STATISTICAL NETWORK ANALYSIS FOR COLLABORATION IN APPLIED MATHEMATICS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SİMGE GÜNERİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
SCIENTIFIC COMPUTING

SEPTEMBER, 2013

Approval of the thesis:

STATISTICAL NETWORK ANALYSIS FOR COLLABORATION IN APPLIED MATHEMATICS

submitted by **SİMGE GÜNERİ** in partial fulfillment of the requirements for the degree of **Master of Science in Department of Scientific Computing, Middle East Technical University** by,

Prof. Dr. Bülent Karasözen
Director, Graduate School of **Applied Mathematics** _____

Prof. Dr. Bülent Karasözen
Head of Department, **Scientific Computing** _____

Prof. Dr. Gerhard Wilhelm Weber
Supervisor, **Institute of Applied Mathematics, METU** _____

Examining Committee Members:

Assist. Prof. Dr. Ceylan Talu Yozgatlıgil
Department of Statistics, METU _____

Assist. Prof. Dr. Berna Burçak Başbuğ Erkan
Department of Statistics, METU _____

Prof. Dr. Gerhard Wilhelm Weber
Institute of Applied Mathematics, 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: SİMGE GÜNERİ

Signature :

ABSTRACT

STATISTICAL NETWORK ANALYSIS FOR COLLABORATION IN APPLIED MATHEMATICS

Güneri, Simge

M.S., Department of Scientific Computing

Supervisor : Prof. Dr. Gerhard Wilhelm Weber

September, 2013, 86 pages

In spite of easy access to academic information, person-to-person contact is still significant so as to provide the academic issues to be discussed in more detail and well-developed. So, throughout the thesis, we investigate one-to-one communication and frequency of activities performed together in the researcher communities from different study fields of applied mathematics. We deal with this problem by taking each researcher community as a network where its entities are considered the researchers and two entities are connected if the two reseachers published an article together. The underlying data were derived from the archives of ArXiv. Firstly, we apply the statistical procedure based on the hypothesis test to model our collaboration networks. By use of the technique combining both maximum likelihood estimation method and Kolmogorov-Smirnov statistic, we investigate the cumulative distribution of its degree sequence for each network which gives the probability of total collaborators of an arbitrarily chosen mathematician being greater than or equal to a specified number. Secondly, we evaluate mean number of collaborators of the researchers, mean size of small groups of connected researchers, size of a giant assemble of connected researchers, mean shortest distance and maximum shortest distance between the researchers in the giant assemble both empirically and theoretically. In addition, we also calculate the degrees of clustering and mutuality, indicative properties of real networks. Finally, as observed in most real networks, we see that relationships on almost all

communities show a small-world effect which is an indication of small mean distance and high clustering. However, the communities are not scale free. Beside scientific collaboration networks, we can also use our findings to make an analysis of the other types of networks such as gene regularity networks and social media networks.

Keywords: Kolmogorov-Smirnov statistic, Maximum Likelihood Estimation method, Scale-free networks, Clustering, Mutuality, Small-world effect

ÖZ

UYGULAMALI MATEMATİKTEKİ İŞBİRLİĞİN İSTATİSTİKSEL AĞ ANALİZİ

Güneri, Simge

Yüksek Lisans, Bilimsel Hesaplama Bölümü

Tez Yöneticisi : Prof. Dr. Gerhard Wilhelm Weber

Eylül 2013, 86 sayfa

Akademik bilgiye kolay erişim olmasına rağmen, akademik konuların daha detaylı tartışılmasını ve iyi gelişmesini sağladığı için birebir görüşme hala önemlidir. Bu nedenle, tez boyunca uygulamalı matematiğin farklı çalışma alanlarındaki araştırmacı topluluklarındaki birebir iletişimi ve aktivitelerin birlikte yapılma sıklığını inceledik. Bu problemle her bir topluluğu birimlerinin araştırmacılar ve iki biriminin ancak bu iki araştırmacı ortak makale çıkardığı zaman bağlantılı olacağı bir ağ olarak ele alarak uğraştık. Veriler ArXiv sitesinin arşivlerinden elde edildi. İlk olarak, işbirlik ağlarımızı modellemek için hipotez testine dayanan istatistiksel bir metoda başvurduk. En çok olabilirlik kestirimi metodu ve Kolmogorov-Smirnov istatistiğin kombinasyonu bir teknik kullanarak, her bir ağ için onların derece dizilerinin kümülatif dağılımlarını hesapladık. Bu dağılım rastgele seçilen bir matematikçinin aranan sayıdan büyük ya da eşit işbirlikçiye sahip olma olasılığını verir. İkinci olarak, araştırmacıların işbirlikçilerinin ortalama sayısını, iletişimdeki araştırmacıların küçük topluluklarının ortalama boyutunu, iletişimde olan büyük topluluğun boyutunu, büyük gruptaki araştırmacılar arasındaki ortalama ve maksimum en kısa mesafeyi hem gözlemsel hem de teorik olarak hesapladık. Ek olarak, gerçek ağların belirleyici özellikleri kümelenme ve karşılıklılık derecelerini de ölçtük. Son olarak, gerçek ağların çoğunda gözlemlendiği gibi, çoğu topluluktaki ilişkilerin küçük dünya etkisi gösterdiğini gördük. Fakat, bu topluluklar ölçüsüz değillerdir. Bilimsel işbirlik ağlarının yanı sıra, bu bulgularımızı gen-düzenleyici ağlar ve sosyal medya ağları gibi diğer

tipteki ağların analizini yapmak için de kullanabiliriz.

Anahtar Kelimeler: Kolmogorov-Smirnov istatistik, En çok olabilirlik kestirimi metodu, Ölçüsü olmayan ağlar, Kümelenme, Karşılıklılık, Küçük dünya etkisi

ACKNOWLEDGMENTS

I owe heartfelt thanks to Prof. Dr. Gerhard-Wilhelm Weber for encouraging and supporting me generously while dealing with such an enourmous task. I would like to express my deepest gratitude to Dr. Erik Kropat who has made a significant contribution to this thesis by doing a great deal of proofreading and clarifying troublespots. Moreover, my greatest debt is to Assist. Prof. Dr. Ceylan Yozgatlıgil and Assist. Prof. Dr. Berna Burçak Başbug Erkan who have made extremely useful suggestions. I am also grateful to Simge Özçelik who has read and commented on the linguistic aspects of certain parts of the material. Finally, I would like to express my appreciation to my family. Without their supports and encouragements, this thesis would never have been completed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Network science is an interdisciplinary research area which studies complex networks using theories and methods arising from graph theory (mathematics), model selection procedure and clustering analysis (statistics and probability), data mining and pattern recognition (computer science) and statistical mechanics (physics). In network science, a complex network is a graph with a nontrivial and complicated topological structure in which patterns of connections between their entities are neither regular nor random. There are some key aspects which are common in all complex systems. All complex networks comprise many interacting parts. Although a complete network shows a macroscopic collective behaviour, its each component has own specific structure and intrinsic functions. Moreover, discarding a small part of the system may affects the whole system substantially. In other words, the complex system show a behavior in such a way that its any feature is not simple sum of its parts.

The most well known classes of complex networks are scale-free networks and small-world networks. Scale-free networks have a heavy tail in their degree distributions. Systems with such a degree distribution are dominated by only a few entities. This brings about two important properties, vulnerability to targeted attacks and preferential attachment. Another type of networks, small-world networks, show high clustering and have a small average number of connections between any two entities.

A potential of the study of real networks is proved by the growing number of research groups occupying with this field. For example, researchers have been studying on how epidemic diseases disperse in certain social communities for many years. These all scientific findings yield many vaccination strategies to reduce the impact of diseases on affected populations and controlling or halting outbreaks of contagious diseases. Likewise, they have also examined the spread of malicious softwares and viruses in worldwide communications systems to fortify Internet security. Concisely, regardless of whether they are natural or human-made, most structures in the world are explained using real networks. Some networks mainly studied could be represented as follows:

- *Biological Networks:* They consider the structures of living organisms or the interactions between them. For instance, gene regularity networks explain

how genes become activated or deactivated and what kind of proteins are produced in a cell at a certain time period. Furthermore, protein-protein interaction networks describe the interaction between proteins in consequence of which process protein complexes are build or a protein binding to another protein is modified. These the two processes play a fundamental role in many diseases such as cancer.

  - Ecological networks (food-webs, predator-prey relationships) [10, 24, 29, 36, 37, 52, 58, 64, 80, 86, 102, 113],
  - Epidemic Spreading [94, 101],
  - Genetic regulatory networks [33, 34, 48, 49, 66, 67, 68, 124, 125, 126],
  - Metabolic networks [44, 62, 103, 119],
  - Neural networks [115, 116, 128],
  - Protein interaction [60, 61, 81, 114].

- *Information Networks:* They represent the exchange of information or services among people, communities or constitutions in order to allow efficient research.

  - Citations between academic papers [38, 74, 105, 108, 111, 117, 127],
  - Network in linguistics [59],
  - World-Wide-Web (WWW) [1, 6, 11, 12, 22, 45, 70, 72] . (You can find its data refreshed in every week on Erdös WebGraph.).

- *Social Networks:* They state a pattern of the relationships between people, groups, companies or institutions. Social networks are also a good way to understand the diffusion of news and rumuors between people.

  - Business relationships between companies [46, 47, 78, 84],
  - Coapperance networks in which individuals are linked by mention in the same context particularly on web pages or in newspaper articles [2, 30, 69],
  - Friendship [31, 43, 89, 106, 107],
  - Human sexual contacts [17, 63, 71, 77, 90, 104],
  - Intermarriages between families [100],
  - Movie-actor collaboration [7, 11],
  - Rumour Spread [18],
  - Science collaboration [14, 15, 21, 32, 55, 75, 83, 87, 91, 92, 93, 95],
  - Social circles from Facebook, Twitter, Google+ [82].

- *Technological Networks (or Transportation Networks):* They are man-made networks which are designed for the distribution of some commodities or natural resources.

  - *Transport of people and goods:* Airline routes [8], networks of roads [65], railways [73, 112], pedestrian traffic [27], river [35, 79, 109, 110],

– *Transport of electric:* Electric power grid [8, 120, 121],
– *Transport of information:* Email-network [56, 98], Internet (the network of physical connection between computers) [23, 26, 42], phone-call network [3, 4].

1.1 Scientific Collaboration

A science collaboration (coauthorship) network is a collection of scientific researchers, each of which is tied to some of the others. In such a network, a tie between any two researchers occurs when they have written at least one paper jointly. Since we omit the relationships between some colleagues who interact with each other but do not write any paper together, the definition of collaboration constructed in science collaboration networks is only moderately robust.

1.1.1 Erdös Number

In the period graph theory focused on only regular graphs, the Hungarian mathematicians Paul Erdös and Alfred Renyi's studies about random graphs attracted great attention. Erdös wrote 1525 mathematical articles [54], having 511 different collaborators in his lifetime [53].

The Erdös number describes the collaborative distance between any scientist and Paul Erdös. Anyone who has published an article with him has Erdös number 1. Continuing with the same thinking, anyone with Erdös number 2 has studied with someone who has studied with Erdös. The highest known Erdös number is 15 [91]. In physics, this number is analogous to Einstein number. We note:

- Albert Einstein (Physics) has Erdös 2 [25].
- Carl Sagan (Astronomy) has Erdös 4 [25].
- John Nash (Economics and Mathematics) has Erdös 4 [25].
- Stephen Hawking (Cosmology) has Erdös 4 [25].

1.2 Contributions of the Thesis

Improvements in technology enable us to access all academic events, publications, books and lectures related to any issue we are interested in. Nevertheless, common activities carried out together and one on one meetings take an indispensible place in arising of more different ideas and increasing the quality of scientific researches. While working together to tackle any problem, every researcher uses a different point of view. This provides common works to be extended by different ideas. In addition, even if some researchers do not work on that subject, by stimulation

of their collaborators they can tend towards that topic and carry on their works on it. Judging from this, we can say that the more person-to-person contact, the faster and wider prevalence of scientific knowledge. In order to investigate the topological structure of the spread of academic information and the frequency of the studies performed together in applied mathematics, we construct science collaboration networks for researcher communities, each of which comprises of researchers studying on a subfield of applied mathematics.

1.3 Outline of the Thesis

In the thesis, collaboration between the mathematicians in each subfield of applied mathematics will be investigated. Applied mathematics can be separated into eight main categories: cryptography and security, financial mathematics, mathematical biology, mathematical physics, numerical analysis, optimization and control, probability and statistics. All data are derived from the archives of ArXiv. The process for obtaining this data is explained concisely in Appendix A.

The thesis is structured as follows:

- *Chapter 1:* We mention briefly different types of networks and Erdös number. Besides, we make a summary of our study and contributions of this study.
- *Chapter 2:* We talk on Erdös-Rènyi random graphs and generating function method used for random graphs analysis. Then, we represent in what properties real networks differ from random graphs.
- *Chapter 3:* We apply the statistical procedure based on the hypothesis test to model our collaboration networks. By use of the method combining both maximum likelihood estimation method and Kolmogorov-Smirnov test, we investigate the cumulative distribution of its degree sequence for each network which gives the probability for an arbitrarily chosen mathematician to have greater than or equal to a searched number of co-workers.
- *Chapter 4:* We assess mean co-workers of the scientists, mean size of small groups of connected scientists, existence and size of a giant assemble of connected scientists, mean shorthest distance and maximum shortest distance between the scientists in the giant assemble, the degree of clustering and mutuality both empirically and theoretically.

  In the light of the evaluation results, we try to estimate the behaviour of observed data. According to their distribution functions, whether they exhibit the Barabási-Albert model which is also known as a scale-free network can be interpreted readily. Besides, judging from the findings about their distances and clustering densities, whether they have a small-world effect asserted by Strogatz and Watts can be determined. In addition, as theory depends mainly on random graphs, we catch an opportunity to see the differences between the Erdös-Rènyi random graph model and real networks.

- *Chapter 5:* Finally, we conclude this thesis by summarizing our contributions and discussing directions for future work.

CHAPTER 2

PRELIMINARIES

In this chapter, we provide some basic facts on important types of networks. Firstly, we will address the theory of random graphs in section 2.1 and provide its some useful characteristic properties by means of generating functions. Then we turn to graph measures of real networks in section 2.2. In particular, we introduce small-world networks in form of the Watts-Strogatz model as well as scale-free networks which are based on the Barabási-Albert model.

2.1 Random Graph

2.1.1 Notation and Basic Facts

Definition 2.1. A *graph* is a pair of sets $G = (V, E)$, where V is a set of n *vertices* $V_1, V_2, \ldots, V_n$ and E is a set of *edges* that connect two elements of V [5].

Example 2.1. A graph where the set of the vertices are $V = \{1, 2, 3, 4, 5, 6\}$ and the set of the edge are $E = \{(1, 2), (1, 3), (1, 4), (2, 4), (3, 5), (3, 4), (4, 6)\}$ (see Figure 2.1).

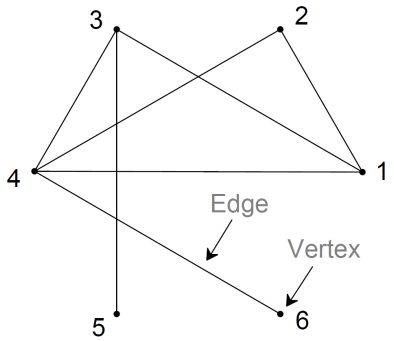


Figure 2.1: An example of a graph.

Vertice or vertex, the fundamental unit of networks, is also called a *site* (physics), a *node* (computer science), or an *actor* (sociology) [96]. Edge, the line connecting two vertices, is also called a *bond* (physics), a *link* (computer science), or a *tie* (sociology) [96].

Definition 2.2. An edge is called *directed* if it runs in only one direction (such as a one-way road between two points), and *undirected* if it runs in both directions [96].

Definition 2.3. The *degree* of a vertex is the number of edges connected to that vertex [96]. Each vertex in an directed network has two degrees, an *in-degree*, which is the number of edges that point into the vertex, and an *out-degree*, which is the number pointing out [96].

A *neighbour* of a vertex is an another vertex connected to that vertex by a simple edge. Second neighbours of a vertex is the neighbours of its neighbours.

Example 2.2. The graph in Figure 2.1 is undirected. It has 2 vertices with degree 1, one vertex with degree 2, and 3 vertices with degree 3. In first graph in Figure 2.2, in-degree of vertex 1 is 1, whereas out-degree of vertex 1 is 0. Vertex 2 has both in-degree and out-degree.

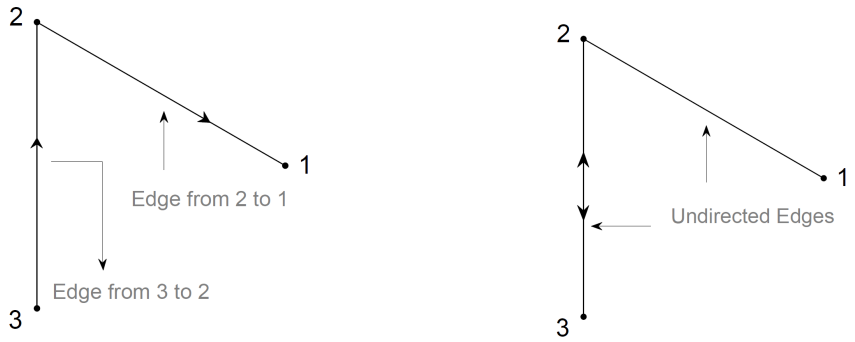


Figure 2.2: Examples of directed and undirected graphs.

Definition 2.4. The *component* to which a vertex belongs is that set of vertices that can be reached from it by paths running along edges of the graph [96]. In a directed graph a vertex has both an *in-component* and an *out-component*, which are the sets of vertices from which the vertex can be reached and which can be reached from it, respectively [96].

Remark 2.1. In graph theory, a component size generally refers to numbers of edges included in a component. Many people focusing on random graphs, however, mean numbers of vertices in a component when studying component size.

Definition 2.5. A *geodesic path* is the shortest path through the network from one vertex to another [96]. We note that there may be and often is more than one geodesic path between two vertices [96].

Definition 2.6. The *diameter* of a network is the length (in number of edges) of the longest geodesic path between any two vertices [96]. A few authors have also used this term to mean the average geodesic distance in a graph, although strictly the two quantities are quite distinct [96].

Definition 2.7. An *adjacency matrix* is a means of representing which vertices of a graph are adjacent to which other vertices [50].

Example 2.3. The graph in Figure 2.3 includes three distinct components. Size of pink marked component is 1, while size of green marked is 3. Blue marked one which is the component including maximum number of nodes (9 nodes) is the giant component for this case. There are three distinct paths between vertices 1 amd 2. These paths are $(2,1)$, $(2,5) \to (5,2)$, $(2,6) \to (6,4) \to (4,3) \to (3,8) \to (8,2)$. Geodesic path is $\{2,1\}$. The shortest distance is 1 between vertices 1 and 2, accordingly. The maximum distance in giant component is between 5 and 9. The path is $(5,2) \to (2,6) \to (6,4) \to (4,3) \to (3,7) \to (7,9)$. There is not any shorter way between the two vertices. So, diameter is equal to 6.

The adjacency matrix of blue marked part of the graph is

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

As first degrees of vertex 1 are vertices 2, 5 and 8, $A(1,2) = A(1,5) = A(1,8) = 1$ and $A(1,i) = 0$ for $i = 1,3,4,6,7,9$.

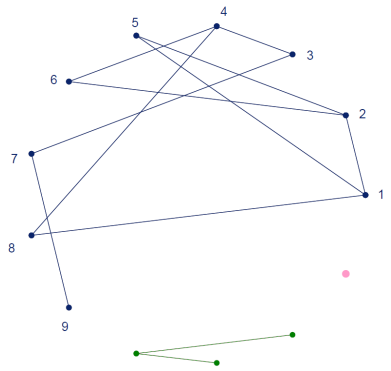


Figure 2.3: A graph including more than one component. (See Appendix C.3)

2.1.2 Random Graph Theory

Graph theory arised in 1736 with the study of Euler known as "Seven Bridges of Königsberg" where only small graphs with a high degree of regularity were taken in consideration. In the last half of the 20^{th} century graph theory started to develop in more statistical and algorithmic ways so that random graphs in which the edges are distributed randomly became the main focus of graph theory. Networks with a complex structure are in fact random, thus random graph theory is frequently preferable in the study of complex networks.

2.1.2.1 Erdös-Rényi Graph

In their studies [39, 40, 41], each edge between two vertices is considered to be present with independent probability p, and absent with probability $1-p$. As the maximum number of edges on the graph with n nodes is $\frac{n(n-1)}{2}$, the expected value of number of edges on this graph is $\frac{n(n-1)p}{2}$. Since each edge has two ends, the mean number of ends of edges in the graph is $n(n-1)p$. So, the mean degree of a vertex is

$$k = \frac{n(n-1)p}{n} = (n-1)p \approx np, \tag{2.1}$$

where the last approximate equality is good for large n. Figure 2.4 illustrates an example of the stepwise Morse code implementation which is based on the

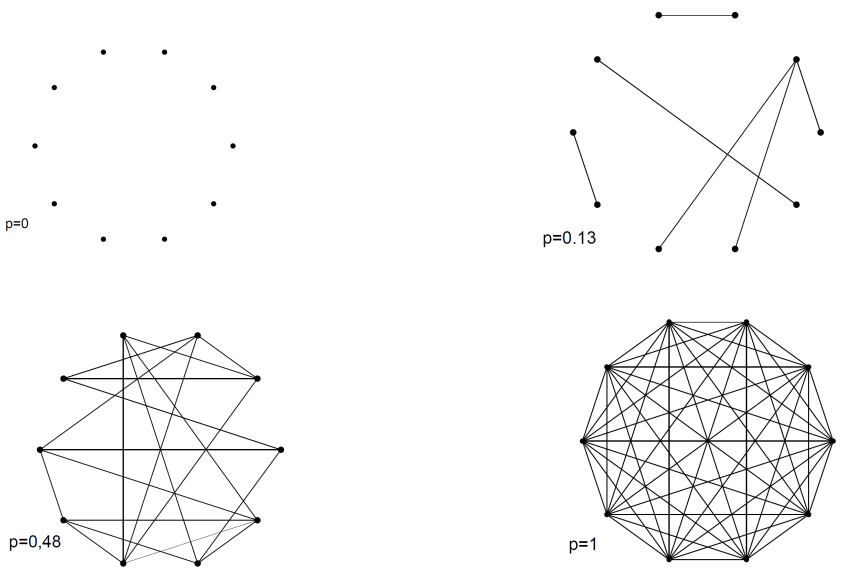


Figure 2.4: Random Graphs. (See Appendix C.3)

probabilistic method used for the proof of Erdös' Theorem [122]. Erdös and Rényi studied on only the distribution of the extremes in a degree sequence of a random graph. The distribution of its degree sequence was handled later by

Bollobás [19]. In a random graph with probability p of an edge being present, the degree k_i of a node i follows a binomial distribution with parameters $n-1$ and p [39],

$$P(k_i = s) = \binom{n-1}{s} p^s (1-p)^{n-1-s}. \tag{2.2}$$

While $n \to \infty$, the binomial distribution turns out to be a Poisson distribution [39, 40, 41]:

$$P(k_i = s) = \frac{k^s e^{-k}}{s!}. \tag{2.3}$$

2.1.2.2 Generating Functions

Newman, Strogatz and Watts [99] have extended the formulas of a Erdös-Rényi graph for random graphs with arbitrarily degree distributions using generating function formalisim. The most essential generating function in their studies is $G_0(x)$ which generates the probability p_s that an arbitrarily chosen vertex in a graph has a first degree k:

$$G_0(x) = \sum_{s=0}^{\infty} p_s x^s. \tag{2.4}$$

In all calculations, we suppose that this function satisties the condition

$$G_0(1) = 1. \tag{2.5}$$

Remark 2.2. It is clear that this function is positive definite. Owing to the fact that the generating function $G_0(x)$ is normalized (see Eqn. (2.5) and positive definite, it also absolutely convergent for all $|x| \leq 1$. Therefore, it has not any singularity in the region $|x| \leq 1$.

Definition 2.8. The *probability* p_s is given by the s^{th} derivative of G_0 according to [97, 99]

$$p_s = \frac{1}{s!} \frac{d^s G_0}{dx^s}\bigg|_{x=0}. \tag{2.6}$$

Looking at Eqn. (2.6), we can say that $G_0(x)$ includes all the probabilities p_s so that k^{th} derivation of $G_0(x)$ gives the probability that any vertex chosen at random has a first degree k. For this reason, we say concisely that $G_0(x)$ "generates" p_s.

2.1.2.3 First and Second Neighbours

Definition 2.9. Mean first degree of a vertex k can be represented as [97, 99]

$$z = \langle k \rangle = \sum_{k=0}^{\text{max degree}} k p_k = G_0'(1). \tag{2.7}$$

We can also compute higher-order moments of the degree distribution using higher-order derivatives of its generating function [99].

$$\langle k^n \rangle = \sum_{k=0}^{\text{max degree}} k^n p_k = \left(x \frac{d}{dx^n} \right)^n G_0(x) \Big|_{x=1}. \tag{2.8}$$

Definition 2.10. The m^{th} power of the generating function G_0 [99]:

$$[G_0(x)]^2 = \left[\sum_{k=0}^{\infty} k p_k \right]^2 = \sum_{j,\, k=0}^{\infty} p_j p_k x^{j+k}, \tag{2.9}$$

where m is the number of arbitrarily chosen vertices in a graph and $G_0(x)^m$ is the generating function for the probability distribution of the sum of the first degrees of these vertices.

Now, we focus on the first degree distribution obtained by pursuing an arbitrarily chosen edge. Vertices in a graph mainly have different degrees. Moreover, a vertex whose degree is high will have a high possibility that we reach this vertex by following a random edge. Therefore, this probability is proportional to the degree of the vertex. Judging from this, the degree distribution of the vertex arrived by following an edge is linearly proportinal to kp_k.

We do not concentrate on the exact degree of a vertex which is arrived by an edge. We only consider the number of edges arising from a vertex and discard the randomly selected edge. So, the number we search will be only one less than first degree of that vertex. In addition, its probability distribution is represented as:

$$q_k = \frac{(k+1)p_{k+1}}{\sum_{j=0}^{\infty} j p_j}. \tag{2.10}$$

Furthermore, its mean degree of a vertex reached by following an edge is given as [97]:

$$\sum_{k=0}^{\infty} k q_k = \frac{\sum_{k=0}^{\infty} k(k+1)p_{k+1}}{\sum_{j=0}^{\infty} j p_j} = \frac{\sum_{k=0}^{\infty} k(k-1)p_k}{\sum_{j=0}^{\infty} j p_j} = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}. \tag{2.11}$$

Theorem 2.1. *Eqn. (2.11) gives the average number of vertices two steps away from our vertex via a particular one of its neighbours [97, 99]. Multiplying this by the mean degree of that vertex, which is just $z = \langle k \rangle$, we find that the mean number of second neighbours of a vertex is [97, 99]*

$$z_2 = \left\langle k^2 \right\rangle - \langle k \rangle . \tag{2.12}$$

The function [97, 99]

$$G_1(x) = \frac{\sum_{k=0}^{\infty} k p_k x^{k-1}}{\sum_{k=0}^{\infty} k p_k} = \frac{G_0'(x)}{G_0'(1)} = \frac{G_0'(x)}{z}, \tag{2.13}$$

generates the probability distribution of outgoing edges. Utilizing the powers feature of the generating function, we can also express $G_1 x$ which is a generating function for the probability distributon of the number of the second-nearest neighbours of a vertex in a graph as

$$\sum_{k=0}^{\infty} p_k [G_1(x)]^k = G_0(G_1(x)). \tag{2.14}$$

Likewise, continuing to apply same procedure for the distribution of third-nearest neighbors, we find that its generating function is described as $G_0(G_1(G_1(x)))$. Based on this, the mean number z_2 of second neighbours can be formualized as below [99]:

$$z_2 = \frac{d}{dx}\left[G_0(G_1(x))\right]_{x=1} = G_0'(1)G_1'(1) = G_0''(1). \tag{2.15}$$

Theorem 2.2. *By extension, the distribution of the numbers of m^{th} nearest neighbours is generated by $G_0(G_1(\ldots G_1(x)\ldots))$, with $m-1$ iterations of the function G_1 acting on itself [97, 99]. So, this generating function $G^m(x)$ can be described as follows [99]:*

$$G^m(x) = \begin{cases} G_0(x) & for\ m = 1, \\ G^{(m-1)}(G_1(x)) & for\ m \geq 2. \end{cases} \tag{2.16}$$

Then the mean number z_m of m^{th}-nearest neighbours is [99]

$$z_m = \left.\frac{dG^m(x)}{dx^m}\right|_{x=1} = G_1'(1)G_{m-1}'(1) = G_1'(1)z_{m-1}. \tag{2.17}$$

Given an initial condition $z_1 = z = G_0'(1)$, we get [97, 99]

$$z_m = [G_1'(1)]^{m-1} G_0'(1) = \left[\frac{z_2}{z_1}\right]^{m-1} z_1. \tag{2.18}$$

2.1.2.4 Component Sizes and Phase Transitions

A quantity in Eqn. (2.18) in fact either diverge or converge when we increase m. As seen clearly from Eqn. (2.18), divergence depends on the condition that z_2 is greater then z_1. Therefore, a phase transition occurs in any graph at the critical point where $z_2 = z_1$ as in the Erdös-Rényi graph [40]. When we use Eqn. (2.12), the condition is also expressed in a different way so that [85, 97, 99]

$$\left\langle k^2 \right\rangle - 2\left\langle k \right\rangle = \sum_{k=0}^{\infty} k(k-2)p_k = 0. \tag{2.19}$$

On account of the factor $k(k-2)$ in Eqn. (2.19), vertices whose first degrees are either zero or two do not change the sum in Eqn. (2.19). So, we conclude that removing or adding them do not alter the critical point for the phase transition and the existence of the giant component. Firstly, vertices with degree zero do not have connections to other vertices in a graph. So, any increase or decrease in their numbers does not make sense for the topological structure of a graph. In addition, vertices with degree two always remain between a pair of vertices. For this reason, we can discard or add such vertices as we desire.

- **Below the Phase Transition**

We always exclude the giant component, if there is one, while calculating the average size of components in a graph. As mentioned before, a graph has not a giant component below the phase transition [85, 97, 99].

The process of calculating the probability distribution of cluster sizes can be summarized as follows [97]. Firstly, we consider a given edge in a graph, follow each edge connected to it and then carry on this until a lap is completed. After a loop, the number of vertices we get gives the size of that cluster. Then we rule out all the vertices in a cluster whose size have been already calculated and continue to apply the same process for remaining vertices.

Now, we define $H_1(x)$ as the generating function which generates the probability distribution of sizes of components found in a graph. As mentioned before, size of a component is the total number of vertices in that component. Firstly, we consider all possibilities. When following an arbitrary edge, we can encounter with a single vertex without a passing edge at its end. Another possibility is that we can find a vertex on which one or more than one edges pass. Then, each edge takes us to another complete component whose size is also generated by $H_1(x)$ [97]. In Figure 2.5, the left-hand side of the equation represents the total probability of all possible sizes [97]. This probability can be expressed as the sum of the probabilities (right-hand side) of having only a single vertex (the circle), having a single vertex connected to one other component, or two other components, and so forth [97].

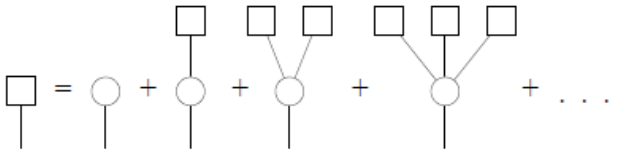


Figure 2.5: Schematic representation of the possible forms for the connected component of vertices reached by following a randomly chosen edge [97]. (We symbolize each component as square and each vertex as cycle.)

Theorem 2.3. *The probability distribution of the number of edges k passing the vertex except the one we reach that vertex by following is q_k in Eqn. (2.10). By*

the way, we can say that the probability distribution of the sum of the sizes of the k components is generated by $[H_1(x)]^k$ *owing to the powers property of generating functions [97, 99]. Using all the information, a generating function* $[H_1(x)]$ *for the total number of vertices accessible by pursuing an arbitrarily selected edge can be expressed as [97, 99]:*

$$H_1(x) = x \sum_{k=0}^{\infty} q_k [H_1(x)]^k = x G_1(H_1(x)), \tag{2.20}$$

where the leading factor of x accounts for the one vertex at the end of our edge, and we have made use of Eqn. (2.13).

In fact, we are mainly interested in the probability distribution of sizes of the component in which a vertex uniformly chosen at random stays. Since p_k is the degree distribution which gives the number of edges a randomly chosen vertex has and $H_1(x)$ is the generating function for the distribution of sizes of the component when we follow such an edge, we make use of them in order to get it.

Theorem 2.4. *Given any vertex, we represent the size of the entire component which includes that vertex as [97, 99]*

$$H_0(x) = x \sum_{k=0}^{\infty} p_k [H_1(x)]^k = x G_0(H_1(x)). \tag{2.21}$$

If we know the generating functions $G_0(x)$ and $G_1(x)$, we can solve Eqn. (2.20) in order to get $H_1(x)$. At second step, we find $H_0(x)$ by substituting $H_1(x)$ and $G_0(x)$ into Eqn. (2.21). The value of s^{th} derivative of $H_0(x)$ at $x = 0$ gives the probability that an arbitrarily chosen vertex is included in a component with size s. However, it is almost impossible to find out exact solutions of $H_0(x)$ and $H_1(x)$ for most cases. Nevertheless, we can obtain the finite taylor expansion of $H_1(x)$ so that $H_1(x)$ becomes a polynomial with a degree of m by iteration Eqn. (2.20). We can summarize this method in such a way. Firstly, we choose any polynomial for $H_1(x)$. For instance, we get $H_1(x) = q_0 x$ where q_0 is the probability that a vertex following a randomly chosen edge has not another connection. Then substituting this polynomial into Eqn. (2.20), we obtain a new approximation for $H_1(x)$. We carry on producing a polynomial for $H_1(x)$ which has a higher degree. As we construct $H_1(x)$ so that its m coefficients are completely accurate, the coefficient of x^{m+1} will be also accurate in the new approximation we get in the next iteration. Continuing on our example, we get an approximation where the first $n+1$ coefficients are exactly accurate after n iteration. Finding a proper $H_1(x)$, we substitue it into Eqn. (2.21) and then obtain an approximate solution for $H_0(x)$. As an alternative, carrying out the preceding steps for various types of values of x close to zero, we can use these numerical results to evaluate the derivatives of $H_0(x)$ at $x = 0$. So, its s^{th} derivative gives P_s.

It is rather plausible to calculate first several hundred derivatives of $H_0(x)$ by means of many computer softwares. When the aforementioned method fails, we

also apply to some numerical differentiation techniques. However, calculations based on numerical differentiation have an inclination to problems related to machine precision. Therefore, it is recommended that the derivatives be calculated by numerical integration of the famous Cauchy formula [88]:

$$P_s = \frac{1}{s!}\frac{d^s H_0}{dz^s}\bigg|_{x=0} = \frac{1}{2\pi i}\oint_{\gamma}\frac{H_0(z)}{z^{s+1}}dz, \tag{2.22}$$

where $\gamma = \{z : |z| = 1\}$ is the circle. From Remark 2.2, we say that $H_0(z)$ is not singular on the region $R = \{z : |z| \leq 1\}$. Using Cauchy formula, we can calculate first several thousand derivatives of the generating function with a high accuracy.

Theorem 2.5. *When there is not any giant component, using Eqns. (2.20) and (2.21), we can represent the mean component size as [97, 99]*

$$\begin{aligned} \langle s\rangle &= H_0'(1) && (2.23)\\ &= [G_0(H_1(x)) + xG_0'(H_1(x))H_1'(x)]_{x=1} && (2.24)\\ &= 1 + G_0'(1)H_1'(1). && (2.25) \end{aligned}$$

From Eqn. (2.20) we have

$$H_1'(1) = 1 + G_1'(1)H_1'(1), \tag{2.26}$$

and thus

$$\langle s\rangle = 1 + \frac{G_0'(1)}{1 - G_1'(1)}. \tag{2.27}$$

By using

$$G_0'(1) = \langle k\rangle = z,$$

and

$$G_1'(1) = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} = \frac{z_2}{z_1}, \tag{2.28}$$

this expression can also be written as [97, 99]

$$\langle s\rangle = 1 + \frac{z_1^2}{z_1 - z_2} \tag{2.29}$$

where $z_1 = z$ is the average number of neighbors of a vertex and z_2 is the average number of second neighbors.

Eqn. (2.29) is indefinite at $z_1 = z_2$. This is the indication of the phase transition, which is the formalization stage of the giant component. This equation gives us an opinion about the position of the critical point and can be seen an alternative to Eqn. (2.19).

- **Above the Phase Transition**

Now, we turn to investigate above the phase transition and estimate the mean component size after excluding a giant component from the graph. Using generating function technique still provides the mean component size to be calculated even if there exists a giant component.

The small components are tree like graphs, but we cannot say the same thing for the giant component. For this reason, from the aforementioned definition, we conclude that $H_0(x)$ generates P_s for only the small components. Eqns. (2.20) and (2.21) are still correct for only the small components. So, we always rule out the giant component during the calculation. Since we discard the part of the graph including the giant component, $H_0(1)$ does not equal to 1 anymore and neither does $H_1(1)$. Our new expression for $H_0(1)$ will be

$$H_0(1) = \sum_{s=0}^{\infty} P_s = \text{fraction of vertices not in giant component.} \tag{2.30}$$

Namely, because of their tree-like topological structures, we can merely make assumptions on the sizes of the small clusters. Thus, sum of all the probabilities P_s is never equal to 1. Nonetheless, we can still find out the size of the giant component. Let S be the fractional representation of the size of the giant component so that [97, 99]

$$S = 1 - H_0(1). \tag{2.31}$$

Using Eqns. (2.20), (2.21) and (2.31), we get [97, 99]

$$v = G_1(v), \quad S = 1 - G_0(v), \quad \text{where } v = H_1(1). \tag{2.32}$$

Incidentally, since sum of all the possibilities P_s is not equivalent to 1, we need to normalize the distribution P_s ourselves. Moreover, we take the derivative of $H_0(x)$ using Eqn. (2.21). Then we get the mean component size [97, 99]

$$\begin{aligned} \langle s \rangle &= \frac{H_0'(1)}{H_0(1)} && (2.33) \\ &= \frac{1}{H_0(1)} \left[G_0(H_1(1)) + \frac{G_0'(H_1(1)) G_1(H_1(1))}{1 - G_1'(H_1(1))} \right] && (2.34) \\ &= 1 + \frac{z v^2}{[1 - S][1 - G_1'(v)]}, && (2.35) \end{aligned}$$

where S and v are the solutions of Eqn. (2.32). It can be readily seen that when the conditions $S = 0$ and $v = 1$ are satisfied, there is not any giant component. Since we are below the phase transition because of these conditions, Eqn. (2.35) turns out to be Eqn. (2.32).

2.1.2.5 Average Path Length

One of the characteristic aspects of networks is the mean shortest distance. By distance between any pair of vertices, we mean the number of connections we need for reaching to the another one starting from one of these vertices. Since there could be more than one path between a pair of vertices, we always choose the path which has the smallest number of connections.

If there is not any giant component, calculating the mean shortest distance will have no meaning. Most pairs of vertices chosen at random in the graph will not be reachable to any other in any way. Therefore, it will be reasonable to consider solely the case where there is a giant component. For this case, we will eliminate all the small components and focus on merely the mean shortest distance in the giant component. So, if there exists n vertices in the whole graph, it will be better to replace n by $N = nS$ where S is the fraction of the graph including the giant component.

Eqn. (2.18), i.e.,

$$z_m = \left[\frac{z_2}{z_1}\right]^{m-1} z_1,$$

gives the mean number of vertices which are m connections away from a randomly chosen vertex in the giant component. We can predict the mean shortest distance l using this formula. Since all the vertices in the giant component are connected somehow, a randomly chosen vertex and all its neighbours give us the total number of vertices in that component. In principle, we can write as [99]

$$1 + \sum_{m=1}^{l} z_m = N. \tag{2.36}$$

Using Eqn. (2.18), we get [99]

$$l = \frac{\ln[(N-1)(z_2 - z_1) + z_1^2] - \ln z_1^2}{\ln(z_2/z_1)}. \tag{2.37}$$

In the common case where $N \gg z_1$ and $z_2 \gg z_1$, this reduces to [99]

$$l = \frac{\ln(N/z_1)}{\ln(z_2/z_1)} + 1. \tag{2.38}$$

Eventhough there are some drawbacks, we cannot omit several notable aspects of Eqn. (2.38): [99].

1. Irrespective of degree distribution, any random graph has the average shortest distance which is *linearly proportional* to logarithm of the total size of

the giant component. In symbols, we can express this as $l \approx A + B\ln(N)$ where A and B are the constants.

2. We see that local features of a random graph are sufficient for us to evaluate its global feature which informs us about its topological structure. According to Eqn. (2.38), local properties are first and second mean degrees while the global property is the mean shortest distance.

3. Two random graphs which have distinct degree distributions but the same values of z_1 and z_2 have also the identical mean shortest distances.

2.1.2.6 Application of the Theory on Erdös Rényi Graph

We are interested in the standard Erdös Rényi random graph, with its Poisson degree distribution given by Eqn. (2.3) [97, 99]

$$p_k = \frac{z^k e^{-z}}{k!}.$$

- **Generating Functions for Erdös-Rènyi random graph:**

Applying Eqn. (2.4) to the poisson distribution, we get the generating function $G_0(x)$ for first neighbours of a vertex [97, 99]

$$\begin{aligned} G_0(x) &= e^{-z}\sum_{k=0}^{\infty}\frac{z^k}{k!}x^k && (2.39)\\ &= e^{z(x-1)}. && (2.40)\end{aligned}$$

The generating function $G_1(x)$ for vertices reached by following an edge is also easily found, from Eqn. (2.13)) [97, 99]:

$$\begin{aligned} G_1(x) &= \frac{G_0'(x)}{z} && (2.41)\\ &= e^{z(x-1)}. && (2.42)\end{aligned}$$

Thus, for the case of the Poisson distribution we have $G_1(x) = G_0(x)$ [97, 99]. This identity is the reason why the properties of the Erdös-Rènyi random graph are particularly simple to solve analytically [97, 99].

- **First and Second Neighbours:**

First degree of a vertex [97, 99]

$$z_1 = G_0'(1) = ze^{z(x-1)}\big|_{x=1} = z, \tag{2.43}$$

and second degree of a vertex [97, 99]

$$z_2 = G_0''(1) = z^2e^{z(x-1)}\big|_{x=1} = z^2. \tag{2.44}$$

- **Component Size:**

By Eqn. (2.32) and the fact $G_0(x) = G_1(x)$, we get $S = 1 - u$. Then

$$\begin{aligned} S &= 1 - G_0(u) && (2.45)\\ &= 1 - e^{z(u-1)} && (2.46)\\ &= 1 - e^{zS}. && (2.47) \end{aligned}$$

Using Eqn. (2.35), the average component size is given by [97, 99]

$$\langle s \rangle = \frac{1}{1 - z - S}. \tag{2.48}$$

- **Average Shortest Path:**

By Eqn. (2.38), average path length in the giant component is [20]

$$l = \frac{\ln(N/z)}{\ln(z^2/z)} + 1 = \frac{\ln N}{\ln z}. \tag{2.49}$$

2.2 Real Networks

Real networks differ in many respects from random graphs. Differences are mainly ascribed to fundamental features of real networks:

- Clustering [121],
- Mutuality [95],
- Degree distribution following power-law (with cut-off) [11].

In this section, we will introduce these concepts and two significant models, a small-world model [121] and a scale-free model.

2.2.1 Clustering (Transitivity)

In random graphs, the probability of an edge being present is same for all edges in these graphs. In real networks, however, the probability of any two neighbours (only first) of a vertex having a first degree connection is more than the probability of any randomly chosen vertices being first neighbours of each other.

As seen in Figure 2.6, first neighbours (vertices 2 and 4) of vertex 1 are also one-path connected. So, clustering brings a problem in calculating second and more degrees of a vertex. For example, in Figure 2.6, although vertices 2 and

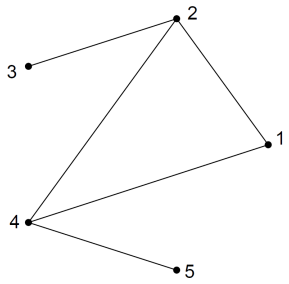


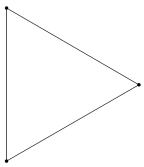
Figure 2.6: An illustration of clustering.

4 are actually first neighbours of vertex 1, the vertices are also taken as second neighbours of vertex 1 when it is calculated without regard to clustering. A simple calculation causes counting its second neighbours over the number it has, adding also first neighbours. Following that, this result affects correctness of the calculation for more degrees.

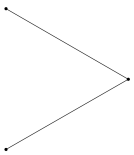
To solve the problem, Watts and Strogatz [121] defined the clustering coefficient (the transitivity ratio) as follows:

$$C = \frac{3 \times \text{ number of triangles in the whole graph}}{\text{number of connected triples of vertices in the graph}}, \tag{2.50}$$

where "connected triple" is an union of edges and vertices in which a single vertex is one-path connected to two others.



(a) A triangle



(b) A triple

Figure 2.7: A triangle has 3 triples. For a graph including only a triangle, $C = 1$.

Remark 2.3. The clustering coefficient for the Erdös-Rényi random graph is $C = p$ where p is the probability of edge being present.

2.2.2 Mutuality

Sometimes one-path connections of a vertex have a mutual first neighbour. When we rule out mutuality, calculation have a drawback so as to over-count the number of second neighbours.

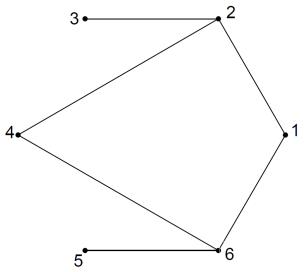


Figure 2.8: Mutuality is the quadrilateral here. Although, in fact, the number of second neighbours of vertex 1 is 3, on account of omitting the mutuality factor and counting the vertex 4 two times, this number ascends to 4.

To deal with this problem, the mutuality coefficient is expressed as [95]:

$$M = \frac{\text{mean number of vertices two steps away}}{\text{mean paths of length two to those vertices}} \tag{2.51}$$

and

$$M = \frac{\overline{k/[1 + C^2(k-1)]}}{\overline{k}}, \tag{2.52}$$

where k is the degree of a vertex.

While calculating the mutuality M using Eqn. (2.51), we must need to know the exact mean number of vertices two steps away from a vertex. This is evidently the mean number of second neighbours for which we want to find the mutuality coefficient. While searching for other options, the author tries to find the most common situations in which both clustering and mutuality occur, then catches the relation between them and gets Eqn. (2.52). That being the case, we can say that the formula does not aproximate all the uncommon situations in which clustering does not occur exactly even if there are not many.

2.2.3 The Second Degree of a Vertex in a Real Network

Eqn. (2.12) does not give a good result on calculating the number of second-nearest neighbours for any real network. Combining Eqns. (2.51) and (2.52) gives a better aproximation to mean second degree [95]:

$$z_2 = M(1 - C)(\overline{k^2} - \overline{k}), \tag{2.53}$$

where C is the clustering coefficient and M is the mutuality coefficient.

2.2.4 Small-World Networks (Watts-Strogatz Model)

A network which shows a *small-world effect* [121] is a network where the shortest distance l between two arbitrarly chosen vertices increases proportionally to the

logarithm of the number of vertices N in the network, that is:

$$l \approx \log(N). \tag{2.54}$$

Random graphs built according to the Erdös-Rènyi model show a small-world effect along with a small clustering coefficient (see Figure 2.9). Watts and Strogatz demonstrated that many real-world networks have in fact a small mean shortest path length, but also a clustering coefficient significantly higher than expected by random chance. Following that, they proposed a new graph model, named the *Watts-Strogatz model or a small-world network* which shows

- a small mean shortest path length (a small-world effect),
- a high clustering coefficient.

Construction of the small-world network is shown concisely in Figure 2.9 [121]. We have circular graph of n nodes in which each node is one-path connected to its k nearest neighbours by undirected edges. Then, we randomly select a node and the edge that connects it to its nearest neighbour in a clockwise direction. With probability p, we remove the edge to reconnect it to a vertex chosen arbitrarily over the entire circle. We repeat this step turning clockwise around the circle. Next, we take the edges that connect vertices to their second nearest neighbours clockwise in consideration. As before, we randomly reunite each of these edges with probability p and carry on circulating around the circle and applying this step for more remote neighbours. As there are $nk/2$ edges in the entire graph, this process is completed after looping $k/2$ times.

2.2.4.1 Six Degrees of Separation

Six degrees of separation is the theory that everyone in the world is six or fewer steps away so that a chain of a friend of a friend statements can be constructed so as to connect any two people in a maximum of six steps.

- Movie-Actor Collaboration (Six Degrees of Kevin Bacon):

The game "Six Degrees of Kevin Bacon" is known as a play based on the concept: its target is to link any actor to Kevin Bacon through six or less than connections, where two actors are connected if they have appeared in a movie together. The shortest distance between Kevin Bacon and any other actor is expressed as that actor's "Bacon number".

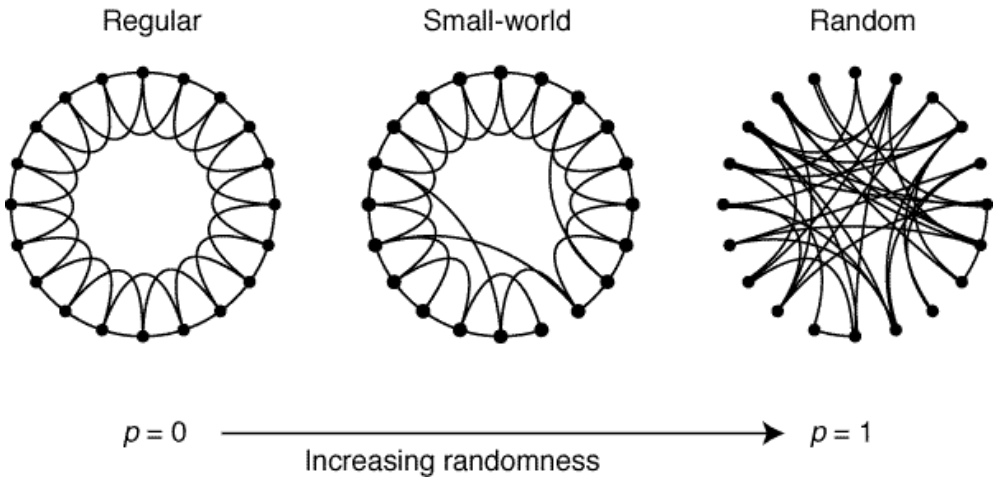


Figure 2.9: An implementation of random reuniting method which interpolates between a regular circle of nodes and a random network, without changing the number of vertices and edges in the graph [121]. For the regular circle, $n = 20$ and $k = 4$ [121]. Three stages of the process are shown here for different values of p. First circle is shown for $p = 0$. When p increases, the graph becomes increasingly disordered until for $p = 1$ where its all edges are reunited randomly [121]. We see that for intermediate values of p, the graph is a small-world network where clustering is dense as in a regular graph and has small average path length like a random graph [121].

2.2.5 Scale-Free Networks (Barabási-Albert Model)

Scale-free networks have a key feature: some vertices, called "hubs", have tremendous numbers of connections to other vertices, whereas most vertices have just a handful [13].

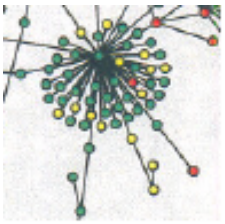


Figure 2.10: A small example of a hub. Hubs, however, can have hundred, thousand or even millions of connections. On the other hand, the vertices one-path connected to the hub have just a few neighbours. In that sense, the network appears to have no "scale" [13].

- Movie-Actor Collaboration:

The network of actors in Hollywood-popularized by the game Six Degrees of Kevin Bacon is scale-free in which it is dominated by hubs [13]. Specifically, although most actors have only a few ties to others, only a few actors, including Rod Steiger and Donald Pleasence, have thousands of connections [13]. By the way Bacon ranked just 876^{th} on a list of most connected actors [13].

- Robustness against accidental failures [13]:

In general, scale-free networks display an amazing robustness against accidental failures, a property that is rooted in their inhomogeneous topology [13]. The random removal of nodes will take out mainly the small ones because they are much more plentiful than hubs [13]. And the elimination of small nodes will not disrupt the network topology significantly, because they contain few links compared with the hubs, which connect to nearly everything (see Figure 2.11) [13].

On the other hand, a reliance on hubs has a serious drawback: vulnerability to attacks (see Figure 2.11) [13]. Recent research suggests that the simultaneous elimination of as few as 5% to 15% of all hubs can crash a system [13].

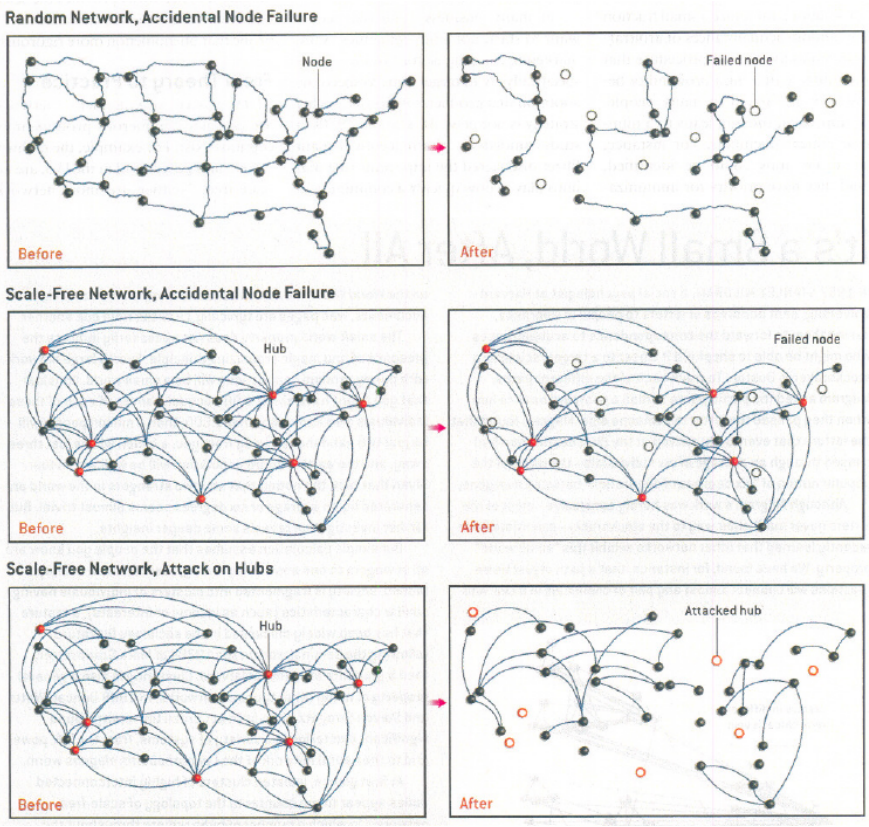


Figure 2.11: Removal of a node from random network, scale-free network and a hub from scale-free network, respectively [13].

- Preferential attachment (the rich get richer):

Scale-free network is generated by a process of "preferential attachment" [13]. When deciding where to establish a link, a new node prefers to attach to an existing node that already has many other connections [13]. These two basic

mechanisms - growth and preferential attachment - will eventually lead to the system's being dominated by hubs (see Figure 2.12) [13].

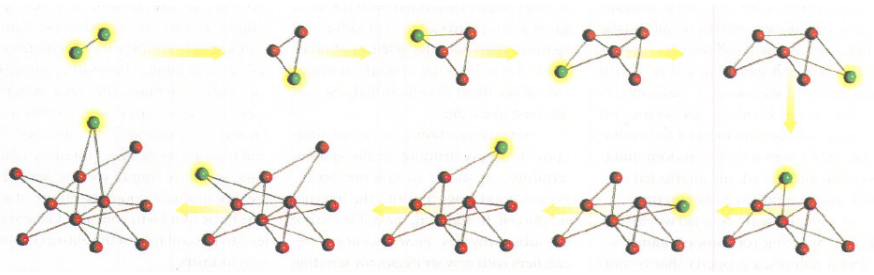


Figure 2.12: Construction of scale-free network [13]. (Green one is a new added node.)

In sociology, prefential attachment is known as "Matthew effect" or "the rich get richer and the poor get poorer".

This process occurs elsewhere. In Hollywood, it is more possible for the more connected actors to be chosen for new roles [13]. On the Internet the more connected websites which have greater users attract more new users [13]. Likewise, the most cited articles in the scientific literature stimulate more researchers to read and cite them [13].

- Degree distribution of scale-free networks:

In a random network, the nodes follow a Poisson distribution with a bell shape, and it is extremely rare to find nodes that have significantly more or fewer links than the average [13]. Random networks are also called exponential, because the probability that a node is connected to k other nodes decreases exponentially for large k [13].

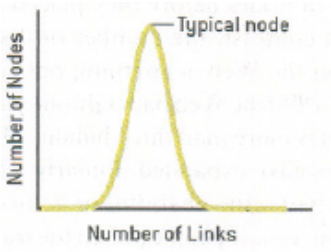


Figure 2.13: A typical bell-curve distribution (node linkages in a random network). All node degrees in the random graph are close to mean degree. Hubs are simply forbidden in random networks [13].

In contrast to the homogeneous distribution of links seen in random networks, "power laws" describe systems in which a few hubs dominate [13]. Two main behaviour they exhibit can be summarized as follows [13]:

1. A power law does not have a peak, as a bell curve does [13]. This explains the fact that there is not a definite limit for degree of any vertex [13].

2. Power law is described by a continuously decreasing function with a long tail [13]. This shows the existence of hubs nodes with very large degree, even if their number are small [13].

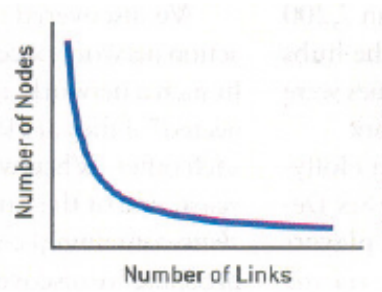


Figure 2.14: A power law distribution. From the graph, we can say that there are huge amount of nodes with small degree. Even if the number of nodes decreases as degree increases, this amount does not vanish until degree does not becomes enough large [13]. The nodes with enough large degree, for instance a hundred, are our hubs which make a network "scale-free" [13].

CHAPTER 3

STATISTICAL INFERENCE FOR COLLABORATION NETWORKS

In this chapter, we will endeavour to find the most suitable distribution functions for the degree sequences of our collaboration networks.

3.1 Use of Degree Distribution

A *degree sequence* is a monotonic nonincreasing sequence of first degrees of all the vertices in the network. The method based on the degree sequences provides global properties of very large networks to be obtained [99].

- *Canonical Ensemble Procedure:* After finding a proper degree distribution for a given degree sequence, many degree sequences are also generated by using that distribution. The elements of each degree sequence drawn from the specified distribution are independent identically distributed (i.i.d.) random integers. For each drawn degree sequence, the graph is chosen uniformly at random from the set of all possible graphs having that degree sequence. We get many sets of graphs with these degree sequences repeating the preceding step for other degree sequences drawn from the distribution. Then, all properties are averaged over the ensemble of graphs generated in this way [99].

- *Microcanonical Ensemble Procedure:* In the limit of large graph size an equivalent procedure is to study only one particular degree sequence and average all features uniformly over all graphs with that sequence, where the sequence is chosen to approximate as closely as possible the desired probability distribution [99].

3.2 Methods for Obtaining Degree Distribution

Many studies have shown that real networks which are also mentioned in Subsection 2.2.5 follow a power law distribution

$$p(k) = Ck^{-\alpha}, \tag{3.1}$$

where C is a normalized constant and α is a power law exponent.

3.2.1 Linear Least Squares Method

In the last decade of 20^{th} century, scientists focused on the problem of estimating α accurately for real networks from different disciplines mentioned in Chapter 1 by means of the least squares method [57].

The method fits the empirical data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to the function

$$y = a_0 + a_1 x + \ldots + a_m x^m, \tag{3.2}$$

where a_i $(i = 1, 2, \ldots, m)$ are the coefficients. The concept of "linearity" comes from the linearity of the variables (coefficients). Estimations of the coefficients, denoted by $\hat{a}_i$, can be easily found as we minimize the square of the residual

$$R^2 = \sum_{j=1}^{n} (y_j - (a_0 + a_1 x_j + \ldots + a_m x_j^m))^2. \tag{3.3}$$

In other words, refering to $\alpha = (a_1, a_2, \ldots, a_m)^T$

$$\hat{a} = \arg\max_{a} R^2. \tag{3.4}$$

There are some ways to represent empirical data [16].

1. The empirical probability distribution is given by

$$p(k) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = k), \tag{3.5}$$

where $I(\cdot)$ is a characteristic (indicator) function

$$I(x_i = k) = \begin{cases} 1, & x_i = k, \\ 0, & x_i \neq k. \end{cases} \tag{3.6}$$

Then,

$$p(k) \approx a_0 k^{-a_1} \tag{3.7}$$

can be replaced by

$$y \approx a_0 - a_1 x, \tag{3.8}$$

where

$$(x_i, y_i) = (\ln(p(k)),\ \ln(k)) \text{ for all possible k} \in \mathbb{N}. \tag{3.9}$$

In logarithmic scale, the slope of the line gives $\hat{a}_1$.

2. The empirical cumulative distribution is given by

$$P(k) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \geq k) \tag{3.10}$$

and, therefore,

$$(x_i, y_i) = (\ln(P(k)),\ \ln(k)) \text{ for all possible k} \in \mathbb{N}. \tag{3.11}$$

3. Logarithmic binning reduces the noise in the tail of the empirical distributions $p(k)$ and $P(k)$ by merging data points into groups [16]. By introducing the logarithmically scaled boundaries [16], we obtain

$$b_i = \text{round}(c^i) \text{ with some } c > 1. \tag{3.12}$$

A linear least squares fit is performed to

$$(x_i, y_i) = \left(\ln\frac{b_{i+1} + b_i - 1}{2}, \ln\sum_{k=b_i}^{b_{i+1}-1} \frac{p(k)}{b_{i+1} - b_i}\right) \tag{3.13}$$

or

$$(x_i, y_i) = \left(\ln\frac{b_{i+1} + b_i - 1}{2}, \ln\sum_{k=b_i}^{b_{i+1}-1} \frac{P(k)}{b_{i+1} - b_i}\right), \tag{3.14}$$

respectively.

3.2.1.1 Drawback of Linear Least Squares Fit

A distribution function always takes values in a range between 0 and 1 if it is normalized. However, using a linear least squares method, we have some problems with an adjustment of a normalization constant of a power law distribution so that all the probabilities do not take place in this range. Some methods are useful to make the regression line incorporate such constraints, but there are not remarkable extensions of these methods to power law distributions.

3.2.1.2 Normalized Form of Power-law

Normalized and continuous pdf is [28]

$$p(x) = (\alpha - 1)x^{-\alpha}x_{min}^{\alpha-1}. \tag{3.15}$$

Normalized and discrete pdf is [28]

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} \tag{3.16}$$

with Hurwitz-zeta function

$$\zeta(\alpha, x_{min}) = \sum_{j=0}^{\infty} \frac{1}{(x_{min} + j)^{\alpha}}. \tag{3.17}$$

For the details, see Appendix B.

3.2.2 Maximum Likelihood Estimation Method

For the i^{th} observation x_i, the conditional probability density function is $p_i(x_i|\alpha)$ [123]. Following that, the joint probability density function for an n-dimensional vector of observations x will be [123]

$$p(x|\alpha) = \prod_{i=1}^{n} p_i(x_i|\alpha). \tag{3.18}$$

We cannot easily compute the integral of $p(x|\alpha)$. Instead, we try to get *maximum likelihood estimation* (MLE) of α, denoted by $\hat{\alpha}$, by maximizing the likelihood function

$$L(\alpha|x) = p(x|\alpha) = \prod_{i=1}^{n} p_i(x_i|\alpha) \tag{3.19}$$

with respect to a.

By the way, the MLE for the continuous case is

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}, \tag{3.20}$$

where x_i $(i = 1, 2, \ldots, n)$ are the observed values of x such that $x_i \geq x_{min}$.

Morever, the MLE for the discrete case is [28, 51]

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min} - \frac{1}{2}} \right]^{-1}. \tag{3.21}$$

For the details, see Appendix B.

3.2.3 Kolmogorov Smirnov Test

Kolmogorov-Smirnov (KS) statistic is a nonparametric test which measures a distance between a theoretical cumulative distribution function (CDF) and an empirical CDF of the data. Given a sample $x_1, x_2, \ldots, x_n$ of i.i.d. random variables with distribution function F, consider the problem of testing $H_0 : F = F_0$ versus $H_1 : F \neq F_0$, where F_0 is some specified distribution. H_0 can be tested using KS statistic

$$d = \max_{x \geq x_{min}} |F(x) - F_{emp}(x)|, \tag{3.22}$$

where F and F_0 are theoretical and empirical CDFs of the sample, respectively.

3.2.4 Combination of MLE Method and KS Test

When using MLE method, we try to fit the data to a specified model with regardless of whether they all respect that model. In order to estimate the parameters of the model more accurately, we rule out some of the data which do not follow that distribution. In other words, we regard x_{min} as a parameter so that we eliminate all data below x_{min}. Because when choosing x_{min} too high we also throw away the data according with that model, predicting x_{min} take an important place in model selection.

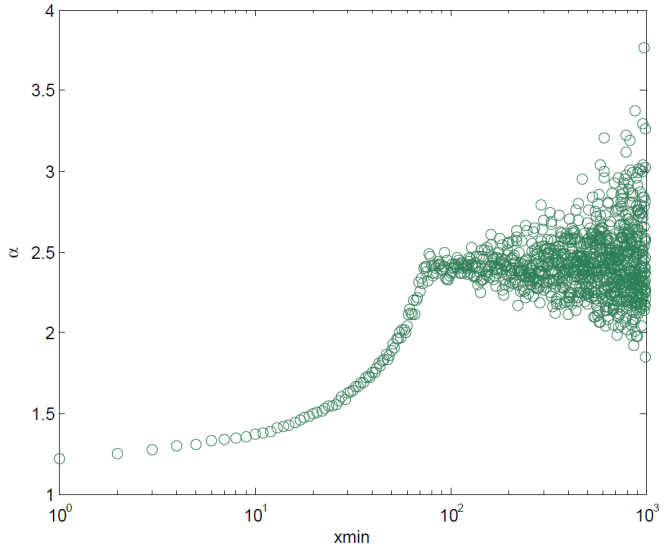


Figure 3.1: We have drawn 1000 different samples. Each sample have $n = 1150$ observations. 150 of the observations are generated by a uniform distribution and their values are less than $x_{min} = 75$. Remaining observations are power law distributed with $\alpha = 2.4$, $x_{min} = 75$. For each sample, α is searched for different values of x_{min} using MLE method. See Appendix C.2.

The significance of finding the true value for x_{min} is shown in Figure 3.1. When x_{min} is close to the exact value, MLE of the scaling-parameter $\hat{\alpha}$ is same with what is expected. Below the point $x_{min} = 75$, $\hat{\alpha}$ changes very rapidly since we also add more observations which are not power law distributed. Above that point, MLEs of $\hat{\alpha}$ assemble in a range including the true value and this range gets wider as chosen x_{min} move away from its true value, since we exclude some observed data which are also power law distributed. All things considered, chosing x_{min} accurately is critical. Some tried to deal with this problem by observing each data as done in Figure 3.1. This method is relative to only visualization and so, it is rather subjective. We need more concrete and thus convenient methods. Clauset, Shalizi and Newman have proposed a technique which consists of both MLE method and KS test.

The method can be summarized as follows:

1. Firstly, we compute KS statistic d for all the subsets of the data $x^* \geq x_{min}$ and $x_{min} = 1$ using MLEs of the parameters. Then we select $\hat{\alpha}$, MLE of the parameter α, which maximizes the KS statistic.

2. We find a set of $\hat{\alpha}$ values by implementing the preceding step for the different x_{min} values. After duplicating this step for all possible x_{min} values, we choose the x_{min} value for which the KS statistic is minimum and get the $\hat{\alpha}$ value which corresponds to that x_{min}.

Example 3.1. Let our sample be

$x = [\, 14,\ 2,\ 6,\ 7,\ 9,\ 10,\ 1,\ 2,\ 2,\ 3,\ 4,\ 25,\ 3,\ 6,\ 6,\ 7,\ 5,\ 11,\ 3,\ 14,\ 16,\ 1,\ 1,\ 9,\ 21,\ 15\ ]^T$.

For $x_{min} = 1$, distance vector will be

$d = [\, 0.00,\ 0.22,\ 0.24,\ 0.21,\ 0.23,\ 0.23,\ 0.14,\ 0.1136,\ 0.05,\ 0.02,\ 0.02,\ 0.04,\ 0.07,\ 0.08,\ 0.1\ ]^T$.

We select $\hat{\alpha}(1) = 1.5942$. The distance corresponding to that $\hat{\alpha}$ is $d_\alpha(1) = 0.2487$.

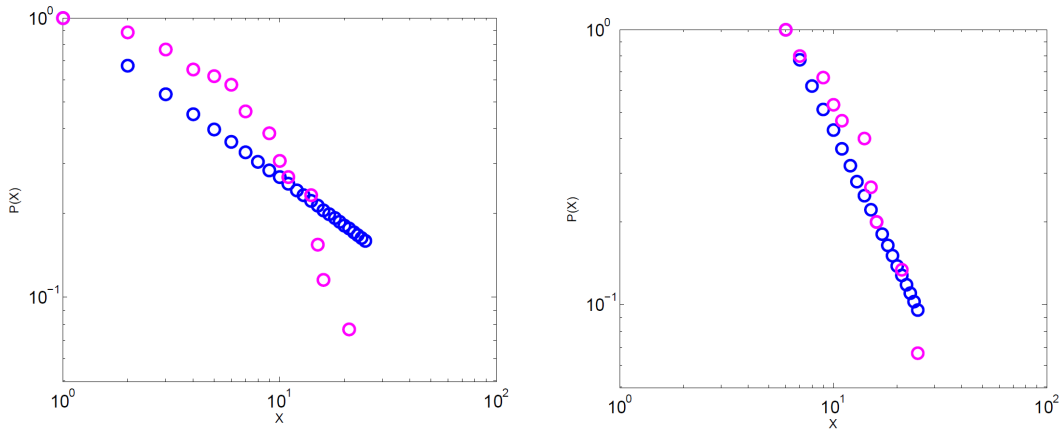
For $x_{min} = 2$, distance vector will be

$d = [\, 0.00,\ 0.00,\ 0.15,\ 0.17,\ 0.22,\ 0.24,\ 0.16,\ 0.14,\ 0.08,\ 0.06,\ 0.06,\ 0.01,\ 0.04,\ 0.05,\ 0.08\ ]^T$.

We select $\hat{\alpha}(2) = 1.8270$. The distance corresponding to that $\hat{\alpha}$ is $d_\alpha(2) = 0.2490$.

When finished, we get the vector whose elements are all possible x_{min}:

$y = [\, 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 9,\ 10,\ 11,\ 14,\ 15,\ 16,\ 21\ ]^T$.



(a) MLE power law fit.

(b) MLE+KS power law fit.

Figure 3.2: Empirical distribution and power law distribution of x are shown as pink and blue circles, respectively.

$\hat{\alpha} = [\ 1.5,\ 1.82,\ 2.01,\ 2.14,\ 2.42,\ 2.76,\ 2.80,\ 3.41,\ 3.42,\ 3.65,\ 6.05,\ 5.38,\ 5.17,\ 12.47\ ]^T$.

$d_\alpha = [\ 0.24,\ 0.24,\ 0.25,\ 0.25,\ 0.19,\ 0.17,\ 0.21,\ 0.25,\ 0.3,\ 0.33,\ 0.2,\ \ 0.27,\ 0.34,\ 0.36\ ]^T$.

Finally, we choose $\min d_\alpha = 0.1774$. So, this brings us $x_{min} = 6$ and $\hat{\alpha} = 2.7631$.

Remark 3.1. During the study, we will not make calculations for large x_{min} values since the situation that most of the data do not follow a power law distribution does not make any sense.

Furthermore, we will always prefer to use a log-log graph, which is a two-dimensional graph that uses logarithmic scales on both the horizontal and vertical axes since power law distributions appear as a straight line in a log-log graph as seen in Figure 3.2.

3.3 Other Distributions

Beside their proposed method, Clauset, Shalizi and Newman have also shown that some real networks in fact are not power law distributed. Namely, there can be alternative distributions that outstrip the power law. So, we will also consider the other possible distributions which are summarized in Table 3.1.

Type	Cont./Disc.	Non-normalized Pdf	Normalized coefficient of Pdf
Exponential	continuous	$e^{-\lambda x}$	$\lambda e^{-\lambda x_{min}}$
Exponential	discrete	$e^{-\lambda x}$	$(1-e^{-\lambda})e^{-\lambda x_{min}}$
Log-normal	continuous	$\frac{1}{x}\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\text{erfc}\left(\frac{\ln x_{min}-\mu}{\sqrt{2\sigma^2}}\right)\right]^{-1}$
Poisson	discrete	$\frac{\mu^x}{x!}$	$\left[e^{\mu}-\sum_{k=0}^{x_{min}-1}\frac{\mu^k}{k!}\right]^{-1}$
PL (cut-off)	continuous	$x^{-\alpha}e^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{min})}$
Weibull	continuous	$x^{\beta-1}e^{-\lambda x^{\beta}}$	$\beta\lambda e^{\lambda x_{min}^{\beta}}$

Table 3.1: Alternative distributions. See Appendix B.

3.4 Power Law Results on Our Data

We consider all papers in the archives of ArXiv until June 2013. The details about getting the data are represented in Appendix A. Before starting data analysis, we discard mathematicians whose degrees are zero. Namely, we do not put any mathematician who wrote his all papers solely by himself into our investigation.

In the study, the nodes (actors) are the researchers who study applied mathematics. Two actors are linked by a simple edge (tie), or two researchers are collaborated, if there exists at least one paper written together. Degree of an actor is the total number of researchers who the actor is in a collaboration with. In the light of all this information, the data used in the thesis are represented in Table 3.2.

Fields of Study	Number of Actors	Number of Ties	Min - Max Degree in the Social Network
Cryp. and Sec.	4267	7238	1 - 47
Financial Math.	2888	4303	1 - 40
Math. Bio.	16971	48570	1 - 115
Math. Phy.	17564	30326	1 - 60
Num. Analysis	3518	5150	1 - 32
Opt. and Cont.	5155	7800	1 - 42
Probability	8137	14774	1 - 99
Statistics	10551	20517	1 - 71

Table 3.2: Data used in our research.

As the degree increases, the number of the actors who have that degree decreases. For instance, according to our calculations, 2987 actors in the network for mathematical biology have degree 2 whereas there are only 2 actors with degree 90.

3.4.1 Power Law and Power Law with Cut-off

It is trivial that non-normalized form of power law is a special case of non-normalized form of power law with exponential cut-off

$$P(x) = x^{-\alpha} e^{-\lambda x}. \tag{3.23}$$

From that, we conclude that they are nested distibutions. Besides, both are heavy-tailed. There are three important subclasses of heavy-tailed distributions, the fat-tailed distributions, the long-tailed distributions and the subexponential distributions. Power law is long-tailed. That is to say, power law has a relatively large population which lies in its tails (right and/or left) when compared with normal distribution [76, 118]. Most of their population is far away the mean, causing the "skewness". Power law with cut-off is a fat-tailed distribution which also shows high skewness. It exhibits a power law decay in its tail, but it may

not follow a power law everywhere [9]. Figure 3.6c is a really good example for understanding the structure of the fat-tailed distribution. In our work, we always observe $0 \leq \lambda \leq 5 \times 10^{-2}$ so that this increases an effect of an exponential function in Eqn. (3.23). The effect, however, decreases while x increases. Thus, as x approaches to ∞, all power law distributions with cut-off must follow a power law even if they do not for any other x. By the way, log-normal is an example of a non-power law heavy tailed distribution, subexponential distribution.

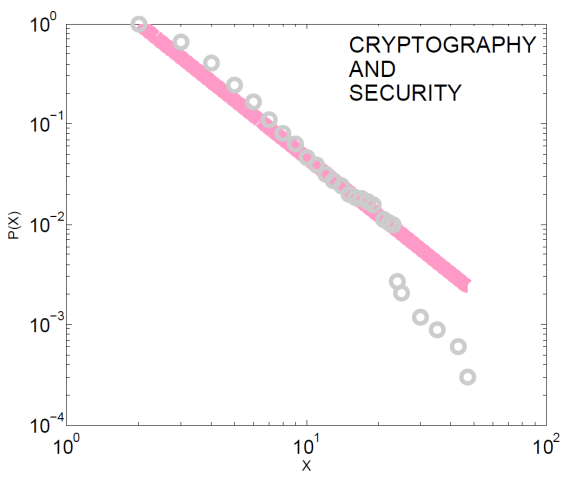
3.4.2 Power Law and Power Law with Exponential Cut-off Graphs for Our Data

Firstly, we draw continuous power law, discrete power law and power law with cut-off distributions for our data shown in Table 3.2. Because a (continuous or discrete) power law distribution function becomes a straight line on a log-log graph, we use a log-log scale while plotting. However, because of the reasons mentioned in Subsection 3.4.1, we see that some of the power law distributions with cut-off appear curved.

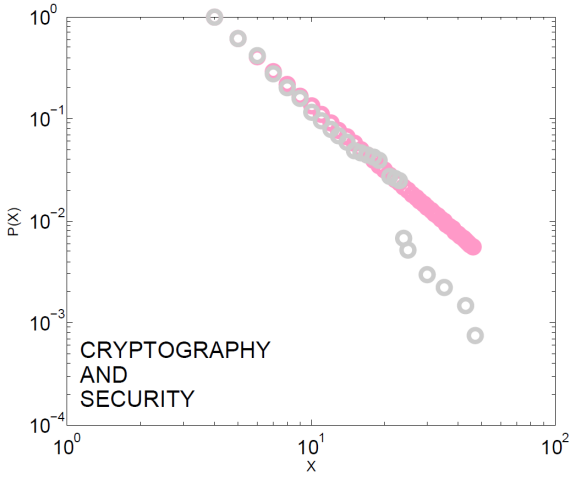
By the way, of course, each calculation gives different values of x_{min}. Because we do not want to put the non-power law data into fitting, we prefer not to intervene in the evaluations of x_{min}. However, as most of the researchers have small degrees, we exclude the majority in a few cases. For example, in power law distribution with exponential cut-off for financial mathematics, we find $x_{min} = 11$ so that we exclude 2118 people out of 2888. However, we observe $x_{min} = 2$ for continuous power law and $x_{min} = 3$ for discrete power law.

It is evident that most data do not fit to the power law well. In fact, most empirical data do not appear as a straight line; namely, almost all of them are curvilinear on a log-log scale. Nevertheless, it is really early to decide for us, since we have not compared them with the other distributions yet. Further, we use a likelihood ratio test to get more better results.

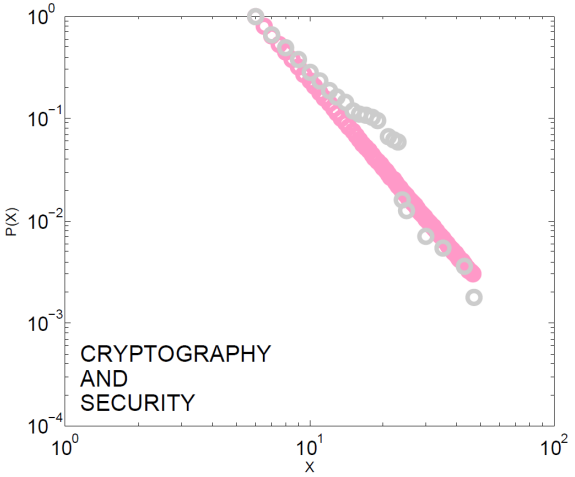
Figure 3.3: Power law data fits for Cryptography and Financial Mathematics. See Appendix C.1.



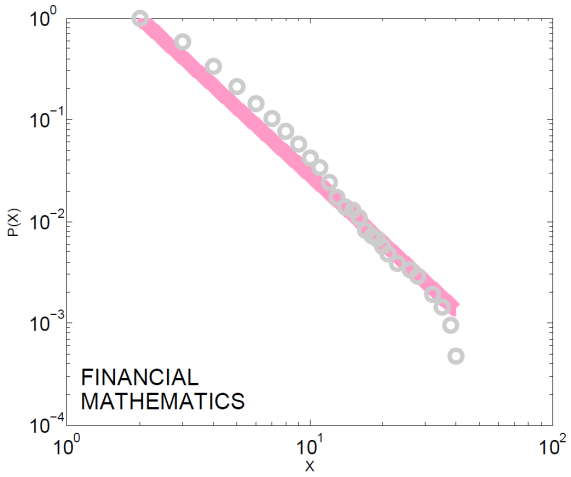
(a) $\alpha = 2.90,\ x_{min} = 2,\ d_\alpha = 0.18$



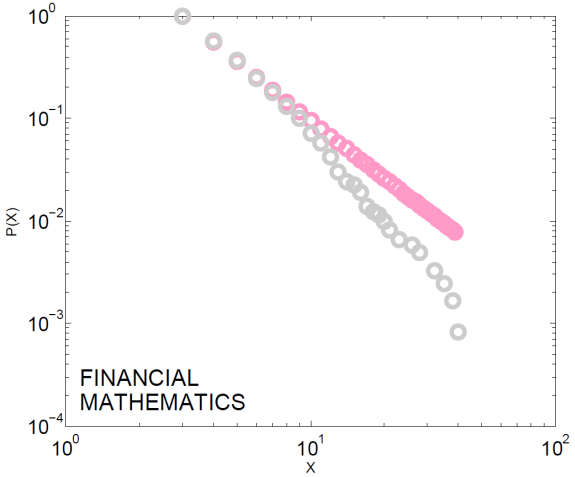
(b) $\alpha = 3.02,\ x_{min} = 4,\ d_\alpha = 0.17$



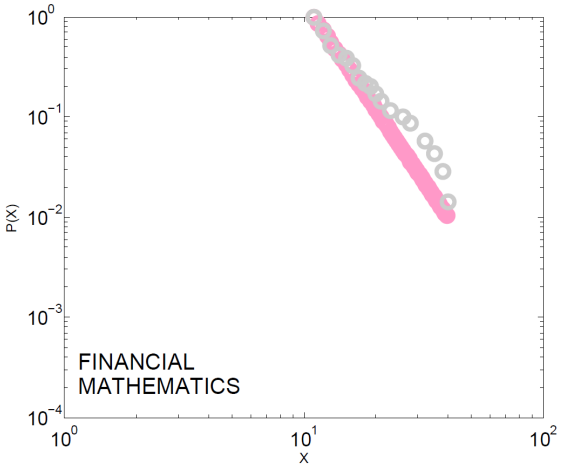
(c) $\alpha = 3.83,\ \lambda = 1.77 \times 10^{-7},\ x_{min} = 6,\ d_\alpha = 0.06$



(d) $\alpha = 3.20,\ x_{min} = 2,\ d_\alpha = 0.17$

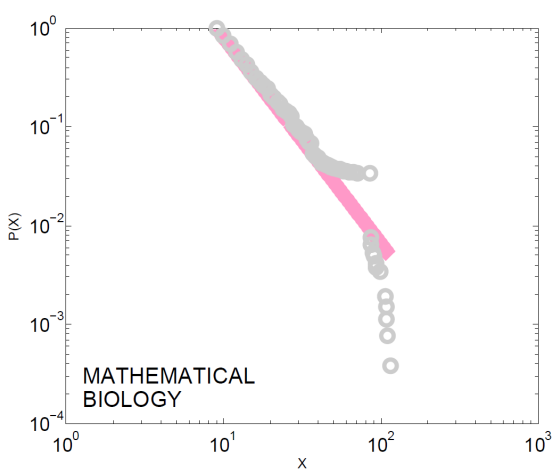


(e) $\alpha = 2.78,\ x_{min} = 3,\ d_\alpha = 0.02$

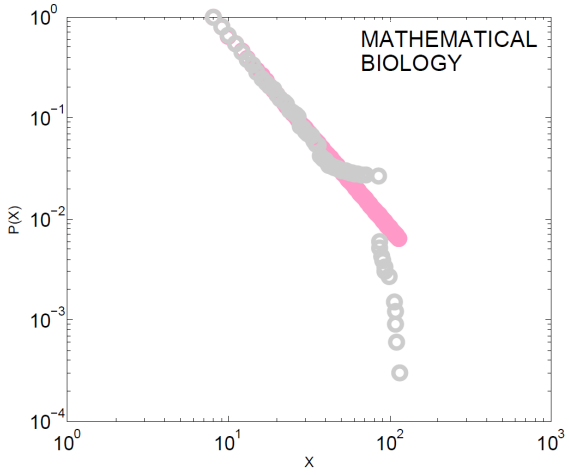


(f) $\alpha = 4.56, \lambda = 1.34 \times 10^{-8},\ x_{min} = 11,\ d_\alpha = 0.06$

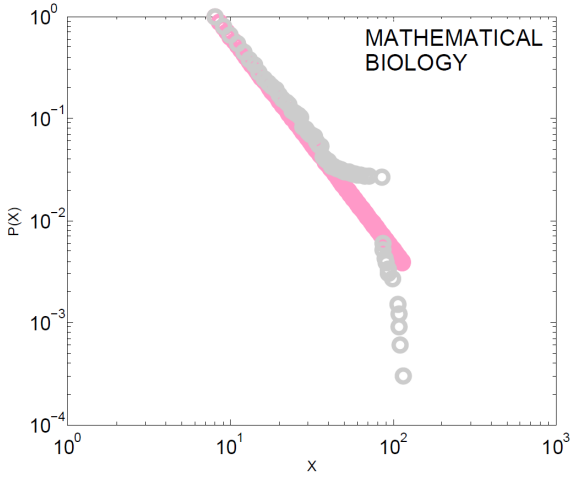
Figure 3.4: Power law data fits for Mathematical Biology and Physics.



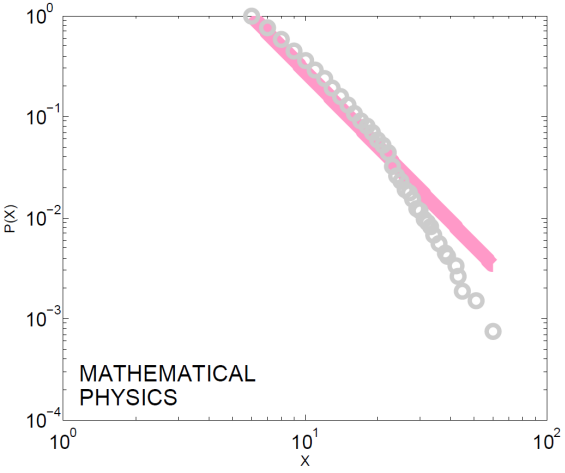
(a) $\alpha = 3.08,\ x_{min} = 9,\ d_\alpha = 0.03$



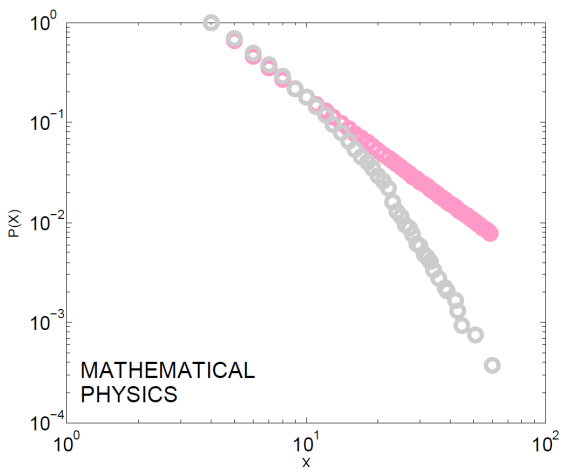
(b) $\alpha = 2.86,\ x_{min} = 8,\ d_\alpha = 0.01$



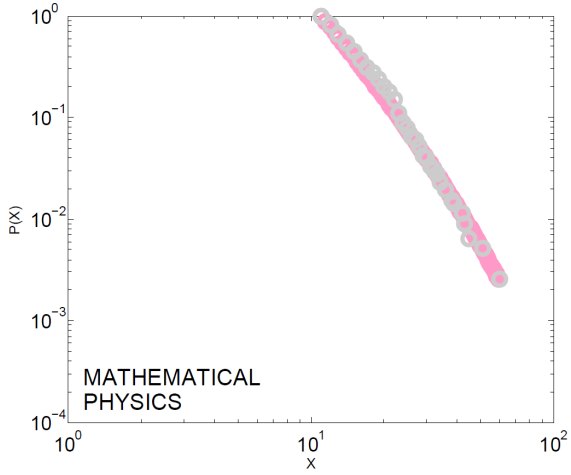
(c) $\alpha = 3.08,\ \lambda = 2.05 \times 10^{-8},\ x_{min} = 9,\ d_\alpha = 0.03$



(d) $\alpha = 3.46,\ x_{min} = 6,\ d_\alpha = 0.08$

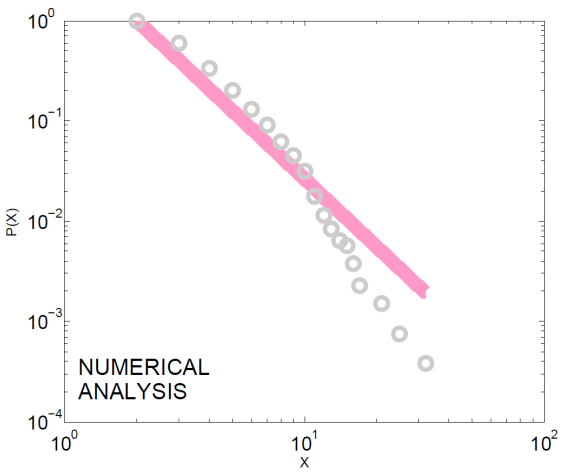


(e) $\alpha = 2.72,\ x_{min} = 4,\ d_\alpha = 0.03$

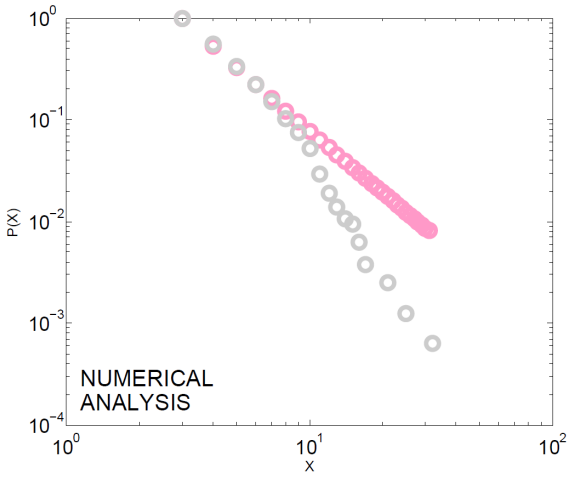


(f) $\alpha = 3.52,\ \lambda = 0.02,\ x_{min} = 11,\ d_\alpha = 0.05$

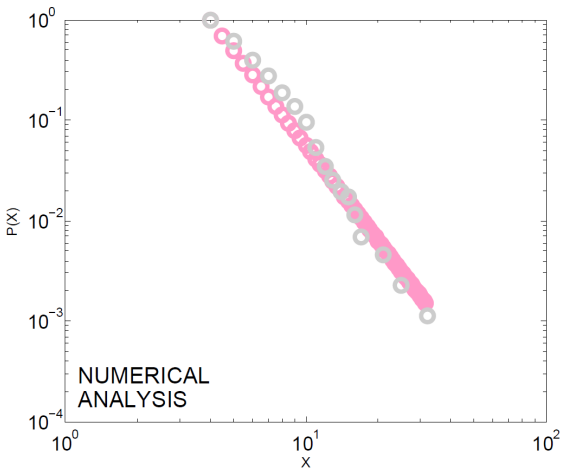
Figure 3.5: Power law data fits for Numerical Analysis and Optimization.



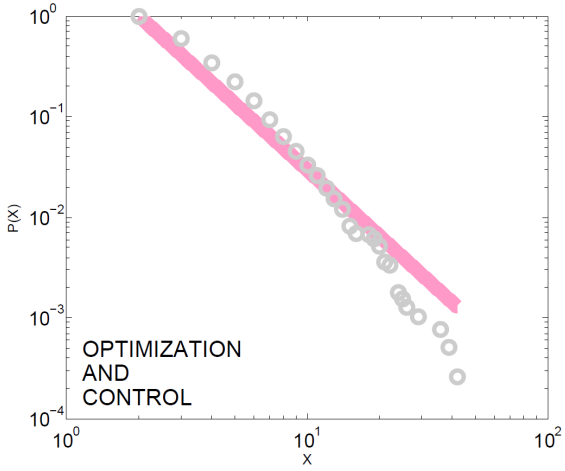
(a) $\alpha = 3.25,\ x_{min} = 2,\ d_\alpha = 0.2$



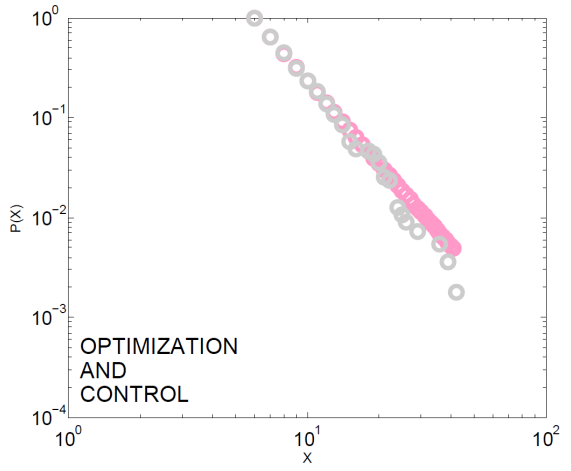
(b) $\alpha = 2.94,\ x_{min} = 3,\ d_\alpha = 0.03$



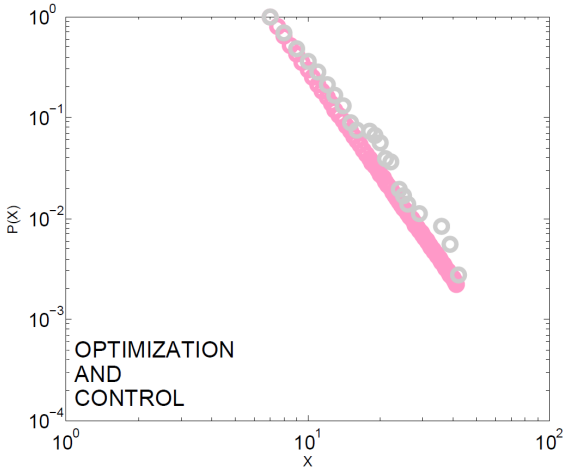
(c) $\alpha = 4.14,\ \lambda = 2.97 \times 10^{-8},\ x_{min} = 4,\ d_\alpha = 0.11$



(d) $\alpha = 3.18,\ x_{min} = 2,\ d_\alpha = 0.18$

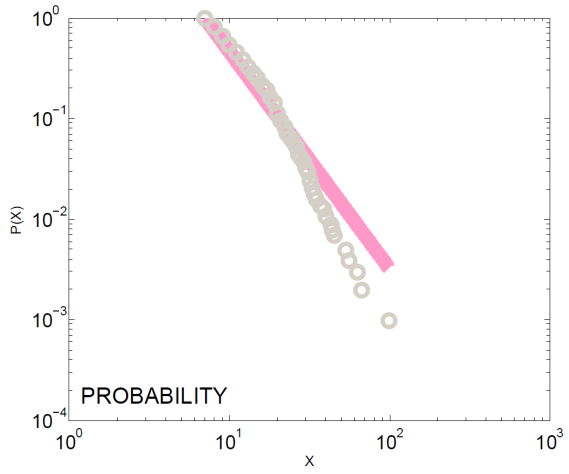


(e) $\alpha = 2.67,\ x_{min} = 6,\ d_\alpha = 0.01$

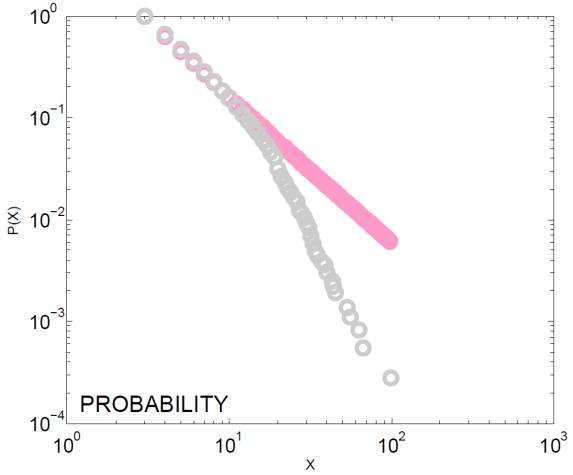


(f) $\alpha = 4.42,\ \lambda = 6.04 \times 10^{-8},\ x_{min} = 7,\ d_\alpha = 0.06$

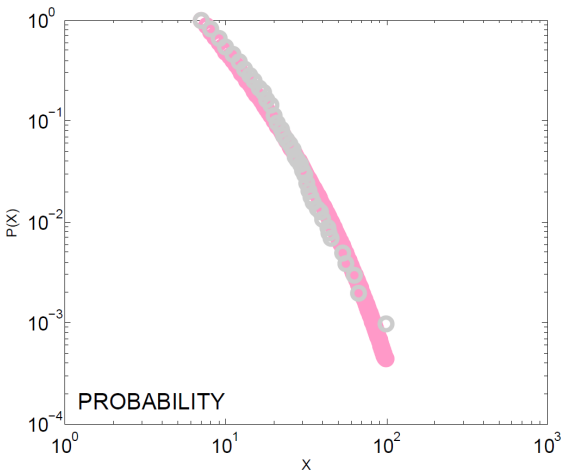
Figure 3.6: Power law data fits for Probability and Statistics.



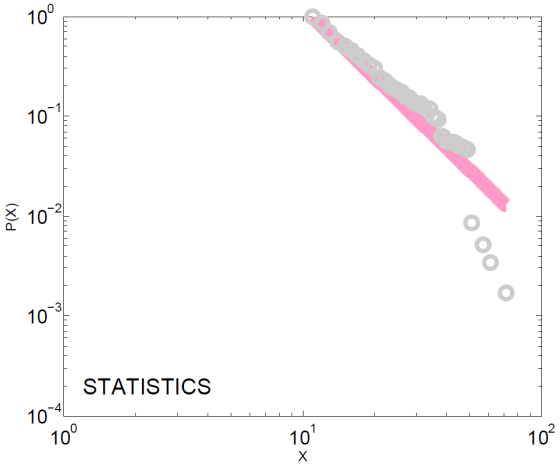
(a) $\alpha = 3.16,\ x_{min} = 7,\ d_\alpha = 0.07$



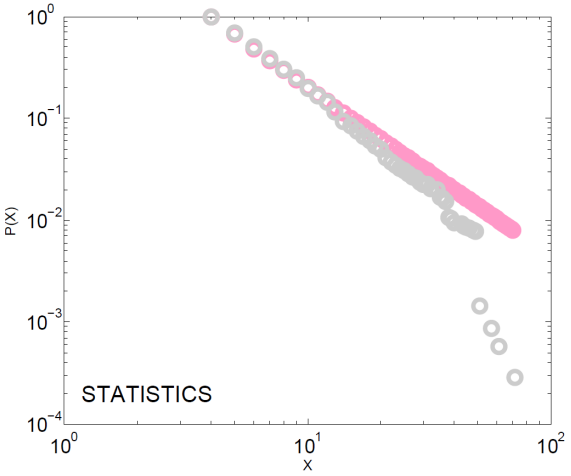
(b) $\alpha = 2.89,\ x_{min} = 3,\ d_\alpha = 0.02$



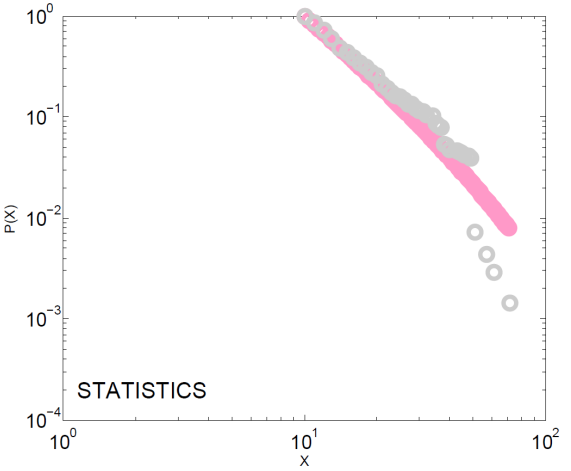
(c) $\alpha = 2.64,\ \lambda = 0.02,\ x_{min} = 7,\ d_\alpha = 0.05$



(d) $\alpha = 3.33,\ x_{min} = 11,\ d_\alpha = 0.05$



(e) $\alpha = 2.62,\ x_{min} = 4,\ d_\alpha = 0.02$



(f) $\alpha = 2.92,\ \lambda = 0.01,\ x_{min} = 10,\ d_\alpha = 0.04$

3.5 Likelihood Ratio Test

A likelihood ratio test is used to compare the fit of two candidate distributions [28]. The likelihoods of the data set within the two distributions with probability density function (PDF) $p_1(x)$ and $p_2(x)$ are [28]

$$L_1 = \prod_{i=1}^{n} p_1(x_i), \quad L_2 = \prod_{i=1}^{n} p_2(x_i) \tag{3.24}$$

and the ratio of the likelihoods is [28]

$$R = \frac{L_1}{L_2} \prod_{i=1}^{n} \frac{p_1(x_i)}{p_2(x_i)}. \tag{3.25}$$

Taking logs, the log-likelihood ratio is [28]

$$\mathcal{R} = \sum_{i=1}^{n} [\ln p_1(x_i) - \ln p_2(x_i)] = \sum_{i=1}^{n} [\xi_i^{(1)} - \xi_i^{(2)}] \tag{3.26}$$

where $\xi_i^{(j)} = \ln p_j(x_i)$ can be thought of as the log-likelihood for a single measurement x_i within distribution j.

But since, by hypothesis, the x_i are independent, so also are the differences $\xi_i^{(1)} - \xi_i^{(2)}$, and hence, by the central limit theorem, their sum $\mathcal{R}$ becomes normally distributed as n becomes large, with expected variance $n\sigma^2$ where σ^2 is the expected variance of a single term [28]. In practice we do not know the expected variance of a single term, but we can approximate it in the usual way by the variance of the data [28]:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\xi_i^{(1)} - \xi_i^{(2)} - (\bar{\xi^1} - \bar{\xi}^{(2)}) \right]^2, \tag{3.27}$$

with

$$\bar{\xi}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(1)}, \quad \bar{\xi}^{(2)} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(2)}. \tag{3.28}$$

Now suppose we are worried that the true expectation value of the log likelihood ratio is in fact zero, so that the observed sign of $\mathcal{R}$ is a product purely of the fluctuations and cannot be trusted as an indicator of which model is preferred [28]. The probability that the measured log likelihood ratio has a magnitude as large or larger than the observed value $|R|$ is given by [28]

$$p = \frac{1}{\sqrt{2\pi n\sigma^2}} \left[\int_{-\infty}^{-|\mathcal{R}|} e^{-t^2/2n\sigma^2} dt + \int_{|\mathcal{R}|}^{\infty} e^{-t^2/2n\sigma^2} dt \right], \tag{3.29}$$

where σ is given by Eqn. (3.28).

This p-value gives us an estimate of the probability that we measured a given value of $\mathcal{R}$ when the true value of $\mathcal{R}$ is close to zero (and hence is unreliable as a guide to which model is favored) [28]. If p is small (say $p < 0.1$) then our value for $\mathcal{R}$ is unlikely to be a chance result and hence its sign can probably be trusted as an indicator of which model is the better fit to the data [28]. (However, this does not mean that the model is a good fit, only that it is better than the alternative.)

3.5.1 Test for Nested Distributions

When the true distribution lies in the smaller family of distributions, the best fits to both families converge to the true distribution as n becomes large [28]. This means that the individual differences $\xi_i^{(1)} - \xi_i^{(2)}$ in Eqn. (3.27) each converge to zero, as does their variance σ^2 [28]. Consequently the ratio $|\mathcal{R}|/\sigma$ appearing in the expression for the p-value tends to $0/0$, and its distribution does not obey the simple central limit theorem argument given above [28]. Nonetheless, we find that adopts a chi-squared distribution as n becomes large,as a consequence of Wilk's Theorem $\mathcal{R}$ [28]. In other words, test statistic p is asymptotically χ^2 distributed.

χ^2 distribution is given by [123]

$$f_{\chi^2}(x) = \frac{x^{\frac{v}{2}-1}e^{-\frac{x}{2}}}{2^{v/2}\Gamma(v/2)}, \tag{3.30}$$

where x is the difference $\xi_i^{(1)} - \xi_i^{(2)}$ and $v = m - n$ is degrees of freedom in which m and n are independent parameter numbers of the candidate distributions. For our case, $v = 1$, since power law has one parameter α and power law with cut-off has two parameters α and λ. Then, p is given by [28]

$$p = \int_{-\infty}^{-|\mathcal{R}|} f_{\chi^2}(x)dx + \int_{|\mathcal{R}|}^{\infty} f_{\chi^2}(x)dx, \tag{3.31}$$

which is converted into

$$p = \int_{z}^{\infty} f_{\chi^2}(z)dz, \tag{3.32}$$

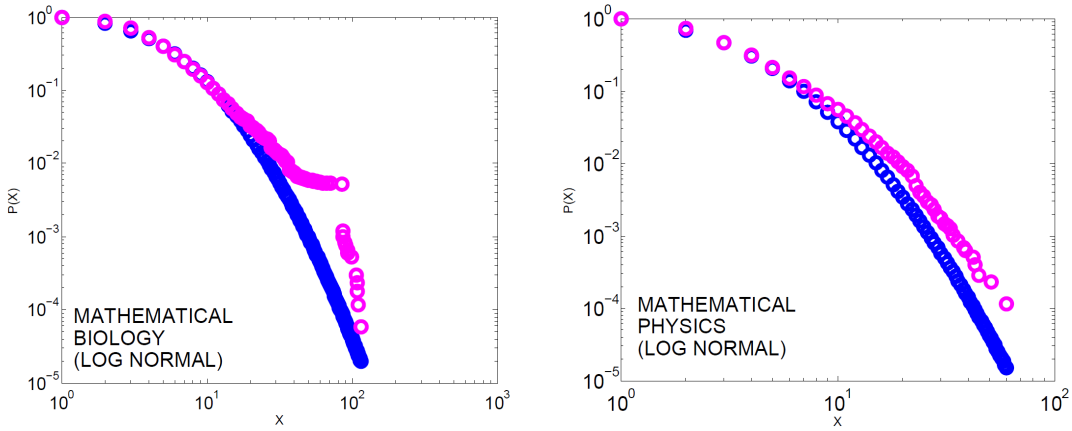
where $z = -2\mathcal{R}$.

3.6 Application of Likelihood Ratio Test on Our Data

In the test, we get null hypotheses the power law distribution with α and x_{min} calculated in Section 3.4. Because any value of a log-likelihood function is negative and the distribution which is favored over the another distribution has a larger likelihood, a negative log-likelihood ratio means that our alternative hypotheses fits better than the power law to the data.

In some cases because of the structure of the integrand, p values do not converge. For instance, chi squared distribution is in fact a special type of a gamma distribution and p can be readily evaluated using its upper gamma function form. However, when the lower bound of the integral is negative, the integrand diverges. For this reason, in such cases we prefer to take the null hypotheses as the power law with exponential cut-off (see Table 3.5).

We find that the log-normal distribution (see Figure 3.7) is favored over the power law distribution for all data whereas the poisson distribution is not a better choice for any data as seen in Tables 3.3 and 3.4. We also find that the power law with cut-off is favored over the ower law for the data except mathematical biology and statistics.



(a) $\mu = 1.3488,\ \sigma = 0.8236,\ x_{min} = 1$ (b) $\mu = 0.9158,\ \sigma = 0.7569,\ x_{min} = 1$

Figure 3.7: Log normal data fit.

Nonetheless, these results do not tell us that the log-normal is a good model for our data. We can only say that the log-normal fits better to the data compared to the power law. That is to say, the log-normal gives better results for x_{min} which is the most proper lower bound for power law distribution according to MLE+KS technique. In addition, we observe that the log-normal fits to all data for $x_{min} = 1$ or $x_{min} = 2$, when applying the MLE+KS method. In fact, when $x_{min} = 1$, MLE+KS method turns out to be only MLE method so that for the data sets except cryptography and security, numerical analysis and optimization and control, the parameters are

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln x_i}{n},\ and \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (\ln x_i - \hat{\mu})^2}{n}}, \tag{3.33}$$

where x_i $(i = 1, \ldots, n;\ x_1 = 1)$ are the observations on the data set (See Appendix B). To sum up, since we rule out the power law on acount of the reasons listed above, it is not very suprising to say that the networks built by the data are not exactly "scale-free".

Fields of Study	C. Exp.	p	D. Exp.	p	Log Normal	p	Poisson	p	Weibull	p
Cryp. and Sec.	481.43	0.00	1194.6	0.00	-655.63	0.00	6750 ($x_{min} = 1$)	0.00	-0.0004 ($x_{min} = 15$)	0.94
Financial Math.	399.36	0.00	-2152	0.00	-258.26	0.00	4323.8 ($x_{min} = 1$)	0.00	-0.0982 ($x_{min} = 13$)	0.11
Math. Bio.	503.48	0.00	37252	0.00	-428.10	0.00	56331 ($x_{min} = 1$)	0.00	-0.0195 ($x_{min} = 15$)	0.12
Math. Phy.	176.76	0.00	-2140.7	0.00	-659.33	0.00	12478	0.00	-0.1098	0.31
Num. Analysis	318.73	0.00	1027.4	0.00	-474.52	0.00	4498 ($x_{min} = 1$)	0.00	-226.8 ($x_{min} = 1$)	0.00
Opt. and Cont.	570.84	0.00	1554.3	0.00	-639.42	0.00	7141.4 ($x_{min} = 1$)	0.00	-1.4353 ($x_{min} = 15$)	0.00
Probability	53.446	0.00	142.74	0.00	-290.64	0.00	10052	0.00	-2.9807	0.00
Statistics	41.938	0.00	21646	0.00	-140.08	0.00	21397 ($x_{min} = 1$)	0.00	-0.0036	0.39

Table 3.3: Comparison of the continuous power law and the other distributions.

Fields of Study	C. Exp.	p	D. Exp.	p	Log Normal	p	Poisson	p	Weibull	p
Cryp. and Sec.	-157.42	0.00	3743.8	0.00	-532.8	0.00	4860.7 ($x_{min} = 1$)	0.00	-29.663 ($x_{min} = 9$)	0.00
Financial Math.	-217.79	0.00	-1573	0.00	-562.45	0.00	2898.4 ($x_{min} = 1$)	0.00	-3.921 ($x_{min} = 13$)	0.00
Math. Bio.	282.14	0.00	35251	0.00	-903.59	0.00	50654 ($x_{min} = 1$)	0.00	-60.494 ($x_{min} = 15$)	0.00
Math. Phy.	-707.86	0.00	-8227.5	0.00	-2384.4	0.00	18811	0.00	-6.0112($x_{min} = 24$)	0.00
Num. Analysis	-421.2	0.00	-1712.7	0.00	-818.92	0.00	2798.6 ($x_{min} = 1$)	0.00	-1926.2 ($x_{min} = 1$)	0.00
Opt. and Cont.	-73.527	0.00	1472.5	0.00	-209.06	0.00	4685.9 ($x_{min} = 1$)	0.00	-4.4391 ($x_{min} = 15$)	0.00
Probability	-302.17	0.00	-3796.2	0.00	-1504.3	0.00	4496.5	0.00	-150.04 ($x_{min} = 7$)	0.00
Statistics	-137.37	0.00	15075	0.00	-1315.8	0.00	16906 ($x_{min} = 1$)	0.00	-58.994 ($x_{min} = 11$)	0.00

Table 3.4: Comparison of the discrete power law and the other distributions.

Fields of Study	Null Hypotheses	Alternative Hypotheses	LLR	p	Null Hypotheses	Alternative Hypotheses	LLR	p
Cryp. and Sec.	C. PL	PL (cut-off)	-4092	0.00	D. PL	PL (cut-off)	-1342.8	0.00
Financial Math.	C. PL	PL (cut-off)	-2674.8	0.00	D. PL	PL (cut-off)	-2095.7	0.00
Math. Bio.	PL (cut-off)	C. PL	-0.0005	0.96	D. PL	PL (cut-off)	-2000.3	0.00
Math. Phy.	C. PL	PL (cut-off)	-4121.6	0.00	D. PL	PL (cut-off)	-10208	0.00
Num. Analysis	C. PL	PL (cut-off)	-2145.6	0.00	D. PL	PL (cut-off)	-1395.4	0.00
Opt. and Cont.	C. PL	PL (cut-off)	-4626.9	0.00	D. PL	PL (cut-off)	-442.23	0.00
Probability	C. PL	PL (cut-off)	-2572.5	0.00	D. PL	PL (cut-off)	-8128.5	0.00
Statistics	PL (cut-off)	C. PL	-278.1	0.00	D. PL	PL (cut-off)	-6292.8	0.00

Table 3.5: Comparison of the power law and the power law with cut-off.

CHAPTER 4

STATISTICAL NETWORK ANALYSIS

In this chapter, we investigate the statistical properties of our collaboration networks.

4.1 The Effect of Clustering and Mutuality on the Mean Number of Collaborators

As mentioned in Subsection 2.2.1, the likelihood of any two coauthors of a researcher being in collaboration with each other is greater than the possibility of any arbitrarily selected researchers coworking on an article. Moreover, any two coworkers of a researcher can have a common coworker who has not studied with the researcher. These two properties separate real networks from any random graph. Without regarding these important properties of real networks, applying to a theoretical approach for only random graphs $(*)$ will give inaccurate results.

On the other hand, computing all degrees of a vertex directly is not easy. When the number of observations n increases, calculating them becomes more compelling and takes considerably more time. Thus, we can use the advantages of the Eqns. (2.50), (2.52), and (2.53). In order to assess the effectiveness of these equations, we compare the results of direct computations with the theoretical ones. Firstly, we compute the total number of triangles and the total number of connected triples in each network (see Appendix C.4 and C.5). Secondly, we get their clustering and mutuality coefficients using Eqns. (2.50), and (2.52). Then, we evaluate average second degree z_2 using the formulas [95]

$$z_2 = \bar{k^2} - \bar{k}, \qquad (*)$$
$$z_2 = (1 - C)(\bar{k^2} - \bar{k}), \qquad (**)$$
$$z_2 = M(1 - C)(\bar{k^2} - \bar{k}), \qquad (***)$$

where k symbolizes first degree of a vertex so that $\bar{k} = z_1$. Formula $(*)$ is constructed for random graphs. For real graphs, we discard the part of the graph where clustering occurs so that we get Eqn. $(**)$. Looking at Eqn. $(**)$, we see that when C goes up, z_2 and higher degrees go down. Due to squares in

the network, a result based on Eqn. $(**)$ will be still an overcount estimation by a factor $1/M$. To deal with this problem, we use Eqn. $(***)$. We compare our theoretical results with actual ones computed with Matlab (see Table 4.2). In addition to this, you can find actual z_1, z_2 and z_3 values in Table 4.1. For matlab codes, see Appendix C.4. According to Table 4.2, when the effect of clustering is not considered, the formula $(*)$ fails to approach the exact value z_2. Because while calculating we count the mathematicians' some coworkers (first neighbours) as the coworkers of their coworkers (second neighbours) owing to clustering, we find the results calculated by $(*)$ relatively large. When clustering is taken into account $(**)$, the results become considerably close to z_2. We admit that they are not perfect, but neither are they ignorable. Moreover, carrying on the observation, we have seen that some approaches computed by $(***)$ to z_2 remain insufficient to give moderately correct results, whereas some are really good. Table 4.1 represents the z_1, z_2 and z_3 values for each network. Looking at these values and the fraction $z_1/z_2 > 1$, we can say that there exists a phase transition for each network (see Subsubsection 2.1.2.4). As seen before, theoretical consequences, however, do not always overlap the computation ones. In the following section, we will investigate an existence of a giant component for each network and calculate the other properties regarding their component sizes.

Fields of Study	actual z_1	actual z_2	actual z_3	z_2/z_1
Cryp. and Sec.	3.3925	4.5451	4.7358	1.3397
Financial Math.	2.9799	6.8912	13.8718	2.3125
Math. Bio.	5.7239	16.3493	46.2676	2.8563
Math. Phy.	3.4531	13.1126	50.1751	3.7973
Num. Analysis	2.9277	3.8681	4.7083	1.3212
Opt. and Cont.	3.0261	6.5493	14.4058	2.1642
Probability	3.6313	20.3205	103.7001	5.5959
Statistics	3.8892	13.8633	55.1299	3.5660

Table 4.1: First three mean degree in our social networks.

Fields of Study	Triangles	Triples	C	M
Cryp. and Sec.	9194	39561	0.6972	0.4200
Fin. Math.	2865	20270	0.4255	0.6472
Math. Bio.	197074	774125	0.7637	0.2445
Math. Phy.	19427	188828	0.3086	0.7052
Num. Analysis	3679	19038	0.5797	0.5473
Opt. and Cont.	4837	33472	0.4335	0.6453
Probability	8824	118915	0.2226	0.7761
Statistics	31484	178559	0.5289	0.4684

actual z_2	$(*)$	$(**)$	$(***)$
4.5451	18.5428	5.6148	2.3582
6.8912	14.0374	8.0645	5.2193
16.3493	91.2292	21.5574	5.2708
13.1126	21.5017	14.8663	10.4837
3.8681	10.8232	4.5490	2.4897
6.5493	12.9862	7.3597	4.7473
20.3205	29.2282	22.7220	17.6345
13.8633	33.8468	15.9452	7.4687

Table 4.2: We have calculated the number of triangles in the network, the number of connected triples in the network, the clustering coefficient, the mutuality coefficient, actual mean second degree z_2, mean second degree of a random graph $(*)$, mean second degree of a random graph by taking the clustering effect into consideration $(**)$, mean second degree of random graph by taking the clustering and mutuality effects into consideration $(***)$, respectively.

4.2 Groups of Connected Mathematicians

In spite of easy access to information, person-to-person contact is still significant so as to contribute to academic issues being discussed in more detail and well-developed. Moreover, some researchers may not be interested in a subject until their coworkers make the subject attractive for them. Therefore, we can say that the larger connected groups, the more spread and share of scientific knowledge.

It is not suprising that all the mathematicians in the networks are not tied to each other. In each network, there are more than 500 distinct groups of connected mathematicians according to our calculations (for matlab codes, see Appendix C.6). With respect to the results presented in Table 4.3, there are not giant components for the networks of cryptography and numerical analysis. We see that there are only 46 people between first two largest components of the network for cryptography and security, while this difference ascends to 322 for numerical analysis. We conclude from these results that it is not sufficient for the proportion to be greater than 1. In addition, it must be greater than a certain amount for networks to have definitely a phase transition. However when we look at the others, their largest components fill almost the majority of the volume of the networks. That is to say, their 2^{nd} largest components remain very small when compared with the 1^{st} ones.

We have calculated the mean component size of every network without excluding the giant component. When comparing it to the mean component size of only the small groups, we have found it relatively large (see Table 4.4). Thus, in order to get reasonable results, it would be better if we exclude the giant one from each network. We have also computed the values of z_1 and z_2 for each network in which the giant component is removed. Then, we have obtained the mean component size of small groups theoretically using Eqn. (2.29). For mathematical physics and probability, the results are compatible with the actual ones.

By the way, for cryptography, removing the largest component in the network, we have found that z_2 is still greater than z_1. Thus, we have shown the theoretical value which is marked by the symbol $\star$ in Table 4.4. Since we cannot mention a certain phase transition for cryptography and numerical analysis, it will not be suprising that they have the greatest mean component sizes of small groups.

As a result, in the field of probability, communication is highest, and so prevalence of any scientific information in that community is faster compared to the others. On the other hand, in the field of cryptography and security, the connections between the mathematicians seems to be lowest.

Fields of Study	Largest C. Size	2^{nd} Largest C. Size	3^{rd} Largest C. Size	4^{th} Largest C. Size	z_2/z_1
Cryp. and Sec.	312 (7 %)	266	55	47	1.33
Financial Math.	1179 (40 %)	17	16	14	2.31
Math. Bio.	9415 (55 %)	42	37	31	2.85
Math. Phy.	11414 (64 %)	24	20	16	3.79
Num. Analysis	714 (20 %)	392	112	65	1.32
Opt. and Cont.	2125 (41 %)	55	21	20	2.16
Probability	6033 (74 %)	22	14	13	5.59
Statistics	6428 (60 %)	32	29	21	3.56

Table 4.3: We represent the first four largest components in each network. We mean the number of mathematicians tied to each other by component size.

Fields of Study	Actual Mean C. Size (including a giant one)	Actual Mean C. Size (only small groups)	Theoretical Mean C. Size (only small groups)
Cryp. and Sec.	4.86	4.51	$\star$
Financial Math.	5.69	3.37	5.69
Math. Bio.	10.36	4.62	9.76
Math. Phy.	9.07	3.17	3.43
Num. Analysis	5.82	4.76	27.8
Opt. and Cont.	6.14	3.61	6.76
Probability	11.34	2.93	4.02
Statistics	8.46	3.31	5.23

Table 4.4: We represent the mean component sizes directly calculated for all the components, for only the small groups, and theoretically calculated for only small groups, respectively.

4.3 Mean Shortest Distance between the Mathematicians and Closeness Centrality

In this section, we search the minimum number of connections (ties) between any two mathematicians in the largest connected group of each network.

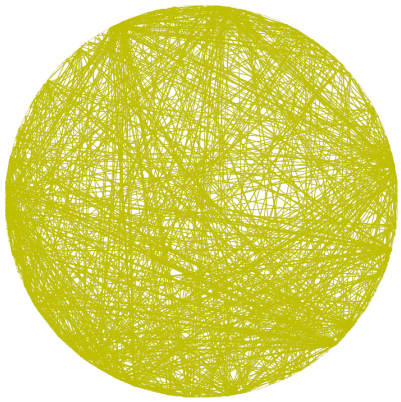


Figure 4.1: The largest connected group of the network for cryptography and security. This group consists of 312 mathematicians (see Table 4.3) and 880 ties. Despite its small size, the network is seen to be complex here. (For its matlab code, see Appendix C.3)

During the assessment, because the network for cryptography and security has not a sufficiently large group of mathematicians, we think that calculating mean distance in its largest group does not make sense and thus rule out it (Figure 4.1). In fact, this idea is also valid for numerical analysis. Nonetheless, in order to see the effects of the sizes, the mean first degrees and the mean second degrees of the networks on the mean distances, we prefer to share the results about the biggest group of people studying numerical analysis on Table 4.5. As seen from the findings represented in Table 4.5, expected value of shortest path between two arbitrarily chosen researchers is inverse proportional to z_2/z_1. We see that the group with the smallest size and the smallest z_2/z_1, which is the numerical analysis group, has the biggest mean distance and diameter. On the other hand, the network which has the largest component with the biggest value z_2/z_1, the probability group, has also the smallest mean distance and diameter. From Table 4.5 we say that in the field of probability, the maximum distance between any two researchers can be only a number up to 17. In the light of these results, we say that all networks except cryptography and numerical analysis exhibit a "small-world effect". As the groups grow in size, the expected values of shortest distance in these groups do not rise in the same proportion as the size.

Closeness centrality of an actor gives how close the actor is to other actors in a social network by quantifying the sum of shortest distance between the actor and all other actors in the network. Low centrality means first several degrees of that actor are high while its other remaining degrees are relatively lower. A researcher who has the lowest centrality is thought to be the first to learn new information, and information arising from him will come to other mathematicians faster, compared to the information originating with other mathematicians. To illustrate, we have calculated cloness centrality of each node in the networks for financial

mathematics and optimization. In financial mathematics, the lowest centrality is 7571 whereas the highest centrality is 17480. Likewise, in optimization and control, the lowest centrality is 10861 whereas 30243 is highest.

With a little adjustment in the algorithm for the average distance, we can find the centrality of the mathematians in these groups. However, the algorithms written by me have a storage problem and thus they are time-taking. So, beside my codes in Appendix C.7, you can find an alternative algorithm which is a modified form of the standart breadth-first search [92].

Fields of Study	Diameter	Mean Distance	Theoretical Mean Distance	z_2/z_1
Fin. Math.	23	9.6996	5.0601	3.7468
Math. Bio.	22	7.3839 *	6.0992	3.8549
Math. Phy.	20	7.4619 *	5.9345	4.7447
Num. Analysis	34	12.6976	6.2436	2.4471
Opt. and Cont.	20	8.4555	5.6798	3.5914
Probability	17	6.4068	4.8215	6.3899
Statistics	20	7.0491	5.5020	4.7195

Table 4.5: We demonstrate diameter (maximum shortest distance), mean shortest distance, theoretically calculated (Eqn. 2.37) mean shortest distance, z_2/z_1 where z_1 and z_2 are mean first degree and mean second degree for the largest component of each network, respectively. By the way, because their sizes are large, while calculating actual mean distances, we sample them over 1000 random people for mathematical biology and mathematical physics.

CHAPTER 5

CONCLUSIONS and OUTLOOK TO FURTHER WORK

5.1 Conclusions

Throughout the study, we have focused on two problems regarding the networks of the study fieds which can be counted as eight main fields of applied mathematics.

Until a few years ago scientists tried to find the well-fitting model for a given degree sequence of any real network using the least squares method and MLE method. However, by these methods, we attempt to it all the data to a model even if some of the data do not in fact comply with that model. To overcome this problem, we use a current statistical technique by which we discard the data that do not follow the distribution we search for. When applying the more robust and confidential method, we find that any network we study in the thesis is not scale-free exactly. From a technical perspective, this means that most of the numbers of collaborators assemble in vicinity of the mean number of co-workers for each network. Namely, even if the number of mathematicians decreases as the number of their collaborations become more, we observe that there are many researchers with co-workers at almost all numbers. Because in scale-free networks, connections of most actors are handful while remaining ones may have thousand or even million neighbours, the collaboration networks being researched in the thesis differ from any scale-free network.

Secondly, we investigate the other attributes of the networks. According to the inquiry, the network for mathematical biology has the biggest average degree which means that one to one communication is highest whereas it is lowest for numerical analysis (see Table 4.1). Then we compute the indispensable facts of real-networks, clustering and mutuality and use the updated version of the formula for z_2 where clustering and mutuality effects are introduced into the random graphs. According to the results, the network mathematical biology has the highest clustering coefficient whereas the network for probability has the lowest. Furthermore, we find that theoretical calculations of average second degrees are considerably close to actual ones (see Table 4.2). Another interesting outcome is that although $z_2/z_1 > 1$ for all the networks, we cannot claim an existence of a giant group for the cryptography and numerical analysis networks (see Table 4.3). We see that there are only 46 people between first two largest

connected groups of the network of cryptography, while this difference ascends to 322 for numerical analysis. We can say that theory does not always match up with the empirical results. So, it is not sufficient for the proportion to be greater than 1. In addition to this, it must be greater than a certain amount for networks to have definitely a phase transition which guarantees the existence of a giant connected group of the mathematicians. The largest connected group of the network for probability comprises of 74 percent of its all population. Hence, the connections between the mathematicians studying on probability are more dense compared with the others. After this stage, we evaluate the actual average group sizes for all the assembles, and only the small assembles taking place in each network separately. So, we comprehend the serious effect of the giant groups on the average group size. In order to get more robust consequences, it is better ruling out the giant groups during the assessment of the mean group sizes (see Table 4.4). Next, we compute the mean group sizes in theory for solely small groups. Measuring theoretical findings against their exact values, we see that they are imminent for mathematical physics and statistics. In addition, we predict the maximum value of the shortest distance and the average shortest distance between the mathematicians in the largest group for each network. Following that, we learn that the assemble with the smallest size and z_2/z_1, which is the numerical analysis group, has the biggest diameter and the average distance (see Table 4.5). Based on this, all the networks comprising of the mathematicians from the fields except cryptography and numerical analysis are small-world networks.

5.2 Outlook to Further Work

It is feasible to find how many papers have been written by a pair of the mathematicians mentioned in the thesis. Collecting the information, we change the collaboration networks into weighted ones. In weighted graphs, each edge of a graph has an associated quantitative value, a non-negative integer, called weight. In our next problem, the *weight* represents the number of papers written together by two certain mathematicians. Using the concept "weight", we predict the strength of collaborative ties. Moreover, we can find the distance of the least total weight from a researcher to each of the other researchers.

Beside scientific collaboration networks, in our future research, we can use our findings to make an analysis of the other types of networks such as gene regularity networks, social media analysis, target marketing strategies, behaviour analysis based on telecommunication data (cell phones) and belief networks.

REFERENCES

[1] L. A. Adamic, *The small world web, in Lecture Notes in Computer Science*, Springer, New York, 1999.

[2] L. A. Adamic and E. Adar, Friends and neighbors on the web, Social Networks, 25(3), pp. 211–230, 2003.

[3] W. Aiello, F. Chung, and L. Lu, A random graph model for massive graphs, in Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, pp. 171-180, Association of Computing Machinery, New York, 2000.

[4] W. Aiello, F. Chung, and L. Lu, Random evolution of massive graphs, in J. Abello, P. M. Pardalos, and M. G. C. Resende (eds.), Handbook of Massive Data Sets, pp. 97-122, Kluwer, Dordrecht, 2002.

[5] R. Albert and A.-L. Barabasi, Statistical mechanics of complex networks, Rev. Mod. Phys., 74, pp. 47–97, 2002.

[6] R. Albert, H. Jeong, and A. Barabasi, Diameter of the world-wide web, Nature, 401, pp. 130–131, 1999.

[7] L. A. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, Classes of small-world networks, Science, 97(21), 2000.

[8] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, Classes of small-world networks, Proc. Natl. Acad. Sci. USA, 97, pp. 11149–11152, 2000.

[9] D. Bahat, A. Rabinovitch, and V. Frid, *Tensile Fracturing in Rocks Tectonofractographic and Electromagnetic Radiation Methods*, Springer, 2005.

[10] D. Baird and R. E. Ulanowicz, The seasonal dynamics of the Chesapeake Bay ecosystem, Ecological Monographs, 59, pp. 329–364, 1989.

[11] A.-L. Barabasi and R. Albert, Emergence of scaling in random networks, Science, 286, pp. 509–512, 1999.

[12] A.-L. Barabasi, R. Albert, and H. Jeong, Scale-free characteristics of random networks: The topology of the World Wide Web, Physica A, 281, pp. 69–77, 2000.

[13] A.-L. Barabasi and E. Bonabeau, Scale-free networks, Scientific American, 50, 2003.

[14] A.-L. Barabasi, H. Jeong, E. Ravasz, Z. Neda, A. Schuberts, and T. Vicsek, Evolution of the social network of scientific collaborations, Physica A, 311, pp. 590–614, 2002.

[15] V. Batagelj and A. Mrvar, Some analyses of Erdos collaboration graph, Social Networks, 22, pp. 173–186, 2000.

[16] H. Bauke, Parameter estimation for power-law distributions by maximum likelihood methods, Preprint, Theoretical Physics, 2007.

[17] P. S. Bearman, J. Moody, and K. Stovel, Chains of affection: The structure of adolescent romantic and sexual networks, Preprint, Department of Sociology, Columbia University, 2002.

[18] S. Belen, E. Kropat, and G.-W. Weber, On the classical maki-thompson rumour model in continuous time, Central European Journal of Operations Research, 19(1), pp. 1–17, 2011.

[19] B. Bollobas, Degree sequences of random graphs, Discrete Mathematics, 33(1), pp. 1–19, 1981.

[20] B. Bollobas, *Random Graphs*, Academic Press, New York, 1985.

[21] M. Bordens and I. Gomez, Collaboration networks in science, Information Today, Medford, NJ, 2000.

[22] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Graph structure in the web, Computer Networks, 33, pp. 309–320, 2000.

[23] A. Broida and K. C. Claffy, Internet topology: Connectivity of ip graphs, in S. Fahmy and K. Park (eds.), Scalability and Traffic Control in IP Networks, no. 4526 in Proc. SPIE, pp. 172-187, International Society for Optical Engineering, Bellingham, WA, 2001.

[24] J. Camacho, R. Guimera, and L. A. N. Amaral, Robust patterns in food web structure, Phys. Rev. Lett., 88, 2002.

[25] R. D. Castro and J. Grossman, Some famous people with finite Erdos numbers, The Erdos Number Project Website, 2013.

[26] Q. Chen, H. Chang, R. Govindan, S. J. Jamin, S.and Shenker, and W. Willinger, The origin of power laws in internet topologies revisited, in Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE Computer Society, 2002.

[27] G. Chowell, J. M. Hyman, and S. Eubank, Analysis of a real world network: The City of Portland, Technical Report BU-1604-M, Department of Biological Statistics and Computational Biology, Cornell University, 2002.

[28] A. Clauset, C. R. Shalizi, and M. E. J. Newman, Power-law distributions in empirical data, SIAM Review, 51(4), pp. 661–703, 2009.

[29] J. E. Cohen, F. Briand, and C. M. Newman, Community food webs: Data and theory, Springer, New York, 1990.

[30] S. R. Corman, M. Kuhn, T., R. D., and K. J. Dooley, Studying complex discursive systems: Centering resonance analysis of organizational communication, Human Communication Research, 28(4), pp. 157–206, 2002.

[31] A. Davis, B. B. Gardner, and M. R. Gardner, Deep south, University of Chicago Press, Chicago, 1941.

[32] R. de Castro and J. W. Grossman, Famous trails to Paul Erdos, Mathematical Intelligencer, 21, pp. 51–63, 1999.

[33] O. Defterli, A. Fugenschuh, and G.-W. Weber, Regression analysis for clusters in gene-environment networks based on ellipsoidal calculus and optimization, Dynamics of Continuous, Discrete and Impulsive Systems Series B, 17(5), pp. 639–657, 2010.

[34] O. Defterli, A. Fugenschuh, and G.-W. Weber, Modern tools for the time discrete dynamics and optimization of gene-environment networks, Communications in Nonlinear Science and Numerical Simulation (CNSNS), 16(12), pp. 4768–4779, 2011.

[35] P. S. Dodds and D. H. Rothman, Geometry of river networks, Phys. Rev. E, 63, 2001.

[36] J. A. Dunne, R. J. Williams, and N. D. Martinez, Food-web structure and network theory: The role of connectance and size, Proc. Natl. Acad. Sci. USA, 99, pp. 12917–12922, 2002.

[37] J. A. Dunne, R. J. Williams, and N. D. Martinez, Network structure and biodiversity loss in food webs: Robustness increases with connectance, Ecology Letters, 5, pp. 558–567, 2002.

[38] L. Egghe and R. Rousseau, *Introduction to Informetrics*, Elsevier, Amsterdamo, 1990.

[39] P. Erdos and A. Renyi, On random graphs, I, Publicationes Mathematicae (Debrecen), 6, pp. 290–297, 1959.

[40] P. Erdos and A. Renyi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci, 5, pp. 17–61, 1960.

[41] P. Erdos and A. Renyi, On the strength of connectedness of a random graph, Acta Mathematica Academiae Scientiarum Hungarica, 12(1-2), pp. 261–267, 1961.

[42] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the internet topology, Computer Communications Review, 29, pp. 251–262, 1999.

[43] T. J. Fararo and M. Sunshine, A study of a biased friendship network, Syracuse University Press, Syracuse, 1964.

[44] D. A. Fell and A. Wagner, The small world of metabolism, Nature Biotechnology, 18, pp. 1121–1122, 2000.

[45] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of web communities, IEEE Computer, 35, pp. 66–71, 2002.

[46] J. Galaskiewicz, Social organization of an urban grants economy, Academic Press, New York, 1985.

[47] J. Galaskiewicz and P. V. Marsden, Interorganizational resource networks: Formal patterns of overlap, Social Science Research, 7, 1978.

[48] J. Gebert, M. Latsch, E. M. P. Quek, and G.-W. Weber, Analyzing and optimizing genetic network structure via path-finding, Journal of Computational Technologies, 9, pp. 3–12, 2004.

[49] J. Gebert, N. Radde, and G.-W. Weber, Modelling gene regulatory networks with piecewise linear differential equations, in the special issue (feature cluster), Challenges of Continuous Optimization in Theory and Applications of European Journal of Operational Research, 181(3), pp. 1148–1165, 2007.

[50] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer, 2001.

[51] M. Goldstein, S.A.Morris, and G. Yen, Problems with fitting to the power-law distribution, The European Physical Journal B, 41(2), 2004.

[52] L. Goldwasser and J. Roughgarden, Construction and analysis of a large Caribbean food web, Ecology, 74, pp. 1216–1233, 1993.

[53] J. Grossman, The Erdos number project data files, The Erdos Number Project Website, 2010.

[54] J. Grossman, Publications of Paul Erdos, The Erdos Number Project Website, 2011.

[55] J. W. Grossman and P. D. F. Ion, On a portion of the well-known collaboration graph, Congressus Numerantium, 108, pp. 129–131, 1995.

[56] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, Self-similar community structure in organisations, Preprint cond-mat/0211498, 2002.

[57] M. T. Heath, *Scientific Computing, An Introductory Survey*, McGraw-Hill, New York, 2d edition, 2002.

[58] M. Huxham, S. Beaney, and D. Raffaelli, Do parasites reduce the chances of triangulation in a real food web?, Oikos, 76, pp. 284–300, 1996.

[59] R. F. i Cancho and R. V. Sole, The small world of human language, Proc. R. Soc. Lond. B, 268, pp. 2261–2265, 2001.

[60] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. A. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl. Acad. Sci. USA, 98, pp. 4569–4574, 2001.

[61] H. Jeong, S. Mason, A.-L. Barabasi, and Z. N. Oltvai, Lethality and centrality in protein networks, Nature, 411, pp. 41–42, 2001.

[62] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, The large-scale organization of metabolic networks, Nature, 407, pp. 651–654, 2000.

[63] J. H. Jones and M. S. Handcock, An assessment of preferential attachment as a mechanism for human sexual network formation, Preprint, University of Washington, 2003.

[64] P. Jordano, J. Bascompte, and J. M. Olesen, Invariant properties in coevolutionary networks of plant-animal interactions, Ecology Letters, 6, pp. 69–81, 2003.

[65] V. K. Kalapala, V. Sanwalani, and C. Moore, The structure of the United States road network, Preprint, University of New Mexico, 2003.

[66] S. A. Kauffman, Metabolic stability and epigenesis in randomly connected nets, J. Theor. Bio., 22, pp. 437–467, 1969.

[67] S. A. Kauffman, Gene regulation networks: A theory for their structure and global behavior, in A. Moscana and A. Monroy (eds.), Current Topics in Developmental Biology, Academic Press, New York, 6, pp. 142–182, 1971.

[68] S. A. Kauffman, The origins of order, Oxford University Press, Oxford, 1993.

[69] H. Kautz, B. Selman, and M. Shah, Referralweb: Combining social networks and collaborative filtering, Comm. ACM, 40, pp. 63–65, 1997.

[70] J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, *The Web as a graph: Measurements, models and methods*, volume 1627, in Proceedings of the International Conference on Combinatorics and Computing, no. 1627 in Lecture Notes in Computer Science,Springer, Berlin, 1999.

[71] A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow, Social networks and infectious disease: The colorado springs study, Soc. Sci. Med., 38, 1994.

[72] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. S. Tomkins, and E. Upfal, Stochastic models for the web graph, in Proceedings of the 42st Annual IEEE Symposium on the Foundations of Computer Science, IEEE Computer, pp. 57–65, 2000.

[73] V. Latora and M. Marchiori, Is the Boston subway a small-world network?, Physica A, 314, pp. 109–113, 2002.

[74] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graphs over time: Densification laws, shrinking diameters and possible explanations, KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 177–187, 2005.

[75] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 2007.

[76] D. M. Levine, D. Stephan, T. C. Krehbiel, and M. L. Berenson, *Statistics for Managers using Microsoft Excel*, Prentice Hall, 2002.

[77] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, The web of human sexual contacts, Nature, 411, 2001.

[78] P. Mariolis, Interlocking directorates and control of corporations: The theory of bank control, Social Science Quarterly, 56, pp. 425–439, 1975.

[79] A. Maritan, A. Rinaldo, R. Rigon, A. Giacometti, and I. Rodriguez-Iturbe, Scaling laws for river networks, Phys. Rev. E, 53, pp. 1510–1515, 1996.

[80] N. D. Martinez, Artifacts or attributes? Effects of resolution on the Little Rock Lake food web, Ecological Monographs, 61, pp. 367–392, 1991.

[81] S. Maslov and K. Sneppen, Specificity and stability in topology of protein networks, Science, 296, pp. 910–913, 2002.

[82] J. J. McAuley and J. Leskovec, Learning to discover social circles in ego networks, NIPS, 2012.

[83] G. Melin and O. Persson, Studying research collaboration using co-authorships, Scientometrics, 36, pp. 363–373, 1996.

[84] M. S. Mizruchi, The american corporate network, Sage, Beverley Hills, pp. 1904–1974, 1982.

[85] M. Molloy and B. Reed, The size of the giant component of a random graph with a given degree sequence, Combinatorics, Probability and Computing, pp. 35–68, 1998.

[86] J. M. Montoya and R. V. Sole, Small world patterns in food webs, J. Theor. Bio., 214, pp. 405–412, 2002.

[87] J. Moody, The structure of a social science collaboration network, Preprint, Department of Sociology, Ohio State University, 2003.

[88] C. Moore and M. E. J. Newman, Exact solution of site and bond percolation on small-world networks, Phys. Rev. E, pp. 7059–7064, 2000.

[89] J. L. Moreno, Who shall survive?, Beacon House, Beacon, NY, 1934.

[90] M. Morris, Sexual networks and HIV, AIDS, 97: Year in Review, 11, pp. 209–216, 2002.

[91] M. E. J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, Phys. Rev. E, 64, 2001.

[92] M. E. J. Newman, Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, Phys. Rev. E, 64, 2001.

[93] M. E. J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA, 98, pp. 404–409, 2001.

[94] M. E. J. Newman, The spread of epidemic disease on networks, Physical Review E, 66(1), 2002.

[95] M. E. J. Newman, Ego-centered networks and the ripple effect, Social Networks, 25, pp. 83–95, 2003.

[96] M. E. J. Newman, The structure and function of complex networks, SIAM Review, 45, pp. 167–256, 2003.

[97] M. E. J. Newman, Random graphs as models of networks, Wiley-VCH Verlag GmbH & Co. KGaA, pp. 35–68, 2005.

[98] M. E. J. Newman, S. Forrest, and J. Balthrop, Email networks and the spread of computer viruses, Phys. Rev. E, 66, 2002.

[99] M. E. J. Newman, S. H. Strogatz, and D. J. Watts., Random graphs with arbitrary degree distributions and their applications, Physical Review E, 64(2), 2001.

[100] J. F. Padgett and C. K. Ansell, Robust action and the rise of the Medici, Am. J. Sociol., 98, pp. 1259–1319, 1993.

[101] R. Pastor-Satorras and A. Vespignani, Epidemic spreading in Scale-Free networks, Physical Review Letters, 86(14), 2001.

[102] S. L. Pimm, Food webs, University of Chicago Press, Chicago, 2nd ed., 2002.

[103] Z. N. Podani, J.and Oltvai, H. Jeong, A.-L. Tombor, B.and Barabasi, and E. Szathmary, Comparable system-level organization of Archaea and Eukaryotes, Nature Genetics, 29, pp. 54–56, 2001.

[104] J. J. Potterat, L. Phillips-Plummer, S. Q. Muth, R. B. Rothenberg, D. E. Woodhouse, T. S. Maldonado-Long, H. P. Zimmerman, and J. B. Muth, Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs, Sexually Transmitted Infections, 78, pp. 159–163, 2002.

[105] D. J. d. S. Price, Networks of scientific papers, Science, 149, pp. 510–515, 1965.

[106] A. Rapoport, Contribution to the theory of random and biased nets, Bulletin of Mathematical Biophysics, 19, pp. 257–277, 1957.

[107] A. Rapoport and W. J. Horvath, A study of a large sociogram, Behavioral Science, 6, pp. 279–291, 1961.

[108] S. Redner, How popular is your paper? An empirical study of the citation distribution, Eur. Phys. J. B, 4, pp. 131–134, 1998.

[109] A. Rinaldo, I. Rodriguez-Iturbe, and R. Rigon, Channel networks, Annual Review of Earth and Planetary Science, 26, pp. 289–327, 1998.

[110] I. Rodriguez-Iturbe and A. Rinaldo, Fractal river basins: Chance and self-organization, Cambridge University Press, Cambridge, 1997.

[111] P. O. Seglen, The skewness of science, J. Amer. Soc. Inform. Sci., 43, pp. 628–638, 1992.

[112] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna, Small-world properties of the Indian railway network, Preprint cond-mat/0208535, 2002.

[113] R. V. Sole and J. M. Montoya, Complexity and fragility in ecological networks, Proc. R. Soc. London B, 268, pp. 2039–2045, 2001.

[114] R. V. Sole and R. Pastor-Satorras, Complex networks in genomics and proteomics, Wiley-VCH, Berlin, 2003.

[115] O. Sporns, Network analysis, complexity, and brain function, Complexity, 8(1), pp. 56–60, 2002.

[116] O. Sporns, G. Tononi, and G. M. Edelman, Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices, Cerebral Cortex, 10, pp. 127–141, 2000.

[117] A. Vázquez, Knowing a network by walking on it: emergence of scaling, Europhys. Lett., 54(23), 2001.

[118] J. Verzani, *Using R for Introductory Statistics*, CRC Press, 2004.

[119] A. Wagner and D. Fell, The small world inside large metabolic networks, Proc. R. Soc. London B, 268, pp. 1803–1810, 2001.

[120] D. J. Watts, Small Worlds, Princeton University Press, Princeton, 1999.

[121] D. J. Watts and S. H. Strogatz, Collective dynamics of small-world networks, Nature, 393, pp. 440–442, 1998.

[122] G.-W. Weber, *Generalized Semi-Infinite Optimization and Related Topics*, Heldermann Verlag publishing house, Research and Exposition in Mathematics 29, Lemgo, eds: K.H. Hofmannn and R. Wille, 2003.

[123] G.-W. Weber, IAM 557 Statistical Learning and Simulation Lecture Notes, METU, 2011.

[124] G.-W. Weber, P. Taylan, Z. Alparslan-Gök, S. Özuğur, and B. Aktekе-Öztürk, Optimization of gene-environment networks in the presence of errors and uncertainty with chebychev approximation, TOP, the Operational Research journal of SEIO (Spanish Statistics and Operations Research Society), 16(2), pp. 284–318, 2008.

[125] G.-W. Weber, P. Taylan, Z. Alparslan-Gök, S. Özuğur, and B. Akteke-Öztürk, A new mathematical approach in environmental and life sciences: gene-environment networks and their dynamics, Environmental Modeling and Assessment, 14(2), pp. 267–288, 2009.

[126] G.-W. Weber and A. Tezel, New views: Generalized semi-in

nite optimization of genetic networks, TOP, the Operational Research journal of SEIO (Spanish Statistics and Operations Research Society), 14(1), pp. 48–55, 2006.

[127] H. D. White, B. Wellman, and N. Nazer, Does citation reflect social structure? Longitudinal evidence from the globenet interdisciplinary research group, Preprint, University of Toronto, 2003.

[128] J. G. White, E. Southgate, J. N. Thompson, and S. Brenner, *The structure of the nervous system of the nematode*, C. Elegans, Phil. Trans. R. Soc. London, 1986.

APPENDIX A

COLLECTING THE DATA

All data is derived from the archives of ArXiv. Process for obtaining this data being compelling and requiring analytical thinking, it is worth mentioning. It can be summarized as follows:

Copying and Filtering: I copied all information to any excel page, filtering authors using excel command *filter*.

Removing Various Expressions: I tried to deal with the problem that some people identified themselves in different ways on their different papers.

Field of Study	Different uses of the names
Cryptography and Security	Heam Pierre-Cyrille Héam Pierre-Cyrille
Financial Mathematics	J.P.Bouchaud J.-P.Bouchaud
Numerical Analysis	Hyeonbae Kang Hyœnbæ Kang
Optimization and Control	A. Agrachev Andrei A. Agrachev Andrei Agrachev Andrey A. Agrachev Andrey Agrachev
Statistics	Abdelouahab Bibi Abdelouhab Bibi

Table A.1: An example of the problem we faced during the performance to get the data about the names of the researchers.

Firstly, all Latin alphabet letters with diacritics and letters using acute accent were replaced by the most imminent ones found in English alphabet. For instance,

without regard to whether they are capital or not, á, à, æ, í, é, ø, ś were all changed with a, i, e, o, s, respectively. Secondly, same logic was followed to get rid of other problems except abbreviation of names. Nonetheless, when assigning each scientist a number, by matching abbreviations of each name with the same number, I solved this problem, too.

Converting collaborator relationships into a matrix: Using the numbers in which each represents a mathematician, a collaborator relationship matrix was constructed via excel command *vlookup*. Then, I utilized a very simple matlab code to form the matrix into a matrix with 2-columns consisting of reciprocal relationships.

APPENDIX B

GENERAL INFORMATION ABOUT DISTRIBUTION FUNCTIONS

B.1 The way to get the normalized form of the distribution functions

B.1.1 Power Law

Normalized distribution function satisfies

$$Pr(x_{min} \leq x \leq x_{max}) = \int_{x_{min}}^{x_{max}} Cx^{-\alpha} dx = 1. \tag{B.1}$$

In this thesis, we will take $x_{max} = \infty$ so that Eqn. (B.1) becomes

$$Pr(x \geq x_{min}) = \int_{x_{min}}^{\infty} Cx^{-\alpha} dx = 1. \tag{B.2}$$

Solving this, we get

$$\frac{1}{C} = \frac{x^{-\alpha+1}}{-\alpha+1} \Bigg|_{x=x_{min}}^{x=\infty} = \frac{x_{min}^{-\alpha+1}}{\alpha-1}. \tag{B.3}$$

Finally, the probability density function is

$$p(x) = (\alpha-1)x_{min}^{\alpha-1}x^{-\alpha}. \tag{B.4}$$

For the discrete case, normalized distribution function satisfies

$$Pr(x \geq x_{min}) = \sum_{x_{min}}^{\infty} Cx^{-\alpha} = 1. \tag{B.5}$$

Following that, we get

$$C = \frac{1}{\sum_{x_{min}}^{\infty} x^{-\alpha}} = \frac{1}{\sum_{j=0}^{\infty}(x_{min}+j)^{-\alpha}} = \frac{1}{\zeta(\alpha, x_{min})}. \tag{B.6}$$

Finally,

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})}. \tag{B.7}$$

In fact, hurwitz zeta function can be transformed into

$$\zeta(s,a) = \frac{1}{\Gamma(a)} \int_0^\infty \frac{t^{s-1}dt}{e^{at}(1-e^{-t})}, \tag{B.8}$$

where $\Gamma(a)$ is a gamma function. This transformation can help you for a faster calculation of Hurwitz zeta function.

B.1.2 Power Law with Exponential Cut-off

$$Pr(x \geq x_{min}) = \int_{x_{min}}^\infty Cx^{-\alpha}e^{-\lambda x}dx = 1. \tag{B.9}$$

Then, we get

$$C^{-1} = \int_{x_{min}}^\infty x^{-\alpha}e^{-\lambda x}dx = \lambda^{-1+\alpha}\int_{\lambda x_{min}}^\infty z^{-\alpha}e^{-z}dz, \tag{B.10}$$

using the property of the gamma function

$$\frac{d}{dx_{min}}\int_{x_{min}}^\infty x^{-\alpha}e^{-\lambda x}dx = -x_{min}^{-\alpha}e^{-\lambda x_{min}} \tag{B.11}$$

$$= \frac{d}{dx_{min}}\lambda^{-1+\alpha}\int_{\lambda x_{min}}^\infty z^{-\alpha}e^{-z}dz. \tag{B.12}$$

So, the probability density function

$$p(x) = \frac{\lambda^{-1+\alpha}x_{min}^{-\alpha}e^{-\lambda x_{min}}}{\Gamma(1-\alpha,\lambda x_{min})}, \tag{B.13}$$

where $\Gamma(\beta,z)$ is known as a upper (incomplete) gamma function. Upper (incomplete) gamma function converges for only $z > 0$. You can use the matlab commands *gammainc*$(\beta, z,$"*upper*"$)$ for $\beta > 0$. For $\beta < 0$, you can find the code in Appendix (C.1).

B.1.3 Log Normal

$$Pr(x \geq x_{min}) = \int_{x_{min}}^\infty C\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]dx = 1. \tag{B.14}$$

Then, we get

$$C = \left[\int_{x_{min}}^\infty \frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]dx\right]^{-1}. \tag{B.15}$$

A transformation as in Eqns. (B.12) and (B.13) done, we obtain

$$\int_{x_{min}}^{\infty} \frac{1}{x} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right] dx = \sqrt{2}\sigma \int_{\frac{(\ln x_{min}-\mu)}{\sqrt{2}\sigma}}^{\infty} e^{-t^2} dt. \tag{B.16}$$

Then,

$$\sqrt{2}\sigma \int_{\frac{(\ln x_{min}-\mu)}{\sqrt{2}\sigma}}^{\infty} e^{-t^2} dt = \frac{\sigma\sqrt{\pi}}{\sqrt{2}} \mathrm{erfc}(\frac{(\ln x_{min}-\mu)}{\sqrt{2}\sigma}), \tag{B.17}$$

where $erfc$ is a complementary error function defined as

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt. \tag{B.18}$$

Finally, the probability density function is

$$p(x) = \frac{2}{\sqrt{\pi\sigma^2}} \left[\mathrm{erfc}(\frac{(\ln x_{min}-\mu)}{\sqrt{2}\sigma})\right]^{-1} \frac{1}{x} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]. \tag{B.19}$$

You can use the matlab commands erf and $erfc$ to calculate the error function and the complementary error function, respectively. These function are useful for $x \in \mathbb{R}$.

The same procedure can be applied for the other distributions.

B.2 Maximum Likelihood Estimations

B.2.1 Continuous Power Law

Given a data set containing n observations $x_i \geq x_{min}$, we would like to know the value of α for the power-law model that is most likely to have generated our data. The probability that the data were drawn from the model is proportional to

$$p(x|\alpha) = \prod_{i=1}^{n} \frac{\alpha-1}{x_{min}} \left(\frac{x_i}{x_{min}}\right)^{-\alpha}. \tag{B.20}$$

This probability is called the likelihood of the data given the model. The data are most likely to have been generated by the model with scaling parameter α that maximizes this function. Commonly we actually work with the logarithm $\mathcal{L}$

of the likelihood, which has its maximum in the same place:

$$\mathcal{L} = \ln(p(x|\alpha)) = \ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{min}} \left(\frac{x_i}{x_{min}} \right)^{-\alpha} \tag{B.21}$$

$$= \sum_{i=1}^{n} \left[\ln(\alpha - 1) - \ln(x_{min}) - \alpha \ln \frac{x_i}{x_{min}} \right] \tag{B.22}$$

$$= n \ln(\alpha - 1) - n \ln x_{min} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{min}}. \tag{B.23}$$

Setting $\frac{\partial \mathcal{L}}{\partial \alpha} = 0$ and solving for α, we obtain the maximum likelihood estimate or MLE for the scaling parameter:

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}. \tag{B.24}$$

B.2.2 Discrete Power Law

Following an argument similar to the one we gave for the continuous power law, we can write down the log-likelihood function

$$\mathcal{L} = \ln \prod_{i=1}^{n} \frac{x_i^{-\alpha}}{\zeta(\alpha, x_{min})} = -n \ln \zeta(\alpha, x_{min}) - \alpha \sum_{i=1}^{n} \ln x_i. \tag{B.25}$$

Setting $\frac{\partial \mathcal{L}}{\partial \alpha} = 0$ we find that

$$\frac{-n}{\zeta(\alpha, x_{min})} \frac{\partial}{\partial \alpha} \zeta(\alpha, x_{min}) - \sum_{i=1}^{n} \ln x_i = 0. \tag{B.26}$$

Thus, the MLE $\hat{\alpha}$ for the scaling parameter is the solution of

$$\frac{\zeta'(\alpha, x_{min})}{\zeta(\alpha, x_{min})} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i. \tag{B.27}$$

This equation can be solved numerically in a straightforward manner. Alternatively, one can directly maximize the log-likelihood function itself, Eqn. (B.25). As the calculations involved are long with results that do not look very structured, however, we omit them here. Eqn. (B.27) is somewhat cumbersome. If x_{min} is moderately large, then a reasonable figure for α can be estimated using the much more convenient approximate formula derived in the next steps.

Given a differentiable function $f(x)$, with indefinite integral $F(x)$, such that $F'(x) = f(x)$,

$$\begin{aligned}\int_{x-\frac{1}{2}}^{x+\frac{1}{2}} f(t)dt &= F(x+\frac{1}{2}) - F(x-\frac{1}{2}) \\ &= \left[F(x) + \frac{1}{2}F'(x) + \frac{1}{8}F''(x) + \frac{1}{48}F'''(x)\right] \\ &\quad - \left[F(x) - \frac{1}{2}F'(x) + \frac{1}{8}F''(x) - \frac{1}{48}F'''(x)\right] + \ldots \\ &= f(x) + \frac{1}{24}f''(x) + \ldots\end{aligned}$$

Summing over integer x, we then get

$$\int_{x_{min}-\frac{1}{2}}^{\infty} f(t)dt = \sum_{x=x_{min}}^{\infty} f(x) + \frac{1}{24}\sum_{x=x_{min}}^{\infty} f''(x) + \ldots \tag{B.28}$$

For instance, if $f(x) = x^{-\alpha}$ for some constant α, then we have

$$\begin{aligned}\int_{x_{min}-\frac{1}{2}}^{\infty} t^{-\alpha}dt &= \frac{(x_{min}-\frac{1}{2})^{-\alpha+1}}{\alpha-1} \\ &= \sum_{x=x_{min}}^{\infty} x^{-\alpha} + \frac{\alpha(\alpha+1)}{24}\sum_{x=x_{min}}^{\infty} x^{-\alpha-2} + \ldots \\ &= \zeta(\alpha, x_{min})[1 + O(x_{min}^2)],\end{aligned}$$

where we have made use of the fact that $x^2 \leq x_{min}^2$ for all terms in the second sum. Thus

$$\zeta(\alpha, x_{min}) = \frac{(x_{min}-\frac{1}{2})^{-\alpha+1}}{\alpha-1}\left[1 + O(x_{min}^2)\right]. \tag{B.29}$$

Differentiating this expression with respect to α, we also have

$$\zeta'(\alpha, x_{min}) = -\frac{(x_{min}-\frac{1}{2})^{-\alpha+1}}{\alpha-1}\left[\frac{1}{\alpha-1} + \ln(x_{min}-\frac{1}{2})\right]\left[1 + O(x_{min}^{-2})\right]. \tag{B.30}$$

We can use these expressions to derive an approximation to the maximum likelihood estimator for the scaling parameter α of the discrete power law, Eqn. (B.27), valid when x_{min} is large. The ratio of zeta functions in Eqn. (B.27) becomes

$$\frac{\zeta'(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})} = -\left[\frac{1}{\hat{\alpha}-1} + \ln(x_{min}-\frac{1}{2})\right]\left[1 + O(x_{min}^2)\right], \tag{B.31}$$

and, neglecting quantities of order x_{min}^{-2} by comparison with quantities of order 1, we have

$$\hat{\alpha} \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min} - \frac{1}{2}} \right]^{-1} . \tag{B.32}$$

which is in fact identical to the MLE for the continuous case except for the $-\frac{1}{2}$ in the denominator.

B.2.3 Log-Normal

Using the same procedure, we get the MLEs of *mean*

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln x_i}{n}, \tag{B.33}$$

and *standart deviation*

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (\ln x_i - \hat{\mu})^2}{n}}. \tag{B.34}$$

By the way, calculating the MLEs for (continuous and discrete) exponential functions are simple. However, following the same way in order to calculate the MLEs of parameters of weibull and PL with cut-off, we get a transcendental equation. For weibull distribution, we apply the well-known numerical rootfinding procedure, Newton-Raphson method. For PL with cut-off, we use the matlab command $fminsearch$. Although $fminsearch$ command is helpful, it is sensitive to the initial conditions. In my thesis, I choose $\alpha_{initial} = 2.5$ and $\lambda_{initial} = 0.0001$. For matlab codes, see Appendix (C.1).

APPENDIX C

MATLAB GUIDE

C.1 MLE+KS Method for Power Law Distribution with Cut-off

Since an implementation of the method for power law distributions with cut-off is the most challenging one, we find sufficient to put only it into Appendix. For all the other codes, please contact by simge.guneri@metu.edu.tr.

```
% This function sorts any vector u in ascending order, counts how many times each element
% is repeated and then removes its duplicated elements.
% Consider the data vector u=[ 3 2 1 1 2 2 ]'. Outputs will be x=[1 2 3]' and y=[2 3 1]'.
% We have implemented this function twice to only the first coloumn of the data mentioned
% in Appendix A. After that operation, i^th gives us the number of the mathematicians
% with x(i) collaborators.
function [x,y] = order(u)
n=length(u);
u=sort(u);
k=1;
i=1;
while(i<=n)
    s=1;
        for j=i+1:n
            if(u(i)==u(j))
                s=s+1;
            end
        end
    y(k)=s;
    k=k+1;
    i=i+s;
end
x=unique(u);
end
```

```
% This function computes the upper complete gamma function for a negative α. This will be
```

```
% useful when applying to MLE+KS method for power law distribution with cut-off.
function [u] = ugamma3(z,a)
v=0;
while(v<abs(a))
      v=v+1;
end
b=gamma(a+v).*gammainc(z,a+v,'upper');
g(1)=b;
for k=1:v
     g(k+1)=(g(k)-exp(-z).*z.^(a+v-k))./(a+v-k);
end
u=g(v+1);
end
```

```
% This function finds best estimations of the parameters α, λ and x_min using MLE+KS
% method.
% Before starting, we need initial predictions for the parameters α and λ. In this thesis,
% we get lb = 0.0001, an initial estimation for λ, and ub = 2.5, an initial value for α.
% plcut(x,y,'p') plots the empirical and theoretical distribution functions.
% plcut(x,y,xmin) gives the estimations of the parameters for xmin you entered.
function [alphaa,lambdaa,xmin,a,L] = plcut(x,y,lb,ub,varargin)
n=length(x);
a=1;
lambda=zeros(n-1,1);
alpha=zeros(n-1,1);
if nargin==4 || ischar(varargin{1})
     fprintf('\n i    lambda    alpha       exitflag    loglikelihood\n');
end
for i=1:round(2.*n./3)
      xminimum(i)=x(i);
      sum1=sum(y(i:n));
      sum2=sum(x(i:n).*y(i:n));
      sum3=sum(log(x(i:n)).*y(i:n));
      f=@(u)-((1-u(2)).*sum1.*log(u(1))-sum1.*...
      log(ugamma3(u(1).*xminimum(i),1-u(2)))-u(2).*sum3-u(1).*sum2);
      [u,fval,exitflag]=fminsearch(f,[lb,ub]);
      lambda(i)=u(1);
      alpha(i)=u(2);
      if nargin==4 || ischar(varargin{1})
         fprintf('%3d %15.4e %20.4f %13d %17.4f\n\n',i,lambda(i),alpha(i),exitflag,-fval);
      end
      if( alpha(i)>1 && alpha(i)<5 )
         for j=i:n
            sum4=sum((y(j:n)./sum1));
            d(j)=abs(ugamma3(lambda(i).*x(j),1-alpha(i))./ugamma3(lambda(i).*...
            xminimum(i),1-alpha(i))- sum4);
         end
```

```
        d1(i)=max(d);
        if nargin==4 || ischar(varargin{1})
           fprintf('distance=%13.4f\n\n',d1(i));
        end
        if nargin==4 || ischar(varargin{1})
            if (d1(i)<=a)
              L=-fval;
              a=d1(i);
              xmin=xminimum(i);
              alphaa=alpha(i);
              lambdaa=lambda(i);
              k=i;
           end
        end
        if nargin>=5 && ~ischar(varargin{1}) && xminimum(i)==varargin{1}
           L=-fval;
           a=d1(i);
           xmin=xminimum(i);
           alphaa=alpha(i);
           lambdaa=lambda(i);
           k=i;
            break;
         end
     end
end
for m=1:length(varargin)
     if nargin>=5 && ischar(varargin {m})
        z=xmin;
        while (z¡x(n))
           loglog(z,ugamma3(lambdaa.*z,1-alphaa)./ugamma3(lambdaa.*xmin,1-alphaa),'o')
           hold on
           z=z+0.5;
        end
        sum1=sum(y(k:n));
        for u=k:n
           loglog(x(u),sum((y(u:n)./sum1)),'mo')
        end
        hold off
     end
end
end
```

C.2 Algorithm for Plotting Figure 3.1

```
% By means of this codes we draw 100 different samples. Each has 1150 observations. 150 of
%the observations are uniformly distributed and their values are less than x_min75. Remaining
% observations are power law distributed with α = 2.4 and x_min = 75.
% The function randht is taken from http://tuvalu.santafe.edu/ aaronc/powerlaws/.
% Space being limited, we did not put the function plfitc here. Please send an email to
%simge.guneri@metu.edu.tr or use the similar code in http://tuvalu.santafe.edu/ aaronc/powerlaws/.
i=1;
xmin=1;
x=2;
while (i<1000 && i<max(x))
x = randht(1000,'xmin',75,'powerlaw',2.4);
x=[x; 10*rand(150,1)]';
[u,y]=order(x);
a=plfitc(u,y,xmin);
plot(xmin,a);
hold on
i=i+1;
xmin=xmin+1;
end
```

C.3 Algorithm for Network Visualization

We have drawn Figures 2.1, 2.2, 2.3, 2.4, 2.6, 2.7, 2.8 and 4.1 using this code.

```
z=accumarray(A,1);  % This finds an adjacency matrix of A which is an nx2 matrix where
% elements of each row represents two vertices which are one-path connected.
% Other part of the code is taken from http://stackoverflow.com/questions/5804468/drawing-
% a-network-of-nodes-in-circular-formation-with-links-between-nodes
theta=linspace(0,2*pi,313);
theta=theta(1:end-1);
[a,b]=pol2cart(theta,1);
[ind1,ind2]=ind2sub(size(z),find(z(:)));
h=figure(1); clf(h);
plot(a,b,'.k','markersize',5);
hold on
arrayfun(@(p,q)line([a(p),a(q)],[b(p),b(q)]),ind1,ind2);
```

axis **equal off**

C.4 Algorithm for Calculating First Three Mean Degrees and Number of Triples of Connected Vertices in a Network

---

```
function [ ] = neighbour(A)
% A is an nx2 matrix where elements of each row represent two vertices which are one-path
% connected.
[u,y]=order(A(:,1));
m=length(A(:,1));
n=length(y);
i=1; s=1;
while(i<=m && s<=n)
    z=[ ]; z1=[ ];
    for k=i:i+y(s)-1
        [r]=find(A(:,1)==A(k,2));
        z=[z; A(r,2)];
    end
    q=sort(z);
    q=unique(q);
    for k=i:i+y(s)-1
        for j=1:length(q)
            if(q(j)==A(k,2) || q(j)==A(i,1))
                q(j)=0;
            end
        end
    end
    for k=1:length(q)
        if (q(k)~=0)
            [r]=find(A(:,1)==q(k));
            z1=[z1;A(r,2)];
         end
    end
    for k=i:i+y(s)-1
         for j=1:length(q)
             for t=1:length(z1)
                 if(z1(t)==A(k,2) || z1(t)==A(i,1) || z1(t)==q(j))
                     z1(t)=0;
                  end
             end
        end
```

```
    end
    z1=sort(z1);
    z1=unique(z1);
    b1(s)=sum(z1>0);
    b(s)=sum(q>0);
    i=i+y(s);
    s=s+1;
end
[u2,y2]=order(y);
mean=sum(u2(1:length(u2)).*y2(1:length(u2))./sum(y2));
[u1,y1]=order(b);
mean1=sum(u1(1:length(u1)).*y1(1:length(u1))./sum(y1));
[u3,y3]=order(b1);
mean2=sum(u3(1:length(u3)).*y3(1:length(u3))./sum(y3));
fprintf('\n mean number of first neighbours of a vertex=%4.8f\n',mean);
fprintf('\n mean number of second neighbours of a vertex=%4.8f\n',mean1);
fprintf('\n mean number of third neighbours of a vertex=%4.8f \n',mean2);
triple=0;
for i=1:length(y2)
    if (u2(i) =1)
       triple=triple+y2(i).*(factorial(u2(i))./(2.*factorial(u2(i)-2)));
     end
end
fprintf('\n number of triples of connected vertices=%4.8f\n',triple);
end
```

C.5 Algorithm for Calculating The Number of Triangles in a Network

```
function [b] = tri(a)
% This function converts an nxm matrix into an nx3 matrix. The input is an nxm matrix
% where elements of each row represent the one-path connected vertices.
% When we put the output matrix b into an excel page and remove its duplicated rows in
% excel, the number of the rows of new matrix gives the number of the triangles.
[n,m]=size(a);
y=[ ];
for j=1:m-2
    for k=j+1:m-1
       x=[a(:,j) a(:,k) a(:,k+1)];
       for s=1:m-k-1
          x=[x;a(:,j) a(:,k) a(:,k+1+s)];
       end
```

```
        y=[y;x];
    end
end
b=sort(y');
b=b';
end
```

C.6 Algorithm for Calculating Component Sizes in a Network

This function is also confirmed by collaboration dataset in http://snap.stanford .edu/data/.

```
function [d,z1,B] = component(A)
% d is a vector in which each element is a component size. z1 is a vector representing the
% vertices in a giant component. B is an n$x$2 matrix representing the connections between
% the vertices in a giant component.
[u,h]=order(A(:,1));
m=length(A(:,1));
n=length(h);
i=1; s=1; u=[ ]; z=[ ]; x=0; B=[ ];
k1=23; k2=5;
while(i<=m && s<=n)
u=[u;z];
[r]=find(u==A(i,1));
    if isempty(r)
        x=x+1;
        z=[ ];
        for k=i:i+h(s)-1
            [r]=find(A(:,1)==A(i,2));
            z=[z; A(r,2)];
        end
        w=z;
        for v=1:k1
            q=order(w);
            w=[ ];
            q=order(z);
            for k=1:length(q)
                [r]=find(A(:,1)==q(k));
                w=[w;A(r,2)];
                z=[z;A(r,2)];
            end
        end
```

```
        z=order(z);
        fprintf('\n%2d. component size is %4d and vertices in the component:\n',x,length(z));
        if (length(z)>1000)
            z1=z;
            j=1;
            k1=k2;
                while (j<=length(z))
                    [r]=find(A(:,1)==z(j));
                    B=[B; A(r,:)];
                    j=j+1;
                end
            end
            d(x)=length(z);
        end
        i=i+h(s);
        s=s+1;
    end
    d=d';
end
```

C.7 Algorithm for Calculating The Mean Shortest Distance and Diameter in a Network

```
function [z] =sm(x,y)
n=length(x);
m=length(y);
if (m>n)
    for i=1:n
        z(i)=x(i)+y(i);
    end
    for i=n+1:m
        z(i)=y(i);
    end
else
    for i=1:m
        z(i)=x(i)+y(i);
    end
    for i=m+1:n
        z(i)=x(i);
    end
end
end
```

```
% This function computes the mean distance and the diameter in a giant component. The
% input A is an nx2 matrix representing the connections between the vertices in the giant
% component.
function [d,d1] = dist(A)
[u,y]=order(A(:,1));
m=length(A(:,1));
n=length(y);
i=1; s=1; d1=[ ];
d(1)=y(1);
while(i<=m && s<=n)
    if s>2
       d=sm(d,d1);
    end
    if s>1
       d1(1)=y(s);
       x=u(1):u(s-1);
       for w=1:length(x)
          [r]=find(A(i:i+y(s)-1,2)==x(w));
          d1(1)=d1(1)-length(r);
       end
    end
    z=[ ]; z1=[ ];
    for k=i:i+y(s)-1
       [r]=find(A(:,1)==A(k,2));
       z=[z; A(r,2)];
     end
q=sort(z);
q=unique(q);
    for k=i:i+y(s)-1
       for  j=1:length(q)
          if(q(j)==A(k,2) ||  q(j)==A(i,1))
             q(j)=0;
          end
       end
    end
    if i==1
       d(2)=sum(q>0);
    else
       d1(2)=sum(q>0);
    end
    if i~=1
       for w=1:length(x)
          [r]=find(q==x(w));
          d1(2)=d1(2)-length(r);
       end
    end
    q1=q;
    for l=1:16
       for k=1:length(q)
          if (q(k)~=0)
             [r]=find(A(:,1)==q(k));
             z1=[z1;A(r,2)];
           end
```

```
        end
        for k=i:i+y(s)-1
            for j=1:length(q1)
                for t=1:length(z1)
                    if(z1(t)==A(k,2) || z1(t)==A(i,1) || z1(t)==q1(j))
                        z1(t)=0;
                    end
                end
            end
        end
        z1=sort(z1);
        z1=unique(z1);
        q=z1;
        q1=[q1;z1];
        q1=order(q1);
        if i==1
            d(l+2)=sum(q>0);
                else
            d1(l+2)=sum(q>0);
        end
        if i~=1
            for w=1:length(x)
                [r]=find(z1==x(w));
                d1(l+2)=d1(l+2)-length(r);
            end
        end
    end
d
d1
i=i+y(s);
s=s+1;
end
x=1:sum(d>0);
mean=0;
for i=1:sum(d>0)
    mean=mean+d(i).*x(i)./sum(d);
end
fprintf('mean shortest distance=%2.4f\n',mean);
fprintf('diameter=%2d \n',sum(d>0));
end
```
