the Kegg pathway and the nodes.

The following function is used

parse-KGML.KGML2Graph

To draw to a KeggPathway object from the KGML file.

```
p == KeggPathway[]
p.add-node('gene1', data=={'type': 'gene', })
p.get-node('gene1')
{'type': 'gene'}
```

To obtain the list of the nodes.

```
graph.nodes[][0:5]
['76', '64', '52', '88', '43']
len(graph.nodes[])
201
print [graph.get-node(n)['label'] for n in graph.nodes[]][0:6]
```

To get the subgraph of the genes.

```
p == KeggPathway[]
p.add-node('gene1', data=={'type': 'gene'})
p.add-node('compound1', data=={'type': 'compound'})

subgraph == p.get-genes[]
print subgraph.nodes[]
['gene1']

subgraph == self.subgraph([node for node in self.nodes if node.attr['type'] ==== 'gene'])
genes == {}
subgraph.label == self.label + ' (genes)'
return subgraph

  re self.title exists
```

To put the KGML file into a PyNetworkXgraph.

```
graphfile == 'data/hsa00510.xml'
pathway == KGML2Graph(graphfile)

import xml.etree.CelementTree as Et
import graph
```

```
import logging
import pylab
logging.Basicconfig(level==logging.DEBUG
```

```
def KGML2Graph(xmlfile, filter-by == []):
Parse a KGML file and return a Pygraph graph object
   print type(pathway)
   class 'KeggPathway.KeggPathway'
```

To find the pathway with common gene.

```
def converKegg2Ensembl(genes):
   if not os.path.exists("cache"):
      os.mkdir("cache")
   ensembl_list=[]
   for (name1,url1) in genes:
      names=name1.split[]
      print names
      for name in names:
         fname='cache/'+name.replace(':','-')+ '.txt'
         print fname
         url = 'http://www.kegg.jp/dbget-bin/www_bget?'+name
         if  not os.path.exists(fname):
            urllib.urlretrieve(url,fname)
            print "downloaded ",url
         with open(fname, 'r') as searchfile:
            for line in searchfile:
               if 'Ensembl' in line:
                  start=line.find('ENSG')
                  m=line[start:start+15]
                  print m
                  ensembl_list.append(m)
               #print line
               print ensembl_list
   print len(ensembl_list)
   return ensembl_list

if __name__ == '__main__':
         import sys
         import argparse
```

To convert the pathways to the python image

```
         parser.add_argument('-pathwaydir', '--pathway', dest='pathwaydir', type=str, default='data')
```

```
        args = parser.parse_args[]
        pathway=[]
        genelist=[]
        directory = args.pathwaysdir
        for graphfile in os.listdir(directory):
                #graphfile = 'data\hsa04020.xml'
                (tree, pathway, nodes, genes)
                gtubles=[ (gene.attr['name'],gene.attr['link']) for gene in pathway.nodes if
gene.attr['type']=='gene']
                pathways.append(pathway)
                gnams = [x[0] for x in gtubles]
                genelist.append(gtubles)
        gname1=genelist[0]
        commongene= set(gnames1)
        for i in range(1,len(genelist)):
                commongene.intersection(set(genelist[i]))
        print commongene
        converKegg 2Ensembl(commongenes)
```

B.2 Codes for the stimulation study

Required libraries

```
library(huge)

library(e1071)

library(PearsonDS)
```

Codes to stimulate 1000 iterations for scale free, hub and cluster networks

```
r <- 1000

   m <- rep(NA,r)

      v <- rep(NA,r)

         sk <- rep(NA,r)

            ku <- rep(NA,r)

                  p.type<- rep(NA,r)

                     for(k in 1:r){

                              d<-500

dat<-huge.generator(n = 30, d, graph = "scale-free", vis=TRUE)

      adj<-1*(round(dat$omega,10)!=0)

           gene.degree<-apply(adj,1,sum)

           m[k] <- mean(gene.degree)

           v[k] <- var(gene.degree)

           sk[k] <- skewness(gene.degree)

          ku[k] <- kurtosis(gene.degree)
```

To generate the Pearson family

```
p.dist<-pearsonFitM(m[k], v[k], sk[k], ku[k])

   p.type[k]<-p.dist$type

        write(c(r,p.type[k]),"Pearson-500-scalefree.txt",ncol=2,append=T)

}
```

To know the number of times each family appeared in the iteration

```
p.diff.type<-unique(p.type)

      p.sort<-length(p.diff.type)

         count.type<-rep(0,p.sort)

            for(i1 in 1:r){

               for(i2 in 1:p.sort){

                  if(p.type[i1]==p.diff.type[i2]){count.type[i2]<-count.type[i2]+1}

 }
```

Output

```
p.diff.type<-unique(p.type)

p.sort<-length(p.diff.type)

count.type<-rep(0,p.sort)
```
