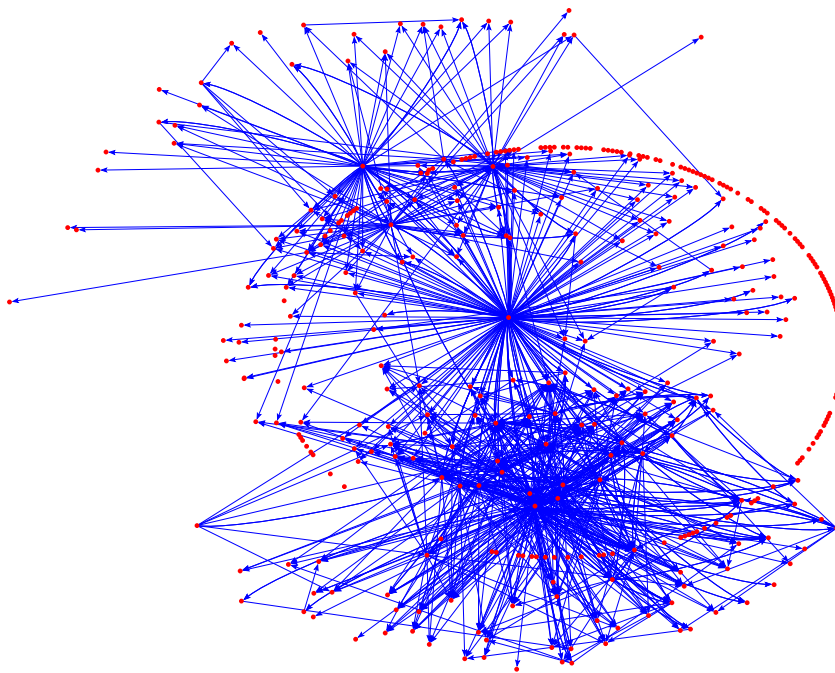


Spires on the Building of Science

Complex Networks and Scientific Excellence

Cand. Scient. Thesis
by Sune Lehmann
Advisor: Benny Lautrup
The Niels Bohr Institute

June 6, 2003



Cover illustration: A portion of the network of papers around Stanley Milgram's 1967 paper on small worlds, 'The Small World Problem' [1], and papers with the words 'Small World' in the title. Many nodes in this network are also present in the list of references of this thesis.

The plot was generated using the *Pajek* freeware program for network visualization and analysis [2]; the data stems from the Garfield collection (<http://www.garfield.library.upenn.edu/histcomp/index.html>). All of these networks are the result of searches in the Web of Science and are used with the permission of ISI of Philadelphia.

Spires on the Building of Science is a thorough investigation of the complex network of scientific publications in the SPIRES hep database. Chapter by chapter, the most important results are:

Chapter 1 Is a general introduction to the field of complex networks. The physics of complex systems and the subject of graph theory (random graphs and the Watts-Strogatz model) are briefly summarized, in order to list some key results and to introduce important nomenclature used in the remainder of the thesis. A number of real world networks are described with a special emphasis on networks related to scientific publications. Finally, a brief history of the SPIRES database is supplied, along with considerations regarding applications of the physics contained in this thesis.

Chapter 2 Here, the network constituted of papers citing papers is investigated. It is found that the probability that a given paper in the SPIRES database has k citations is well described by simple power laws, $P(k) \propto k^{-\alpha}$, with $\alpha \approx 1.2$ for k less than 50 citations and $\alpha \approx 2.3$ for 50 or more citations. A consideration of citation distribution by subfield, shows that the citation patterns of high energy physics form a remarkably homogeneous network. Further, the knowledge of the citation distributions is utilized to demonstrate the extreme improbability that the citation records of selected individuals and institutions have been obtained by a random draw on the resulting distribution. This work was done in collaboration with Benny E. Lautrup and Andrew D. Jackson and it is largely contained in [3].

Chapter 3 The more complicated network that arises when the level of authors is included in the description of the network of scientific publications is discussed. The basic statistics of the resulting network are described, and the two power-law structures from Chapter 2 turn out to re-emerge from the distributions of total citations per author and total number of papers per author. The impact on the distributions of paper- and total citations, when removing the minimally publishing authors is also discussed. The analysis in this chapter is, to a high degree, independent work and the main results will be published in a paper currently in preparation.

Chapter 4 This chapter introduces the brand-new concept of *author citation histories*: We shall discuss the properties of the distribution of first papers, second papers, and so-

forth. It turns out that if we consider authors with more than 25 publications, the quality of their publications is remarkably constant (on average) throughout their careers. Thus, the author citation histories teach us that a highly homogeneous group of authors (with 25+ publications) constitute the backbone of SPIRES; less than 10% of the authors are the target of around 50% of the citations. Further, because of the constant quality of their publications, these authors stand out from the day that they publish their first paper. This chapter is entirely my own and its contents are a part of a paper in preparation.

Chapter 5 A very interesting property of SPIRES is the *longitudinal* correlations in the distribution of citations of papers. Here, Principal Component Analysis (PCA) is utilized to analyze the SPIRES data in order to learn more about these correlations. The new-found group of authors (25+ population) is ideally suited for just this type of investigation, since it turns out that this group of authors is responsible for the vast majority of these longitudinal correlations.

First, the theory behind PCA is reviewed. The covariance matrix for the ‘25+ population’ in SPIRES is established and diagonalized, hereby uncovering the principal components. We then discuss PCA as a means of ‘reverse’ quality control in order to pinpoint interesting authors, and demonstrate the use of PCA on the selected authors from Chapter 2—thus using PCA as an augmentation of the probability measure put forth in that chapter. The idea of employing PCA stems from my advisor, but all of the work and the interpretations are my own. The material here is also intended for publication.

Chapter 6 The growing networks (GN) model proposed by Barabási and Albert [4] is thoroughly reviewed with a focus on the analytical results. Many properties of the model are discussed. The GN model is then used as the starting point when creating a model for the SPIRES data; the resulting model is very successful in recreating the topologies seen in the distribution of citations of papers in SPIRES (discussed in Chapter 2). The work on this simple model for SPIRES is original, and solely the work of the author. In the latter part of this chapter, the subject of several intrinsic problems of the model is discussed.

Chapter 7 This chapter continues where Chapter 6 left off. We begin by discussing other forms of preferential attachment in the GN model and their implications regarding the network topology. After this, some recently proposed alterations of the model are reviewed: We learn how to account for carelessly compiled lists of references and how compensating for ageing of papers affects the citation distributions. The work done on these modifications of the model is not original. Finally, the topic of constructing a model that includes the level of authors, is discussed (this is original work).

Chapter 8 In this postscript, I outline what I believe to be promising venues for future research of the SPIRES database.

Acknowledgements

I owe great thanks to my advisor Benny Lautrup. He has provided more encouragement and support than I had ever dared to hope for, both on physics and on life in general. Andrew

D. Jackson has also been an inexhaustible source of inspiration and knowledge. Closer to home, my family and especially so my girlfriend, Aino, have always been supportive of me and my work. This is not the case for the young scientists whom I have had the pleasure of sharing my office with, while writing this thesis; Søren S. Juhl and Jesper Levinsen have been a continuing source of distraction and inspiration.

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 1.1 | Complex Networks | 1 |
| 1.1.1 | Complex Systems | 1 |
| 1.1.2 | The Study of Networks | 2 |
| 1.2 | Real World Networks | 7 |
| 1.2.1 | The Networks | 7 |
| 1.2.2 | Linking Back to Statistical Physics | 11 |
| 1.3 | Networks of Scientific Publications | 11 |
| 1.3.1 | Co-author Networks | 11 |
| 1.3.2 | Citation Networks | 13 |
| 1.4 | The SPIRES Database | 13 |
| 1.4.1 | A Brief History of SPIRES | 14 |
| 1.5 | Applications | 14 |
| 2 | The Paper Distribution | 17 |
| 2.1 | Basic Properties | 18 |
| 2.1.1 | A Raw Plot | 18 |
| 2.1.2 | Subfields | 18 |
| 2.2 | Homogeneity | 20 |
| 2.2.1 | Skewed Distributions | 20 |
| 2.2.2 | Prior Expectations | 20 |
| 2.2.3 | Visual Comparisons | 21 |
| 2.2.4 | Quantifying Differences | 21 |
| 2.2.5 | Tests? | 24 |
| 2.2.6 | A Potential Source of Inhomogeneity | 25 |
| 2.3 | Form of the Distribution | 26 |
| 2.3.1 | The Asymptotic Tail | 27 |
| 2.4 | The Power of Excellence | 29 |
| 2.4.1 | Examples of the Application of r | 30 |
| 2.5 | Summary | 31 |

| | | |
|----------|--|-----------|
| 3 | The Author Level | 33 |
| 3.1 | A New Level of Complexity | 34 |
| 3.1.1 | Disconnecting the World Wide Web | 35 |
| 3.1.2 | Data Structure and Notation | 35 |
| 3.2 | The Data | 36 |
| 3.2.1 | Problems | 36 |
| 3.2.2 | Quantitative Comparisons | 38 |
| 3.3 | The Frequency Distributions | 40 |
| 3.3.1 | Total Citations per Author | 40 |
| 3.3.2 | Papers per Author | 41 |
| 3.3.3 | Average Number of Citations per Paper per Author | 42 |
| 3.4 | The Scientific Staff | 43 |
| 3.4.1 | The Drop in Total Citations | 44 |
| 3.4.2 | The Paper Citation Distribution | 45 |
| 3.4.3 | Average Number of Citations per Paper per Author | 48 |
| 3.5 | Summary | 49 |
| 4 | Author Citation Histories | 51 |
| 4.1 | Average Citation Histories | 51 |
| 4.1.1 | Plotting the Different Averages | 52 |
| 4.1.2 | Drawing the Line | 54 |
| 4.2 | Total Distribution Histories | 56 |
| 4.2.1 | Unfolding the Averages | 56 |
| 4.2.2 | Percentiles | 58 |
| 4.3 | Summary | 59 |
| 5 | Principal Component Analysis | 61 |
| 5.1 | Preliminaries | 62 |
| 5.1.1 | Multivariate Statistics | 62 |
| 5.1.2 | Matrix Algebra | 63 |
| 5.2 | The Method of Principal Components | 63 |
| 5.2.1 | Conservation of Variability | 64 |
| 5.3 | The SPIRES Covariance Matrix | 65 |
| 5.3.1 | Finding the Right Bins | 65 |
| 5.3.2 | Interpreting Σ_f | 67 |
| 5.3.3 | Diagonalizing the Covariance Matrix | 68 |
| 5.3.4 | Knowing When to Quit | 70 |
| 5.3.5 | Residual Analysis | 71 |
| 5.4 | An Example | 72 |
| 5.5 | Summary | 73 |
| 6 | A Model for SPIRES | 75 |
| 6.1 | Introduction to the Growing Network Model | 75 |
| 6.1.1 | Continuum Solution | 77 |
| 6.1.2 | The Rate Equation Approach | 78 |
| 6.1.3 | Master Equation | 79 |
| 6.2 | Limiting Cases | 82 |

| | | |
|----------|--|------------|
| 6.2.1 | No Preferential Attachment | 82 |
| 6.2.2 | No Growth | 83 |
| 6.2.3 | Finite Size Effects | 84 |
| 6.3 | A Simple Model for SPIRES | 85 |
| 6.3.1 | Numerical Results | 86 |
| 6.3.2 | Analytical Results | 86 |
| 6.4 | Discussion | 91 |
| 6.4.1 | Longitudinal Structure | 91 |
| 6.4.2 | The Age Distribution | 92 |
| 6.4.3 | Measuring Preferential Attachment | 92 |
| 6.4.4 | The Cut-off(s) | 93 |
| 6.5 | Summary | 93 |
| 7 | Modelling and the Real World | 95 |
| 7.1 | Initial Attractiveness | 95 |
| 7.2 | Edge Redirection | 97 |
| 7.3 | Ageing | 99 |
| 7.3.1 | Measuring Ageing in SPIRES | 99 |
| 7.3.2 | Analytical Results | 100 |
| 7.4 | An Author Model | 103 |
| 7.4.1 | Defining the Author Model | 103 |
| 7.4.2 | Results | 104 |
| 7.5 | Summary | 105 |
| 8 | Concluding Scientific Postscript | 107 |
| 8.1 | Future Directions | 107 |
| 8.1.1 | More Data | 107 |
| 8.1.2 | Temporal Aspects | 108 |
| 8.1.3 | Paper Citation Histories | 108 |
| 8.1.4 | Refining the Model(s) | 108 |
| 8.1.5 | Down the Rabbit Hole | 109 |
| A | The Data | 111 |
| A.1 | The Acquisition of Data | 111 |
| A.1.1 | Navigating in SPIRES | 111 |
| A.1.2 | Retrieving the Data | 112 |
| A.1.3 | Problems Regarding the Quality of the Data | 112 |
| A.2 | An Example of a Typical Run | 113 |
| A.3 | A PERL Program | 114 |
| A.4 | Longitudinal Cleaning | 115 |
| B | Miscellanea | 119 |
| B.1 | Residual Analysis | 119 |
| B.2 | Special Functions | 120 |

CHAPTER 1

Introduction

The best way to introduce this *Cand. Scient. Thesis* is to answer a simple question that has arisen in countless discussions since the process of writing began, namely,

What in the world does the distribution of citations in scientific publications have to do with physics!!???

The short answer to this question is simply that the study of citation distributions is a part of *the physics of complex systems*. To answer this question more precisely, a short review of the genesis of the field of complex networks is appropriate:

1.1 Complex Networks

The field of complex networks stems from a union of the fields complex systems and network theory. In this section, we shall briefly study these two fields.

1.1.1 Complex Systems

The physics of complex systems emerged from statistical physics in the 1980's, rooted in ground-breaking work on phase transitions a decade earlier. It is a very inhomogeneous field that is difficult to characterize precisely. In broad strokes, the evolution is as follows. In their seminal paper from 1987, Bak, Tang, and Wiesenfeld [5] suggested that the $1/f$ -noise¹ seen in a number of transport systems normally considered outside the realm of statistical physics—resistors, the hour glass, the luminosity of distant stars, *etc.*—has the same origin as the self-similar fractal structures observed in spatially extended objects such as coast lines, mountain ranges, or cosmic strings. In turbulence, self-similar structures appear both in space and time.

Bak, Tang, and Wiesenfeld hypothesized that the power-law spectra emerge, when the system in question naturally self-organizes into a critical state akin to the microscopic state

¹This name is terrible for more than one reason. First of all, the ' $1/f$ ' part of the expression, should have been $1/f^\beta$, since it refers to the power-law nature of the noise spectra. Second of all, as we shall see in the following, whether 'noise' is the correct classification of these spectra is highly questionable.

of a physical system (gas, liquid, magnet, *etc.*) at the critical point of a phase transition. According to the theory of phase transitions, many physical systems are characterized by *universal* behavior around the critical point, meaning that when we find ourselves on scales that are much larger than the intrinsic scales, the behavior of the system is scale invariant and *does not depend on the microscopic details*. Near the critical point, correlations are distributed according to power-laws with exponents determined by the universality class of the system². The criticality in Bak, Tang, and Wiesenfeld's theory is fundamentally different from the critical point in the theory of phase transitions in one important aspect, namely in its robustness; no fine tuning is needed. In equilibrium statistical physics, the critical point is reached by tuning a parameter, *e.g.* the temperature; here, the critical point is an attractor reached by starting far from equilibrium.

The argument begins with proposing a simple computer model that produces power-law behavior, and their main idea roughly consists in hypothesizing that if in their simple model, a system naturally grows into a self-organized critical (SOC) state and is kept there by the internal dynamics of the system, then this may also be the case for the more complicated macroscopical systems described above. This may not seem like the best of arguments, but it gains strength from the knowledge that if two systems belong to the same universality class, many microscopical details of the system are unimportant: Most features of the specific microscopical dynamics will also be insignificant. Hence, certain general properties will be seen in vastly different systems. Universality classes are typically defined by more fundamental quantities like conservation laws and symmetries.

Thus, with Bak, Tang, and Wiesenfeld it is possible to apply the concept of SOC to a host of wildly disparate phenomena that have the common property that they are characterized by scale-free behavior; in other words, governed by power-laws. At present, complex phenomena are believed to appear almost everywhere in our daily world: earthquakes, evolutionary biology, congested traffic, the economy, *etc.* Because of the diverseness of these physical systems, the physics of complex systems naturally traverses across the boundaries of traditional fields of research, thriving especially in newly formed areas of physics, such as *biophysics*, *geophysics*, and even creating new ones, such as *econophysics*. In this vein the study of citation patterns could be coined *sociophysics*.

Another common denominator is the methodology: Throughout the history of science, physicists have been immensely successful in developing tools to predict the behavior of a system as a whole by understanding the properties of its constituents. This approach is known as *reductionism*, and it is widely believed³ that the almost unbelievably successful application of reductionism in physics and its failure in other fields is due to the simplicity of the interactions in the systems traditionally considered a part of physics. As it turns out, however, the methods of statistical physics seem as if tailor-made for describing systems with more complex interactions, *cf.* the examples mentioned above.

1.1.2 The Study of Networks

The study of networks originated from the mathematical field of graph theory, where Erdős and Rényi lay the foundations of the theory of *random graphs* [8]. As the deficiencies of the

²To limit the size of this introduction, I have been forced to omit countless details. For an excellent and complete introduction to the statistical physics of phase transitions, the interested reader is referred to [6].

³For example by most philosophers of Science and people in the humanities in general. For an introduction to this way of thinking, *cf.* [7]

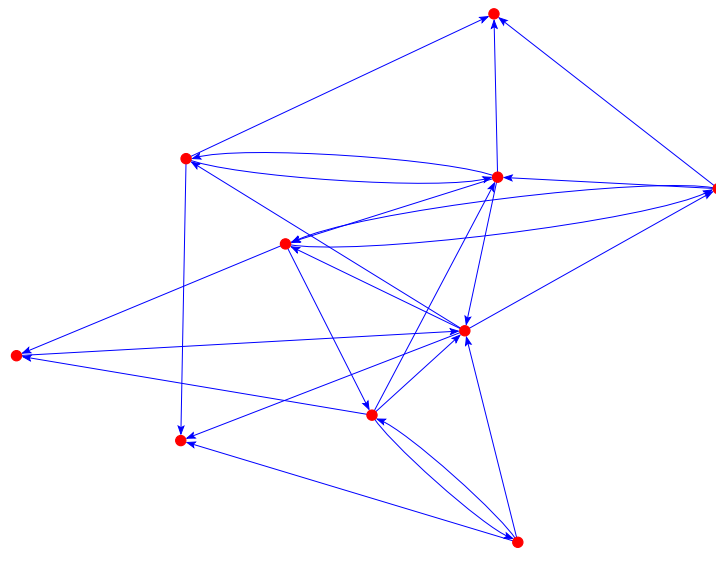


Figure 1.1: A random network with 10 vertices and 25 edges; this corresponds to $p = 5/9$.

random graph as a model for real networks became apparent, Watts and Strogatz designed an elegant model that closely follows our intuitions about social networks [9]. The most recent idea is Barabási's growing network model, characterized by power-law distributions of links [4]. Here, I will review all of these models briefly and introduce common terms and concepts from network theory.

Random Graphs

A random graph is the simplest realization of a complex network: We start out with N points (in the language of network theory, these are called *nodes* or *vertices*) and draw lines (*edges* or *links*) between every pair of nodes with probability p , creating a graph with approximately $pN(N-1)/2$ edges distributed randomly. With this definition, it is clear that the limit of large N , the *degree* (number of edges) of nodes on a random graph is Poisson distributed. Figure 1.1 is an illustration of a random network.

Although random graphs have been studied extensively in the mathematical community, physicists have taken an interest in the study of networks because they wish to understand *real* physical systems. Switching the focus to actual networks, the question inevitably appears whether real complex networks – citation networks or the internet – are fundamentally random. When it comes to citations it is clear that our intuition tells us that the number of citations a given paper receives is a *not* random distribution. On the contrary: The number of citations is considered a measure of the quality of a given paper; therefore the expected degree distribution of scientific publications should be far from random. Analogously, we expect some kind of non-random underlying principles to be reflected in the topology of most other real world systems.

Small-world Networks

One property that is central to all complex networks is what is called the *small-world effect*. The notion of small-world networks was introduced by Harvard social psychologist Stanley Milgram in the 1960's [1]. Milgram conducted a simple experiment, in which a number of letters addressed to an acquaintance of his in Boston, Massachusetts, were distributed to random people in Nebraska—which he considered to be the farthest imaginable place (at least socially speaking) from Boston. Each initial recipient was instructed that the letters were to be handed over to a person known by first name by the sender and most likely to know somebody familiar with the addressee. The letters reached their destination in surprisingly few steps (on average 6), which gave rise to the term ‘six degrees of separation’, which has passed on to popular folklore. Though Milgram’s experiment was very primitive, the general result that any two people can be connected in very few steps is beyond questioning, and it is exactly this property that has resulted in the name small-world networks.

The small-world effect is the reason that random graphs were used as a first approximation to real world networks. It is easy to see that a random graph must display the small-world effect. If a person A is a node in a random network, and has k neighbors, then A has about k^2 second neighbors, and extending this argument, A has k^3 third neighbors and k^4 fourth neighbors and so on. Since most people have between a hundred and a thousand acquaintances, k^4 is already between 10^8 and 10^{12} which is comparable to the population of the world. As explained above, the problem when trying to apply the random graph to the real world is that it does not appeal to our every day experiences—our friends are not chosen at random⁴—our friend’s friends tend to be our friends as well. This means that in a real social network it is not true to say that a person A has k^2 second neighbors, since many of these friends are likely to be his own friends. This property is called *clustering*.

To deal with the subject of clustering in real world networks, one can define a *clustering coefficient*, C , which is defined as the average fraction of pairs of neighbors of a node that are also neighbors of each other [9]. Consider a single node i , with k_i edges. Now, if i ’s neighbors were all connected, there would be $k_i(k_i - 1)/2$ edges between them; the clustering coefficient is defined as the number E_i of actual connections, divided by the total number yielding,

$$C_i = \frac{2E_i}{k_i(k_i - 1)}. \quad (1.1)$$

C is simply the average of the C_i ’s. In a fully connected network (everybody knows everybody), $C = 1$; in a random network $C_{rand} \approx \mathcal{O}(n^{-1})$ [9, 11], which is of course very small for large networks. This was the next step for network theory: To include models that incorporate clustering. This has been done by Watts and Strogatz [9], as we shall see in the following.

The Watts-Strogatz Model

In the class of models proposed by Duncan J. Watts and Stephen H. Strogatz [9], the focus is on creating a model of a small-world network that has a structure that mimics real social networks. The clustering in real social networks appears because most people usually have a

⁴This statement is made very precise in a recent paper by Newman [10] entitled: *Ego-centered networks and the ripple effect—or—Why all your friends are weird*. In this paper Newman demonstrates how the small-world argument made for random graphs above, does not apply to a variety of real networks.

group of friends that are close to them in some sociological sense, our own little sub-culture: co-workers, neighbors, fellow philatelists, *etc.*

The simplest model one can think of that displays clustering, is simply a one dimensional lattice in which each node is connected to its k nearest neighbors. We give it periodic boundary conditions so that it wraps around in a ring. The result is shown in Figure 1.2 a. For the

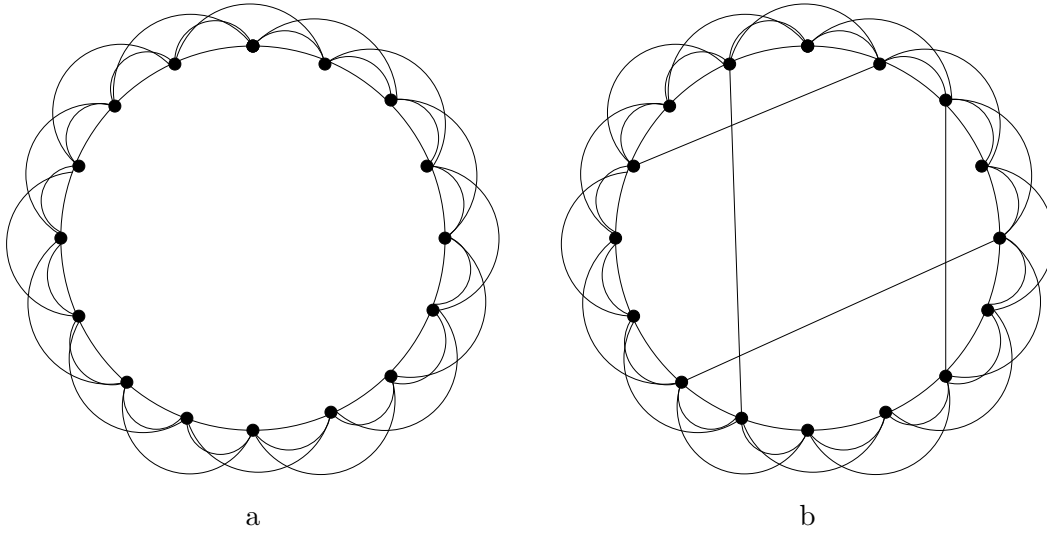


Figure 1.2: a. A one dimensional periodic lattice with $k = 1$. b. The Watts-Strogatz model.

lattice shown in Figure 1.2 a, we can calculate the clustering coefficient exactly; for $k < \frac{2}{3}N$, which will be the case for almost any graph, we find that

$$C = \frac{3(k-2)}{4(k-1)} \quad (1.2)$$

which $\rightarrow \frac{3}{4}$ for $k \rightarrow \infty$. For higher dimensional lattices, the corresponding value of C is

$$C = \frac{3(k-2d)}{4(k-d)}, \quad (1.3)$$

where d is the dimension of the lattice. This expression also tends to $\frac{3}{4}$ for $k \gg 2d$.

Obviously, low-dimensional regular lattices *do not* show the small-world effect, since the typical node-node distance increases too rapidly with the system size. A regular lattice with the shape of a hyper-cube of side L has $N = L^d$ nodes; the average node-node distance for the hyper-cube increases with L or equivalently with $N^{1/d}$. Thus, for a one-dimensional lattice $L \sim N$. If we let d become sufficiently large $N^{1/d}$ becomes a slowly increasing function—in other words, real social networks might roughly be lattices of very high dimensions.

The idea of a social network having the same structure as a regular lattice does not correspond very well to our intuitions. Aside from the group of people in our particular ‘subculture’ (our nearest neighbors on the lattice), most people also have acquaintances in wholly different walks of life, people that are not close to us on the social lattice: The punk rocker has a boyhood friend who became a classical cellist, the Oscar winning actor grew up having a best friend who is now in the army, *etc.* In figure 1.2 b, the Watts and Strogatz

model [9], in which this aspect of the social structure is included, is illustrated. Their proposal is to consider a low dimensional lattice—for instance, a one-dimensional lattice—where, with some probability p , we rewire each of the links. Rewiring means that we move one end of a link to a new position chosen at random from the rest of the lattice; this rewiring models the occasional friend from another subculture. For a small p this results in a graph that is still mostly regular but has a few connections which stretch over long distances. The coordination number is still k on average, although the number of neighbors of any particular node can be greater or smaller than k .

Clearly the strength of the Watts-Strogatz model is the strong correspondence to our intuitions. The model is tailor-made to correspond to how sociologists perceive the structure of human acquaintances.

Results in the Watts-Strogatz Model

Since the topic of this thesis is rather removed from the Watts-Strogatz model, I will only very briefly review the key results. The most important realization in the context of the SPIRES network, is that the Watts-Strogatz model and its variations, have been demonstrated to have exponential degree distributions⁵. It is clear that the clustering coefficient C is close to that of a perfectly ordered lattice for small values of p . Watts and Strogatz have shown, by numerical simulation, that the average node-node distance ℓ is comparable to that of a random graph, even for quite small values of p .

For example, for a random graph of $N = 1000$ and $k = 10$, the value of ℓ is 3.2. For $p = 1/4$ the Watts-Strogatz model gives us $\ell = 3.6$, which is only slightly higher than the random graph value. If we put $p = \frac{1}{64}$, which is quite small, we will find that $\ell = 7.6$ —a number that is still considerably lower than the $\ell = 50$ which is the corresponding result for the perfectly ordered lattice. Thus, these initial simulations show that the Watts-Strogatz model indeed displays both clustering and small-world properties.

There has been a great deal of theoretical interest in the Watts-Strogatz model in the late 90's, which has generated a wide variety of analytical results for this model; these are periphery in the context of this thesis⁶. This due to the fact that the growing networks model (introduced below) has many advantages over the Watts-Strogatz model: The most important of which is that it generates power-law degree distribution of the edges.

Growing Networks

Another—and by far the most successful at present—class of network models is the Growing Networks (GN) models. These incorporate some clustering and also maintain the small-world properties, while emphasizing the fact that networks *grow*! Just like the internet grows (rapidly) over time as webpages are added one after the other, their model grows node by node, with new nodes linking to older nodes according to a ‘rich get richer’ principle. This model was proposed by Barabási and Albert [4]. The degree distributions of many variations of their growing networks have power law tails and are, therefore, often referred to as *scale free networks*. The power-law tail is, as we shall see in the following, a common trait of many real world networks—and an interesting trait as well, because it unexpectedly sheds light on

⁵They are sometimes referred to as *exponential* networks in the literature.

⁶The interested reader is referred to an excellent review by Newman [12], or to one of the more general reviews that can be found in [13, 14].

yet another connection to the physics of complex systems. The details of these ideas are outside the scope of this initial review, but I will discuss this in much detail in Chapter 6, in which I propose a model for the network of scientific publications.

Summary

Defining exactly what a small world network is, is difficult—and still a matter of debate. A reasonable definition, however, is that the average distance between two nodes ℓ (the average path length) should be comparable to the value it would assume on a random graph of the same size and average degree. The diameter is defined as the maximum distance between two nodes of a network. In summary we find that three key results seem to crystallize in the investigation of complex networks:

- Complex networks are small worlds, *i.e.* the average distance between two nodes is comparable to the value it would assume on a corresponding random graph.
- Real networks display a degree of clustering that is higher than what is expected for a random graph.
- The degree distribution of most real world networks are radically different from the Poissonian distributions seen in random and Watts-Strogatz networks. Often the degree distribution of real world networks are power-laws, $P(k) \sim k^\alpha$.

1.2 Real World Networks

Another important step forward for the study of networks is the technological progress. First, the problem of data: Early sociological investigations of networks have all been carried out via field studies, interviewing the members of a certain community, *e.g.* a school [15] or a business community [16]. Interviewing, as a method of collecting data, is surely useful for sociological and anthropological purposes, but from the point of view of statistical physics it suffers from a very serious drawback, namely the relative sparseness of data—to a sociologist a thousand actors is a huge network, to a statistical physicist a data set of a thousand nodes equals poor statistical accuracy.

Secondly, increased computational power: It is now possible to analyze networks containing millions of nodes on a regular personal workstation, enabling physicists to explore questions that could not possibly have been answered just a few years ago.

1.2.1 The Networks

It is interesting to take a look at some of the real world networks that have been studied by physicists, and review some of the properties of these networks that have been unveiled during this study.

World Wide Web

There is no doubt that the world wide web is the largest network for which data is available—it was estimated to consist of some 800 million pages in 1999 [17]. In this network the nodes are the individual web pages and the edges are the hyperlinks connecting them. There is one

important difference between the www and the social networks that we have encountered so far—the www is *directed*. A hyperlink points from one page to another, but not necessarily in the opposite direction.

The internet is a scale free network. Let $P_{in}(k)$ denote the probability that a web page has k incoming links and $P_{out}(j)$ the probability distribution of the outbound links, j ; it has been shown that

$$P_{in}(k) \sim k^{-\alpha_{in}} \quad \text{and} \quad P_{out}(j) \sim j^{-\alpha_{in}}. \quad (1.4)$$

The numerical value of α_{out} varies, assuming the values $\alpha_{out} = 2.45$, $\alpha_{out} = 2.38$, and $\alpha_{out} = 2.72$ in the studies [18, 19, 20], respectively—seeming to increase with the sample size. This is not the case for α_{in} which takes on the same value, $\alpha_{in} = 2.1$ in all of the before-mentioned investigations. Because of the directed nature of the internet, the clustering coefficient C is not directly determinable. In spite of this, it is clear that the internet displays small-world behavior: The average path length in a sample of 325,729 pages was found to be $\ell = 11.2$ ($\ell_{rand} = 8.3$) [18] and in a 50 million sample, Broder *et al.* [20] found an average path length of $\ell = 16$ ($\ell_{rand} = 8.8$). The www is indeed a small world.

Email Network

Another example of a complex network that has recently emerged is the email network. Knowing the structure of an email network is interesting when one has to decide on strategies for fighting computer viruses. In [21] the address books from a large university system were analyzed. The email network consisted of 16,881 address books (nodes); the edges of a node are, of course, email addresses contained in each node (address book). The email network is also directed and both the in and out degree distributions are markedly faster decaying than the power law distributions, found in many other networks. The in-degree is well described by a simple exponential, $P_{in}(k) \sim \exp[-k/k_0]$, and the out-degree by a stretched exponential with exponent $\frac{1}{2}$, $P_{out}(j) \sim (1/\sqrt{j}) \exp[-\sqrt{j/j_0}]$, with $k_0 = 8.58$ and $j_0 = 4.15$. If the system is considered a *semidirected* network (since some of the edges are bi-directional), a clustering coefficient can be calculated, using only these bi-directional edges. It turns out that C_{email} is around one order of magnitude higher than C_{rand} , the clustering coefficient for a random graph with the same parameters.

Internet

The final example from the world of IT is the internet, the network of physical links between the computers that make up the www. We can study the internet at two different levels:

- The level of routers. Each router is a node, and the edges are the physical connections (cables) between these.
- Inter-domain level. At this level each node is a domain (a domain usually consists of many routers and computers).

Both levels are studied in [22] and in both cases the degree distribution follows power-laws. Based on a 1995 study, the slope router level distribution had slope $\alpha_{router} = 2.48$ and for the inter-domain level a slope of $\alpha_{domain} \approx 2.2$ was found.

It has also been established [23, 24] that on the domain level, the clustering coefficient C_{domain} takes on a value somewhere between 0.18 – 0.3—which is orders of magnitude higher

than the $C_{rand} \approx 0.001$ found for a complex graph with the same parameters. With regard to the average path length, ℓ was determined to be somewhere in the range between 3.70 and 3.77, fully in agreement with the corresponding random graph.

Movie actor collaboration network

Surfing to the webpage entitled the ‘*Oracle of Bacon*’⁷ one can calculate the *Bacon number* of any actress or actor. The Bacon number is defined as an actors degree of separation from the actor Kevin Bacon. If you have acted in a movie with Kevin Bacon your Bacon number is 1, if you have been in a movie with someone who has acted with Kevin Bacon, your Bacon number is 2, *etc.* A network centered around a single person is called an *ego-centered network*. In the network of movie actor collaborations, each actor is represented by a node and an edge is established between two actors whenever they co-star in a movie. This type of network is called an *affiliation network*.

The data for this co-star network stems from the very comprehensive *Internet Movie Database*⁸. This database is spectacular in its completeness (Bulgarian art films from the fifties are included!). Thus, the IMDb gives us a unique chance to study a closed and, more importantly, *complete* network.

This network was studied most recently in [25]. For this network, the average path length is $\ell = 3.65$ which corresponds nicely with the value $\ell_{rand} = 2.9$ for a random graph with similar parameters. The actor collaboration network is very connected, with a clustering coefficient that is more than 100 times higher than C_{rand} for the corresponding random network. The degree distribution $P(k)$ once again turns out to be a power law $P(k) \sim k^{-\alpha_{actor}}$, $\alpha_{actor} = 2.3 \pm 0.1$, *cf.* [4].

The Small World of Paul Erdős

As an amusing analogy to the actor collaboration, the network centered around the great (and prolific) mathematician Paul Erdős, who published more than 1500 papers with 507 co-authors, has caught the attention of many scientists (who would like to possess a central place in the small world of mathematics). Again, the structure is that if you have published with Erdős you have Erdős number 1, and so forth. This network has not been charted completely although Barabási *et al.* discusses part of it (papers published 1991-98) in [26, 27]. It is also worth mentioning that Erdős has a Bacon number of four, by virtue of the movie *N is a Number*, a documentary about him in which he plays himself. In the cast, we find Gene Patterson, who later had a small role in the movie *Box of Moonlight*, yielding a Bacon number of three.

Networks of the Cell

Looking at the metabolism of 43 different organisms, Jeong *et al.* have discovered another interesting example of a complex network [28]. In their network representation, the nodes are substrates (such as *ATP*, *ADP*, *H₂O*, *etc.*) and the biochemical reactions, in which the substrates participate, make up the edges of the network. It is in the directed nature of these chemical reactions that the edges are directed, rendering calculation of the clustering

⁷<http://www.cs.virginia.edu/oracle/>

⁸<http://www.imdb.com/>

coefficient impossible. This metabolic network also displays scale free behavior (with both incoming and outgoing distributions having slopes ranging between 2.0 and 2.4). The average path length assumes the value $\ell \approx 3.3$ for all the 43 organisms.

Related to this subject, Sergei Maslov and Kim Sneppen⁹ have recently mapped the networks in yeast *Saccharomyces cerevisiae* [29]. The living cell actually has two levels of networks. One, is the network of metabolic and signaling pathways, shaped by the network of interacting proteins. These are, however, regulated by the genetic regulatory network. Maslov and Sneppen's approach to quantifying the topological properties of both networks consisted in comparing the degrees of interacting nodes to a null model of the network, in which all links were randomly rewired. Again, a familiar pattern of fat tailed (power-law) distributions appear. Also, for both interaction and regulatory networks it was found that links between highly connected proteins are systematically suppressed, whereas those between highly connected and low-connected pairs are favored. This effect is highly important, since it decreases the likelihood of cross talk between different functional modules of the cell; simultaneously, it increases the overall robustness of the network by localizing the effect of deleterious perturbations.

Phone Call Networks

The long distance phone calls made during a single day creates another directed complex network. Here the telephone numbers are nodes and an edge appears whenever one person calls another, the direction of the link going from the caller to the recipient. Both incoming and outgoing degree distributions are power laws, both with slope $\alpha_{call}^{in/out} \approx 2.1$, cf. [30].

Power Grid

The last network described in this section is the power grid of the western United States. The nodes of the power grid are the generators, transformers, and sub-stations; the edges are the transmission lines. Watts and Strogatz [9], considered this network of 4,941 nodes to be undirected¹⁰ and found that, compared to a random graph of same size and average degree, the clustering coefficient $C_{powergrid} = 0.08$ is significantly higher than $C_{rand} = 0.005$, and the the average path lengths are comparable ($\ell_{powergrid} = 18.7$ vs. $\ell_{rand} = 12.4$). The degree distribution of the power grid remains undetermined.

Other Networks

The networks mentioned above are only some of the networks that have been studied recently. Other examples are: The web of human sexual relations, knowing the structure of this network is extremely relevant when considering the spread of sexually transmitted diseases; ecological networks; networks in linguistics, where single words are the nodes and words are connected if they appear next to each other in a sentence; and protein folding networks, where each node represents a different conformation state, and an edge is formed if two states can be obtained from each other in a single elementary move. All of these networks share properties with the ones described above. References can be found in, for example, [13, 14, 31]. Other examples include scientific collaboration networks and citation networks, but since these are

⁹From the Niels Bohr Institute.

¹⁰Although it is surely not! Power lines are directed.

closely related to the topic of this thesis, they will be discussed in further detail later (Section 1.3).

1.2.2 Linking Back to Statistical Physics

Of course the documentation of these networks has spawned a surge of *theoretical* interest within the physics community. As I mentioned above, when analyzing a variety of the networks presented in the preceding section, the methods of statistical physics constitute a particularly well-suited tool: The application of the theoretical methods of physics is what constitutes the direct connection from complex networks to statistical physics. Examples of the statistical physics methods used in this endeavor include: Monte Carlo simulations [18, 32, 33, 34], mean field theory [11, 35], scaling and renormalization group methods [11, 35], percolation theory [36, 37, 38], generating functions [25, 36, 38], the replica method [39], exact solutions [40, 41, 42, 43, 44, 45] and a variety of other techniques. Also, modelling techniques that are well known in statistical physics have been applied [4, 9, 46]. Reviews with an even more comprehensive list of references, than the one presented here, can be found in [12, 13, 14, 31].

1.3 Networks of Scientific Publications

Because of the importance of citations as a measure of quality in most sciences, large databases of scientific publications exist, and entire industries like ISI [47] have arisen from scientists' and universities' avid interest in citations. At least two kinds of networks can be reconstructed from the databases of scientific publications.

1.3.1 Co-author Networks

In discussing the ego-centered network around Erdős, we have already begun considering a (very important) node in a scientific co-author network. When we consider the *total* network of scientists in a field with links defined as in the Erdős case, the resulting network is denoted a co-author network. A visualization of this network can be found in Figure 3.1 (a) in Chapter 3. Like the actor collaboration network, the network of scientific co-authorship is an example of an affiliation network. Needless to say, this network is especially interesting as a *social* network, since two authors are likely to be well acquainted when co-authoring a paper, making this network an important example of a true social network with many actors¹¹. Whether or not the movie-collaboration is a true social network is questionable: Actors usually do not choose to work together in a given film, but are casted by producers or directors.

Mark Newman was the first person to consider scientific collaboration networks [48, 49, 50]. Newman's investigations are interesting, first of all because one of the networks he considers is the SPIRES network, which is the topic of this thesis. Secondly, Newman's work is interesting because it focuses on aspects of the collaboration network that are closely related to upcoming work in the present thesis; therefore, Newman's most important results will be reviewed in some detail in the following.

The networks that Newman investigates stem from 4 different publicly available databases, where the data used is restricted to the years 1995-1999. Upon investigating the properties of these networks, the first thing that springs to mind is that the clustering coefficients for

¹¹As mentioned above, the lack of data has been one of the problems in using statistical physics to analyze real *social* networks.

| Database | # of papers |
|---|-------------|
| Los Alamos e-Print Archive (self-submitted physics preprints) | 98,502 |
| Medline (biomedical research) | 2,163,923 |
| SPIRES (Stanford Public Information REtrieval System) | 66,652 |
| NCSTRL (Networked Computer Science Technical Reference) | 13,169 |

Table 1.1: The databases that Newman considers in [49, 50, 48].

all 4 networks, are extremely high. They range from 0.7 in the case of SPIRES to 0.3 – 0.4 for the other databases. If the reader recalls that $C_{rand} = \mathcal{O}(n^{-1})$ for a random graph and $C_{connected} = 1$ for a fully connected graph, it becomes clear that the co-authorship networks are truly *very* small worlds. Also, the average path lengths are short—from 4.0 for SPIRES to 5.9 in LANL, the only exception being NCSTRL with an average path length of 9.7.¹² As for the topologies of the database, Newman makes several interesting discoveries:

Number of Papers per Author

The average number of papers per author is between 3 and 6 over the 5 year period, the only exception being SPIRES with an average of 11.6 papers per author¹³. When these distributions are plotted, the tails of the Medline and NCSTRL data sets follow power laws¹⁴. The corresponding exponents for Medline and NCSTRL are $\alpha_{medline} = 2.86$ and $\alpha_{NCSTRL} = 3.41$, respectively. In the case of the Los Alamos Archive, he found that the probability distribution of papers-per-author is well-described by an exponentially truncated power-law,

$$P_{lanl}(k) \sim k^{-\tau} e^{-\frac{k}{\kappa}}, \quad (1.5)$$

where τ and κ are constants. The SPIRES database is not fitted very precisely by either function, but it has a violent bump at about 100 papers.

Number of Authors per Paper

The distributions of the number of authors per paper, is well described by power-laws, $P_{authors}(k) \sim k^{-\alpha_{authors}}$ for all 4 databases—although the exponents differ substantially: 6.2 (Medline), 3.3 (Los Alamos Archive), 4.6 (NCSTRL), and 2.2 (SPIRES).

Number of Collaborators per Author

This is the degree distribution for the co-author network. In the case of the SPIRES data, this distribution is very well described by a power law with slope 1.20, but the average number of collaborators for SPIRES raises the question whether or not high energy physics can be

¹²Newman argues that this number is artificially inflated, because NCSTRL has a poorer coverage of its subject matter than the other databases.

¹³This is of course because of the huge collaborations in experimental high energy physics - much more will be said on this subject as a part of the investigation carried out in this thesis.

¹⁴These power laws are in rough agreement with a power law distribution of papers-per-author found by Alfred Lotka, in 1926 [51], known to sociologists and bibliometricians as *Lotka's law*. In a data set completely compiled by hand, Lotka found a power-law relationship for the number q of papers per author $P_{lotka}(q) \sim q^{-\alpha_{lotka}}$ with slope $\alpha_{lotka} = 2$.

regarded as a *social* network, in the sense mentioned earlier, because of its extremely high average number of collaborators: It is unlikely that many authors know $\langle k \rangle_{\text{SPIRES}} = 173$ people well.

For the three remaining databases, the graphs display some curvature. This may be due to the limited time window that creates the cut-off, or it may, simply, be the dynamics intrinsic to the systems. Two power laws with slopes 2 (for the low regime) and 3 (for the high regime) seem to provide a reasonably good fit.

As mentioned in connection with the Erdős network, Barabási's group [26,27] has worked on the network of mathematicians¹⁵, but I have chosen to review Newman's work only, for several reasons. Newman's material includes SPIRES, his investigations were the first of their kind, and Newman has better statistics, making the Barabasi work (although important in its own right) redundant in this context.

1.3.2 Citation Networks

Given the level of interest in citation data and complex networks demonstrated above, surprisingly few serious studies of citation networks have been performed by physicists. In a citation network, the nodes are scientific publications and the edges represent citations. The structure of this network on the paper level is very similar to the internet and, like the internet, it is also directed. References point *out* from a node (to another node) and citations point *to* the node. Information on the out-going distribution is difficult to get a hold of¹⁶, while the citation information (the incoming distribution), as I mentioned earlier, is vigorously documented.

A few investigations of the structure of citation networks have been made in the past. In 1957, Shockley [52] argued that the publication rate for the scientific staff at Brookhaven National Laboratory was described by a log-normal distribution. In 1998, Laherrere and Sornette [53] suggested that the probability of an author (node) to have k citations (edges), $P(k)$, of the 1120 top-cited physicists from 1981 to 1997, is described by a stretched exponential ($P(k) \propto \exp[-(k/k_0)^\beta]$, $\beta \approx 0.3$). Also, in 1998, Redner [54] considered data on papers published in 1981, from journals catalogued by the ISI [47], as well as, data from Phys. Rev. D vols. 11-50. Redner concluded that the large- k citation distribution is described by a power-law such that $p_k \propto k^{-\alpha}$ with $\alpha \approx 3$. In 1999, Tsallis and Albuquerque fitted Redner's data to a slightly different curve $\sim (k + \text{const})^{-\alpha}$ [55]. Sven Bilke and Carsten Peterson calculated the so called *spectral dimension*, d_S , which reflects diffusion processes in the corresponding graphs, for the SPIRES citation network [56].

1.4 The SPIRES Database

In this thesis, the goal is to shed new light on the citation network. The outset is that the statistical material is of a much higher quality than in the papers mentioned above. The ISI data set studied in [54] is materially larger (783,339 papers) than the present SPIRES data set. However, the ISI data used by Redner, contains papers published in a single year in a variety of scientific disciplines (including medicine, biology, chemistry, physics, *etc.*). There are neither *a priori* arguments nor data to indicate that citation patterns in these fields are

¹⁵The network of neuro-scientists were also included in their investigations.

¹⁶It is of course easy to read the references at the end of a given paper, but to actually get a hold of *all* lists of references, that is the tricky part.

sufficiently uniform to justify their treatment as a single data set. On the contrary, Newman's co-author papers [48, 49, 50] indicate that there are significant differences in the publishing habits of different scientific communities.

1.4.1 A Brief History of SPIRES

The SPIRES hep (high energy physics) data is collected from a well-defined area within physics, *i.e.* high energy physics, and the database itself is one of the oldest and most comprehensive in all of science: Since its foundation in 1962, the SLAC (Stanford Linear Accelerator Center) has been collecting new preprints and, as one of the world's centers for theoretical physics, it has attracted some 3000 new papers per year. In the same year DESY (Deutsches Elektronen-Synchrotron) in Hamburg, Germany, began publishing a record called 'High Energy Physics – An Index' (HEPI).

In 1967, computer scientists at Stanford University began working on a new computerized database that was designed to be able to handle (in principle) a limitless number of large bibliographical records. Come March 1968, the SLAC Library, being in possession of a large database that was perfect for testing the new system, began participating in this project. Thus, SPIRES (Stanford Physics Information REtrieval System¹⁷) was born.

What is extremely important in the present context is, that the SPIRES database allowed the SLAC librarians to add the *reference list* of all papers to the database, thus making possible the extraction of *citation data*. Further, it is important to note that, to make this information as reliable as possible, only references to published papers were included. It was only natural for the DESY and SLAC libraries to cooperate, and by June 1969, the conversion of the DESY data to SPIRES format was complete.

By 1974, SLAC and DESY¹⁸ were comprehensively collecting preprints (and by extension published articles) and cataloguing them in a single SPIRES hep database. The next important step for the SPIRES database was the 1991 creation of the LANL (Los Alamos National Laboratory) e-Print server. This allowed authors to self-publish their preprints on one common server, assigning to each paper a unique number of the form *archive/0211210*; the number signifying the 210th paper of November 2002. The unique labelling allowed systematic referencing to unpublished papers and now allows citations of preprints to be registered in the SPIRES hep database. This point is relevant to the present thesis, because conditioning the database to recognize that the e-Print and the published article is the same paper, could be problematic. This could lead to an over-estimation of author productivity and an underestimation of the number of citations per paper. The SLAC library goes to some length to avoid this problem. This review of the history of SPIRES, is based on a paper by Heath O'Connell [57].

1.5 Applications

There is yet another reason that the subject matter of this thesis is interesting to physicists: *It is about them!*. Now, it is always interesting to read scientific investigations in which the subject matter is oneself, and surely many physicists have glanced through Newman's

¹⁷SPIRES was later renamed Stanford Public Information REtrieval System.

¹⁸Later, also CERN, University of Durham, KEK, Yukawa Institute, and Fermilab participated in the collection of papers.

co-author papers [48, 49, 50], simply because they were—in a very direct way—the subject matter of the investigations. In the present work, this aspect of the investigation becomes even more interesting, since the investigation concerns the distributions of *citations* in high energy physics. As we know, the number of citations (at least in the eyes of the world) equals the quality of a given publication.

Richard Feynman once said:

Physics is like sex. Sure, it may give some practical results, but that's not why we do it.

This statement captures an important issue related to physics, *viz.* that there are two reasons to study physics. The first and boring reason is the engineering—the practical results in the quote. The other reason is the sheer pleasure of a beautiful theory, the sensation of establishing solid facts about the world; this is the reason we *really* study physics.

There is no doubt, that there is plenty of *practical* use for the charting of unknown network territory. The specific practical outcome of studying networks depends on the network in question. Studying the www will help us design new protocols for surfing the web, studying the email network will help us fight computer viruses, knowing the topology of the network of human relations will help us fight viruses like AIDS or SARS. In the case of the citation network, the practical outcome is insight regarding the fitness landscape of human excellence. This is due to the fact that the number of citations that a scientific paper receives, has been regarded as a measure of the quality of the paper¹⁹. Citations as a measure of quality has been accepted by the scientific community at least since the 1960's [58, 59].

In the beginning of this introduction, I raised the question: Why are citation networks relevant to physics? The reason this question has been raised so many times during the writing, is that seeing beyond the practical layer of the investigation of SPIRES is difficult and requires knowledge of the physics of complex systems. Throughout this introduction, I have tried to assimilate the reader into the field and demonstrated that *the tools and theories of physics are abundant here*. The physics of complex networks contains lots of 'pure physics' and I urge the reader to keep in mind, that throughout this thesis, both the practical and the (aesthetically pleasing) theoretical layer of the study of physics, pointed out in the Feynman quote, will be present in the work and should be evaluated separately.

Since this is a physics thesis, the primary focus naturally falls on the *physics* of the complex networks, but along the way, I will also touch on how these results influence the practical level, *i.e.* how the knowledge gained by the 'pure physics' approach dramatically changes how we think about quantifying scientific excellence; how the physics of complex systems is utilized to once again turn our intuitions about how the world work upside down.

¹⁹Of course, there are many complications when dealing with this subject, but throughout this thesis we shall exhibit great caution, and distinguish carefully between qualitative statements on the distribution of citations and value judgements.

CHAPTER 2

The Paper Distribution

We are now ready to begin investigating the network constituted of papers in the SPIRES database¹. Sometimes the terms ‘references’ and ‘citations’ are used inconsistently in the literature; here, the (natural) definition that SPIRES supplies on their webpage is used:

The references of your paper are those that you list at the end; they’re the previous papers you’ve cited. The citations of your paper are all the papers that mention your work, that is all the papers that have your paper in their reference list².

Thus the citation network is rather simple: The nodes are scientific publications and edges arise when one papers cites another paper. To get an intuitive feel for the citation network, recall that the network of papers shares certain properties with the world wide web. The analogy is: paper \sim webpage, reference \sim outgoing link, and citation \sim incoming link. As we will learn from the following, another common property is that just like there is a massive number of webpages that nobody reads, the vast majority of papers in high energy physics are not cited by anybody. The subcultures that are found on the www have a clear analogy in the subfields of the citation network. With the distinction between references and citations, it is also clear that the citation network shares the property of being a directed network with the www—even on the papers-citing-papers level, discussed in this chapter. As it is the case for the www, this directedness makes calculating a clustering coefficient for the citation network impossible. There are, however, also important differences when comparing the network of publications to the www: Scientific papers are printed on paper, and therefore the citation network instantly freezes in a tree structure when a paper is published, *cf.* Figure 2.1; this is in stark contrast to the web, a network that is characterized by fluxus; the topology of the internet is constantly changing, constantly being modified. When we include the authors of papers in the analysis in the following chapters, we will see further differences between the network of scientific publications and the www.

In this chapter, we will first investigate the functional form of the citation distribution for papers in SPIRES—*i.e.* the probability $P(k)$ that a paper has k citations. We will proceed

¹A large portion of the contents of this chapter is presented in a paper that I have co-authored with Benny Elley Lautrup and Andrew Dumont Jackson in 2002 (to be published in Physical Review E) [3].

²*cf.* <http://www.slac.stanford.edu/spires/hep/references.html>.

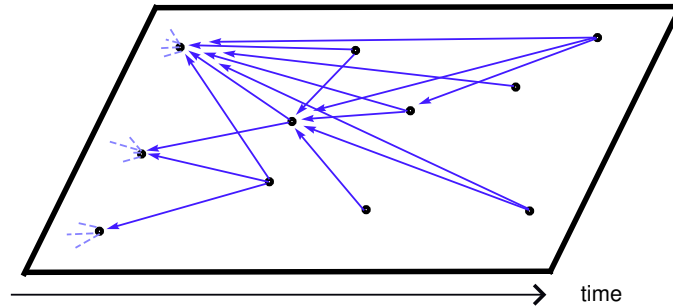


Figure 2.1: An excerpt of the network of papers in SPIRES. The \bullet 's are papers, and (directed) links are represented using arrows. Note the time-line. In this illustration the tree structure is clear; papers can only link back in time, and once they are published, no new outbound links can arise. Another type of representation of the network of papers can be found on the front page of this thesis.

to verify that the SPIRES database is indeed a very homogeneous database, by considering the citation distributions for papers, subfield by subfield. Finally, we will demonstrate the extreme improbability that the citation records of selected individuals have been obtained by a random draw from the resulting distribution.

2.1 Basic Properties

The SPIRES database³ contains 501,531 papers. For a considerable fraction of these papers, however, no citation information is available: 196,432 papers are preprints and conference proceedings for which no citation information is available, other papers seem to have been removed from the database, and in other cases no subfield information is available. All in all we are left with 281,717 papers for which both subfield designations and citation information are accessible. This number corresponds to about 56% of the database.

2.1.1 A Raw Plot

In Figure 2.2, an 'atomic' histogram of the citation distribution of the total data set is displayed. Note the log log scales. This (normalized) histogram can be interpreted as the normalized probability $P(k+1)$ that a paper has $k+1$ citations. The two straight lines with slopes -1.29 and -2.32 , respectively, show that the probability distribution is well described by two power-laws: $(k+1)^{-1.3}$ for $0 \leq k \leq 49$ and $(k+1)^{-2.3}$ for $k \geq 49$.

2.1.2 Subfields

SPIRES is divided into 5 subfields: Theory (159,946 papers), Phenomenology (68,549 papers), Experiment (28,527 papers), Instrumentation (19,637 papers), and Review (5,058 papers); in total corresponding to the before mentioned 281,717 papers.

So at this point, a question presents itself: Is it justified simply to collapse all of the citation data from the various subfields of high energy physics into a single data set, as it is

³For detailed information on how and when the data used in this section is collected, the reader is referred to Appendix A.1. A brief history of the SPIRES database can be found in Section 1.4.

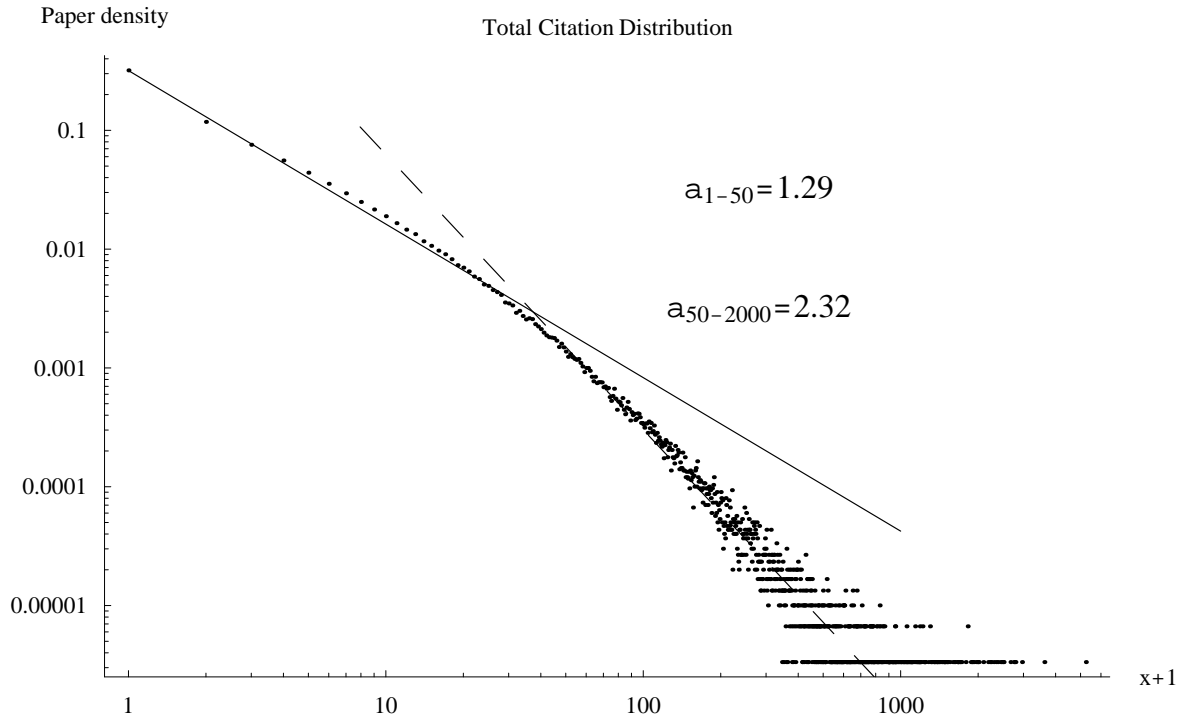


Figure 2.2: An ‘atomic’ histogram of the citation distribution of the total data set showing the normalized probability, $P(k+1)$, that a paper has $k+1$ citations. The straight lines in the low and high citation regimes have slopes -1.29 and -2.32 , respectively. Note the logarithmic scales.

done in Figure 2.2? Answering this question merits an investigation. As a first guess, one

| | Theory | Phenomenology | Experiment | Instrumentation | Review | Total |
|-----------------|-----------|---------------|------------|-----------------|---------|-----------|
| Papers | 159,946 | 68,549 | 28,527 | 19,637 | 5,058 | 281,717 |
| Citations | 2,362,400 | 1,088,269 | 453,803 | 55,927 | 169,165 | 4,129,564 |
| Mean | 14.77 | 15.88 | 15.91 | 2.85 | 33.45 | 14.7 |
| Median | 2.12 | 3.75 | 3.17 | 0.81 | 4.61 | 2.27 |
| Un-cited (%) | 29 | 21 | 27 | 62 | 22 | 29 |
| ≤ 10 (%) | 75 | 69 | 69 | 94 | 63 | 74 |
| ≥ 50 (%) | 6.0 | 7.2 | 7.4 | 0.74 | 14.1 | 6.2 |
| ≥ 1000 (‰) | 0.54 | 0.32 | 0.32 | 0 | 2.5 | 0.46 |

Table 2.1: This table summarizes some important basic statistics of the SPIRES database and the sub fields of high energy physics into which it is divided. Note that the ‘Total’ data is obtained directly from the subfield data.

would expect the SPIRES database to be relatively homogeneous, since it contains papers published by a homogeneous population of authors publishing on a very well bounded subject matter, *viz.* high energy physics. In the following section, we will discuss the question of SPIRES’ homogeneity in some detail.

2.2 Homogeneity

2.2.1 Skewed Distributions

The first thing that springs to mind, when inspecting Table 2.1 is the approximately 29% of the total papers that are un-cited, and the 75% of the total number of papers that have less than or equal to 10 citations! Note that no corrections have been made for self-citations; the removal of self citations would make the fraction of un-cited and minimally cited papers even higher. In the other end of the citation distribution, we find that only about 6.0% of the papers have more than 50 citations and that only 131 papers of the total data set have 1000 or more citations $\approx 0.54\%$. For the data divided into subfields, the situation is analogous—*cf.* Table 2.1.

For the total data set, the mean is 14.7 citations per paper, whereas the median is 2.3 citations per paper, and for the subfields a similar discrepancy between the mean and median can be observed; again, the reader is referred to Table 2.1 for the exact numbers. Thus, we can make the—strictly speaking meaningless—point that *a paper with the average number of citations is substantially more cited than the average paper*, where the first instance of ‘average’ refers to the mean value of citations, and the second instance referring to the everyday use of ‘average’, meaning the type of paper one comes across more often, *i.e.* the median paper. The large factor difference between mean and median indicates that the distributions are highly skewed with long tails: A small fraction of highly cited papers accounts for a significant part of the total number of citations.

It is natural to expect this type of statistics from the power-law structure seen in Figure 2.2, but some numbers on the highly cited tail merits mention, in order to quantify just *how* different the populations of minimally- and highly cited papers are. Approximately 50% of the citations in the database are generated by the top 4% of papers; in contrast, the lowest 50% supply only 2% of the total citations. The rates of citation production by these two parts of the data set, differ by a factor of approximately 310. An interesting number related to these, is the 18% of the *papers* that produce $(100 - 18)\% = 82\%$ of the *citations* in the database.

2.2.2 Prior Expectations

It is easy to think of mechanisms that could cause the topologies of the different subfields to differ in a variety of ways. This is already evident from Table 2.1, where it seems obvious that even though both instrumentation and review data sets share the power law properties (heavy tail, large factor between mean and median, *etc.*), it is also clear that their statistics are radically different from the other subfields.

The subfield that intuitively differs the most from the rest of the SPIRES, is the experiment subfield. Experiments in high energy physics are comprehensive and manpower extensive—it is not unusual that experimental papers have more than a thousand co-authors. As a consequence of this, program committee approval is more or less the same as a pre-review of the work. Compared to the theory and phenomenology subfields, where a paper typically has one to three authors and does not require expensive experiments, it is easy to imagine papers being written that would not survive this kind of pre-reviewing. Thus, it is natural to expect a larger probability for minimally cited papers in the theory and phenomenology subfields, while the experiment subfield is expected to have rather fewer minimally cited papers.

Review papers are often commissioned by journals and are often written by recognized experts. Thus, publications that are part of this subfield could easily be imagined to have a higher probability of receiving many citations. Papers in the instrumentation subfield could be imagined to be minimally cited for a number of reasons. First of all, instrumentation papers are usually specific to an experiment and, therefore, not of general interest; that instrumentation papers are specific to certain experiments also implies that they become dated, as technology evolves—this is not the case for more fundamental theoretical or empirical discoveries; papers containing these, naturally seem likely to keep receiving citations. With all this in mind, the importance of investigating the subfield distributions and pinpointing homogeneities and differences is clear.

2.2.3 Visual Comparisons

In Figure 2.3, the binned subfield distributions are plotted. At first glance these look very similar; the two power-law structure is evident in all subfields except maybe for the instrumentation data. Looking closer, however, one starts to notice differences. The odd one out is clearly the instrumentation plot; this data looks like a single power-law, the characteristic two power-law structure, seen in the other subfields, is absent. Furthermore, this relatively large data set (19,637 papers) is minimally cited. The highest cited instrumentation paper has 627 citations; the result is a steep power-law with an usually high probability for minimally cited papers. Consulting Table 2.1, this suspicion is amply confirmed—the instrumentation data stands out in a variety of ways that are described in Section 2.2.4

The review data has about the same slope as the phenomenology and theory subfields, for the minimally cited regime. This indicates that the minimally cited papers, in these fields, have approximately the same citation rates. For the highly cited regime, the slope of the distribution of the review subfield is radically flatter than the same slope for any other data set. This means that the density of highly cited review papers is considerably higher than for any other subset. Thus, our expectations towards the instrumentation and review subset are confirmed.

A surprise with regard to our *a priori* considerations, is that the experimental and phenomenological subfields are very alike. This is a surprise for two reasons: Firstly because we expected the experiment subfield to be markedly different from the rest of the database—which it is certainly not. And, secondly, because we expected the phenomenology data to be almost indistinguishable from the theory data, because the *modus operandi* in these two fields are similar (no pre-reviewing, no big experiments, *etc.*), which we expected to be able to see in the distributions of papers.

The theory subfield does not have as pronounced a ‘dent’ in the distribution around 50 citations, as it is the case for the experiment and phenomenology subsets. This is in part because the slope of the minimally cited regime for the theory data is steeper than what is the case for the experiment and phenomenology data, and because the slope of the highly cited part of the distribution is less steep.

2.2.4 Quantifying Differences

Indications of the differences between the five categories can also be seen from Table 2.2. This table concerns the probabilities for the minimally cited regime. The probability of having ≤ 4 citations is 59.9%, 53.6%, 51.2%, 47.7% and 86.5%, for theory, experiment, phenomenology,

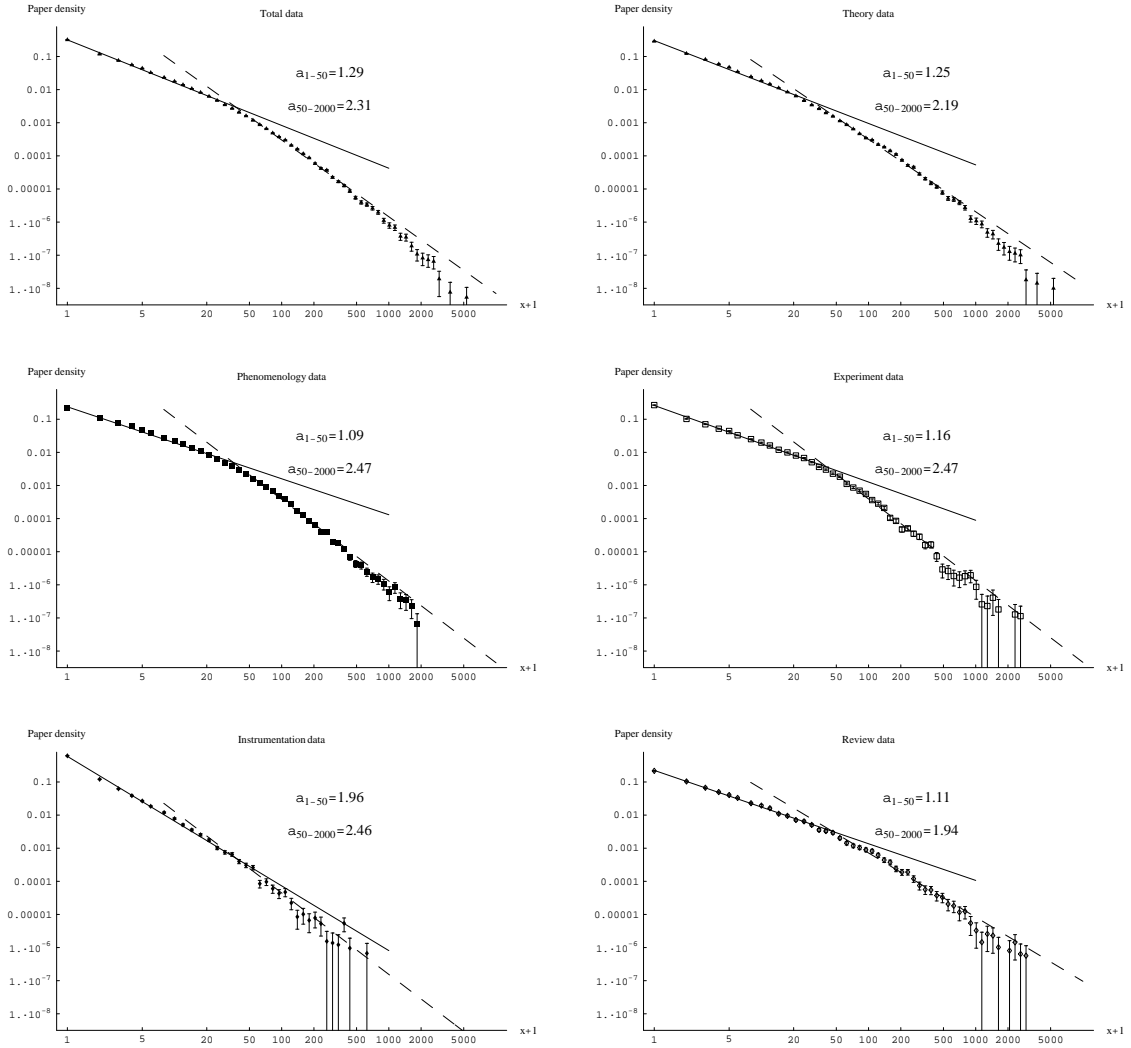


Figure 2.3: Array of binned subfield distributions. The plots are normalized and can be interpreted as probability distributions. Along with the distributions, I have plotted straight lines for the minimally- and highly cited regime, using the dashed ‘—’ and ‘- - -’ respectively. The slopes of the lines are plotted along with each subfield.

review, and instrumentation, respectively. These numbers corroborate with the conclusions drawn from inspecting the plots. The fraction of minimally cited papers is clearly smaller for the review subfield in comparison to the total data set—but, the effect is not dramatic. Instrumentation, however, stands out. The probability that an instrumentation paper will receive ≥ 5 papers is *almost 3 times smaller than for the rest of the data set*. When we look at the differences between citation probabilities in theory, experiment, and phenomenology, we find that they are surprisingly small. The probability of having 0 citations is virtually the same for theory and experiment and a little lower for phenomenology. Regarding the probability of having 2, 3, and 4 citations, the theory data is consistently higher than the experiment

| | $P(k=0)$ | $P(k=1)$ | $P(k=2)$ | $P(k=3)$ | $P(k=4)$ | $P(k \leq 4)$ |
|-----------------|----------|----------|----------|----------|----------|---------------|
| Theory | 0.2884 | 0.1226 | 0.0815 | 0.0590 | 0.0472 | 0.5987 |
| Phenomenology | 0.2150 | 0.1103 | 0.0762 | 0.0618 | 0.0488 | 0.5364 |
| Experiment | 0.2677 | 0.1023 | 0.0704 | 0.0518 | 0.0441 | 0.5122 |
| Instrumentation | 0.6169 | 0.1206 | 0.0622 | 0.0385 | 0.0267 | 0.8650 |
| Review Articles | 0.2167 | 0.1038 | 0.0670 | 0.0496 | 0.0403 | 0.4775 |
| Total | 0.2901 | 0.1171 | 0.0775 | 0.0574 | 0.0458 | 0.5877 |

Table 2.2: The probability of a paper in the SPIRES database having k citations for $0 \leq k \leq 4$ as a function of subfield. The total number of papers in each subfield is: 159,946 (theory), 68,549 (phenomenology), 28,527 (experiment), 19,637 (instrumentation), and 5,058 (review articles). The ‘total’ data entries are obtained directly from the subfield data. The total number of papers in the data set is 281,717.

and phenomenology data sets—that are almost identical, although the probabilities of having a minimally cited phenomenology paper is a *little* higher, than for the experiment data set. As stated above, the probability of having ≤ 4 citations is 59.9, 53.6, and 51.2 for the three data sets, respectively. This illustrates the large statistical weight of the first data points; note, for example, that the medians of these three subfields are 2, 12, 3.17, 3.75 for the theory, phenomenology, and experiment subfields, respectively. The 3 subfields are explicitly compared in Figure 2.4.

Table 2.1 allows us to compare properties of the data sets taken as wholes, and we find that the trends seen in the first 4 citations are supported here. Theory, experiment, and phenomenology are very close to each other. Taking the entire distribution into account, the instrumentation and review data set distinguish themselves as having properties that are quite different from the total data. Recall that a stunning 61% of all instrumentation papers are un-cited, resulting in a median of 0.81 (!), which is markedly lower than the median of the entire data set. With regard to the minimally cited end, the review data is reasonably close to the total set.

Discussing the entire range of citations, we find that our expectations with respect to the review data were correct. Approximately 14% of review papers have ≥ 50 citations—compared to 6.2 percent for the total set. The 3% of review papers with ≥ 1000 citations is also significantly larger than the probability of 0.05% for the complete data set. For the instrumentation papers, the opposite picture is being drawn. Only 146 of the 19,637 instrumentation papers ($\approx 0.7\%$) have 50 or more citations. No instrumentation papers have more than 1000 citations⁴.

In short, instrumentation and review papers, which account for some 9% of the full data set, clearly follow different citation distributions. This can reflect a different underlying dynamical picture for citations in these categories; it can also be an indication that review papers have a higher average quality and instrumentation papers a lower. Whatever the explanation, these two small categories will be excluded from further consideration. As was emphasized in Section 1.5, it is clear that any decision to use citation data as a measure of scientific ‘quality’ should not be made so lightly. Ultimately, however, it must be based on a

⁴The most cited instrumentation paper has 627 citations. It was written by the CDF Collaboration (227 authors).

subjective evaluation of the relative quality and importance of papers published in the various categories.

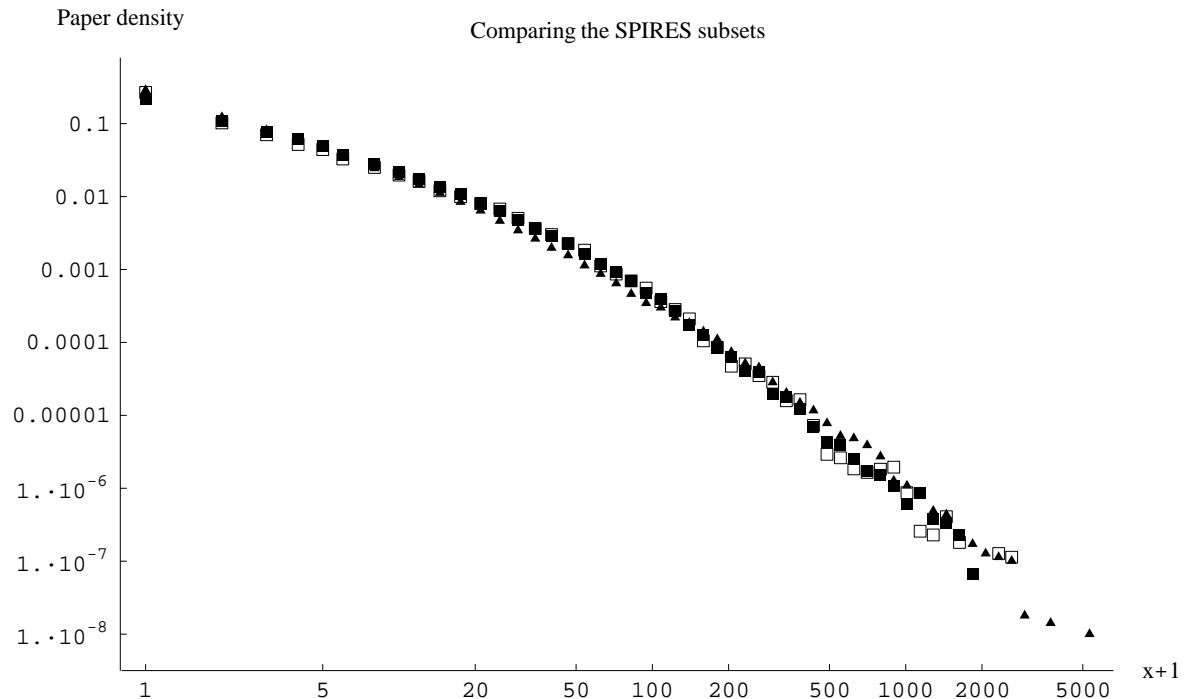


Figure 2.4: (Citation probabilities for the categories theory (\blacktriangle), phenomenology (\blacksquare), and experiment (\square)).

2.2.5 Tests?

The homogeneity of the theory, experiment, and phenomenology subfields is supported by the binned and normalized histograms in figure 2.4. On the logarithmic scales, the three subsets are virtually indistinguishable over the entire range of 0 to 5000 citations. Given that the distribution spans almost 8 orders of magnitude, this is a very good agreement. In order to evaluate whether or not the differences between the 3 remaining subsets are ‘statistically significant’, one’s first impulse would be to assign errors to each bin proportional to the square root of the number of papers in each bin and perform a χ^2 -fit. On second thought, however, it is clear that this exercise would be meaningless.

The χ^2 -test is based on the assumption that the data in the various bins is statistically independent. For the SPIRES data set this assumption can be demonstrated to be false—a large part of this thesis consists in doing exactly that *cf.* Chapters 3, 4, 5. However, merely pointing to an author whose publication record is highly correlated (for the extreme example in high energy physics, take Edward Witten) is sufficient to show that the data indeed contains ‘longitudinal’ correlations. Keep in mind that our study of the citation network is partially motivated by a wish to study scientific excellence. We believe that there is a positive correlation between the intrinsic quality of a scientific paper and the number of citations which it receives. Also, we believe that ‘good’ papers are produced by ‘good’ scientists. One

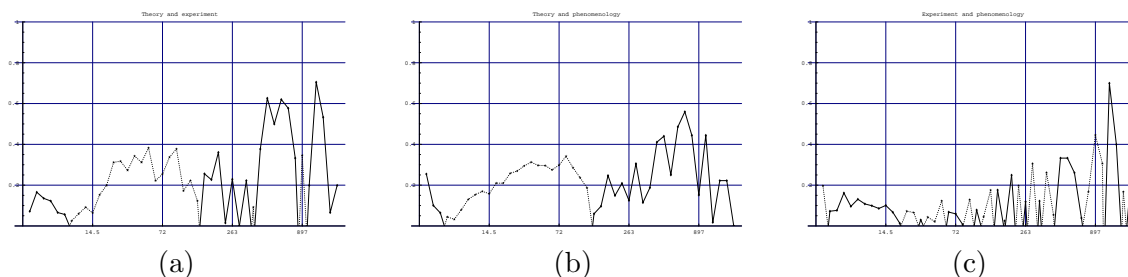


Figure 2.5: Percentage differences. The x -axis shows number of citations, bin by bin and the y -axis shows the percentage difference. (a) Theory and Experiment compared. (b) Theory and phenomenology subfield data compared. (c) This shows the differences between experiment and phenomenology.

way of quantitatively comparing the data, is to look at the percentage differences between the data sets. In Figure 2.5, the percentage differences between (a) theory and experiment, (b) theory and phenomenology, and (c) experiment and phenomenology are plotted. These plots emphasize the differences we could observe directly in Figure 2.4. Theory and experiment (Figure 2.5 (a)) are in 20 – 40% disagreement around 20 to 100 citations, corresponding to the fact that the ‘dent’ in the theory data is not as pronounced as is it in the experiment and phenomenology data. Also, the disagreement in the data from around 500 citations and up, visible in Figure 2.4, is clear from these plots—here the disagreement is between 40 and 50%. Almost the same structure is evident, when comparing theory and phenomenology as it is done in Figure 2.5 (b). In case of the comparison between the phenomenology and experiment subfields, the agreement is excellent, *cf.* Figure 2.5 (c). For the first part (0-250 citations) of this figure, the percentage differences are very small (5-15%)—only for the highly cited papers (1000+) does the disagreement grow to more than 40%.

It is, however, easy to understand the origin of the ‘large’ percentage differences for the highly cited papers—and, thus, why they are not important. They are due to the large ‘arm’ of the normalization. The citation counts in these bins are small (1 – 10 citations), and since they are divided by the total number of papers in each of the subsets, small fluctuations are blown out of proportion. This effect also accounts for some of the differences between the large theory subset and the two smaller subsets.

In summary, it is clear that dividing the data into subfields does reveal differences between the subfields. On the basis of Figure 2.4 the conclusion is, however, that the consistency of the theory, phenomenology, and experiment subfields is sufficient for many applications—one could argue that this is a case of a picture (Figure 2.4) saying more than a thousand words. In the rest of this chapter, we will be discussing a final data set of 257,022 papers consisting of the papers from these three subfields; accordingly, the two small fields of review and instrumentation will be disregarded. The resulting distribution is shown in Figure 2.6.

2.2.6 A Potential Source of Inhomogeneity

So far, the form of the distributions of citations of the different SPIRES subfields have been checked for homogeneity. But, there is another potential inhomogeneity in the SPIRES database that should be mentioned here. The distribution of the number of authors who have written x papers is a monotonically decreasing distribution of x . (As we will see in Section 3.3.2, it is actually a power-law distribution.) In the case of the theory subfield, about

91% of the authors have published less than 20 papers! This is presumably because of the relatively large number of physicists that leave academic physics just after, or within a few years of finishing their Ph.D.'s. Now, one could imagine that the citation distribution would drastically change if ‘minimally publishing’ authors were removed. Exactly how the distribution changes, when the ‘minimally publishing’ authors are removed is discussed in great detail in Chapters 3 and 4. For now, however, it is sufficient to state that the distribution is relatively independent of the removal of minimally cited authors—the differences are similar to those of Figure 2.4; the two power-law structure is intact and the considerations found in the rest of this section are certainly relevant.

2.3 Form of the Distribution

Resting assured that the bulk of the database is homogeneous, we can take a closer look at the form of the distribution. It is clear from the figures that the citation distribution is not described by a single power-law over the entire range of citations. As was stated earlier, it is well approximated by two independent power-laws: One in the low ($k \leq 50$) citations regime, and one in the highly cited ($k \geq 50$) regime, with $\alpha_{low} \approx 1.20$, and $\alpha_{high} \approx 2.31$. If we

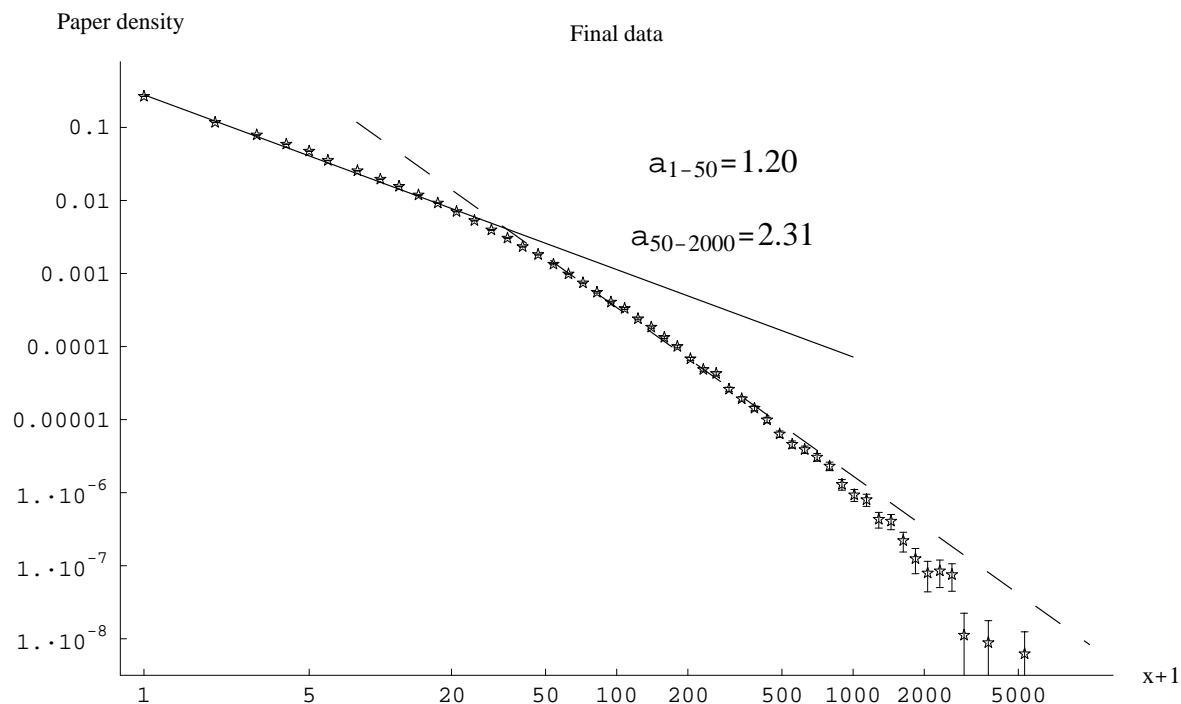


Figure 2.6: A binned histogram of the total data set without review and instrumentation papers.

set the relative normalization such that the two fits are equal in $k = 50$, and set the global normalization such that a total probability of 1 is ensured, the data is reproduced with a surprisingly high accuracy.

It is natural to expect that the different power-laws are caused by different dynamics in the two regimes. In the low k regime, the bulk of the papers are ‘dead’, in the sense that they

have not been cited in the past year(s); these dead papers will most likely never be cited again. This dead population constitutes the vast majority of the database. In the minimally cited regime, there are, of course, also vigorous ‘young’ papers of high quality that are collecting citations at a steady rate, making their way through the population, waiting to join the highly cited regime. In the highly cited regime, virtually all papers are ‘alive’, with the oldest of them collecting citations at a regular basis. The temporal evolution of the database will be discussed in detail in Chapter 6, when I discuss a new model for the citation network. In summary: The tail is where the action is.

2.3.1 The Asymptotic Tail

Since one of the major points of interest in this thesis is investigating scientific excellence, and since I have argued that the highly cited end of the distribution is where the dynamical papers of high quality are found, it is natural to take a closer look at the asymptotic tail of the distribution of citations.

In [54], Redner argues that the asymptotic tail of the distribution of citations⁵ is well described by a power-law with a slope of -3 . To this end, Redner uses a *Zipf* plot. The Zipf plot was introduced by Harvard linguistics professor, George Kingsley Zipf [60]⁶. A Zipf plot is a plot of the n th most ranked paper versus the number of citations of this paper, Y_n . The most cited paper is assigned rank 1.

The intuitive reason why the Zipf plot is well suited for analyzing the highly cited end of the distribution, is that it provides a much higher resolution of this part of the plot. On the log log scale, the Zipf plot places the high citation data in the beginning and, thus, it is not as compressed as in the plots of $P(k)$ vs. k in Figures 2.2, 2.3, 2.4, and 2.6. Figure 2.7 is a Zipf plot of the final data set. It is clear from the figure that the asymptotic power-law, $1/k^3$, that Redner found is not present in the SPIRES data. On the contrary, Figure 2.7 clearly shows that the tail of the final data set is *not* described by any asymptotic power-law. The same conclusion is clear from Figure 2.6, where the second power-law tracks the data accurately through four decades, after which the data begins to cut off.

Even though the high citation end of the population is sparsely populated, it is possible to present quantitative indications of this cut-off. If the power-law from Figure 2.6 was valid for arbitrarily large k , as proposed by [54], we would expect to find 33 papers with more than the maximum of 5,242 citations, actually found in the data set. The most cited paper should have about 55,000 citations. Or, put inversely, if we assume that an asymptotic power-law is valid, then the probability of drawing 257,022 papers at random from this distribution, with no paper having more than 5,242 citations is approximately 10^{-14} .

⁵In [54] papers from Physical Review and from ISI are used to reach this conclusion. See Section 1.3 for further details.

⁶Zipf (1902-1950) was one of the first people to recognize the ubiquity of power-laws. The most famous example of Zipf’s law is the frequency of English words. The second example Zipf included in his book, was the population of cities (or population of communities). The population of the cities plotted as a function of rank (the most popular city is ranked number one, etc) is a power-law function with an exponent close to 1. The income or revenue of a company, as a function of the rank is also an example of the Zipf law (also in Zipf’s book [60]).

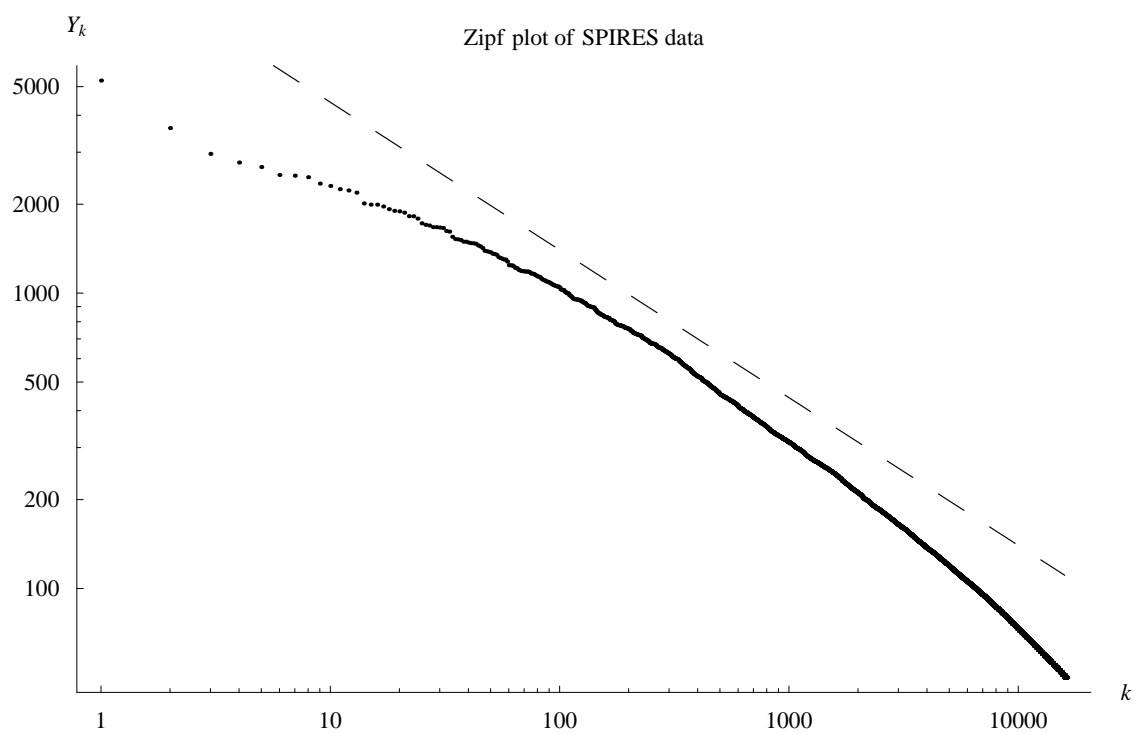


Figure 2.7: A Zipf plot of the citation distribution. For visual reference a line of slope $-\frac{1}{2}$, corresponding to $\alpha = 3$, is also plotted.

Scientific ‘Saints’

There is a simple way to explain the cut-off for the large- k data that seems reasonable for a data set—like SPIRES—which contains a significant number of truly important papers. It is a fact that papers of high quality and fundamental importance can literally be ‘canonized’; fundamental papers can pass into the received wisdom of physics that no longer requires citation. An everyday example of this is that many high energy physicists publish papers about ‘Goldstone Bosons’, but only a few feel the need to cite the original papers by neither Bose nor Goldstone. To state an even stronger example, the attentive reader would surely stop to consider what special point was being made, when citing Einstein on special relativity [61]. In the absence of a cut-off, a paper like [61] should have been cited by approximately $55,000/257,022 \approx 21\%$ of the papers in the database—this does not seem unreasonable.

2.4 The Power of Excellence

So far, we have focused on the properties of the SPIRES database, seen as a network. In this section, I will consider an application of this work; that is, using the SPIRES data to find a new measure of scientific excellence. The reason this is interesting, will be evident from the following. This measure, r , quantifies the ‘improbability’ of excellent authors’ citation records being drawn at random from the citation distribution.

| Paper category | Citations | Probability |
|-------------------|-----------|-------------|
| Unknown papers | 0 | 0.267 |
| Less known papers | 1–9 | 0.444 |
| Known papers | 10–49 | 0.224 |
| Well-known papers | 50–99 | 0.0380 |
| Famous papers | 100–499 | 0.0250 |
| Renowned papers | 500+ | 0.00184 |

Table 2.3: The search option ‘citation summary’ at the SPIRES website returns the number of papers for a given author in the categories of this table. The probabilities of getting citations in these intervals are listed in the third column.

The ‘citation summary’ option in the SPIRES database returns the number of papers for a given author with citations in each of six intervals. These intervals and the probabilities revealed by our analysis, that papers will fall in these bins, are given in Table 2.3. The probability, P , that an author’s actual citation record of M papers was obtained from a random draw on the citation distribution, is readily calculated by multiplying the probabilities of drawing the author’s number of papers in the different categories, m_i , and correcting by the number of permutations

$$P = M! \prod_i \frac{p_i^{m_i}}{m_i!}. \quad (2.1)$$

If a total of M papers were drawn at random on the citation distribution, the most probable

result, P_{\max} , would correspond to $m_i = Mp_i$ papers in each bin. The quantity⁷

$$r = -\log_{10}(P/P_{\max}) , \quad (2.2)$$

is a useful measure of this probability which is relatively independent of the number of bins chosen. Since r provides completely objective information about the probability of drawing a given citation record at random, given knowledge of citation patterns in that field; *r is particularly well-suited for comparisons between fields.* It is equally meaningful to calculate r for authors who publish in several fields—this idea is truly novel—never before has any attempt been made to compare authors across field boundaries.

The careful reader is right in thinking that a *caveat* should be voiced here. There are intrinsic problems associated with the leap from the improbability of a given author's citation record to drawing conclusions regarding author quality. This leap requires certain assumptions which cannot be tested. For example, in order to compare citation records from the instrumentation category with those in the remainder of our data set, it is necessary to make some a priori assumption about the relative intrinsic quality of the two data sets. While the 'democratic' assumption of equal intrinsic quality is easiest, it may or may not be accurate. In a Bayesian sense, it is necessary to establish a prior distribution.

2.4.1 Examples of the Application of r

Consider the following two authors in the SPIRES database. Author A has a total of 201 publications with 18, 70, 82, 22, 9, and 0 publications in each of the bins above and an average of 26 citations per article. Author B has a total of 178 publications with 19, 79, 58, 10, 9, and 3 publications in each bin and an average of 46. A simple calculation reveals that $r = 17.8$ for Author A and 9.9 for Author B.

The minimum value of r is evidently 0. The maximum value of r , in the current data set, is found for Author C, who has a total of 252 publications with 5, 25, 47, 37, 102, and 36 publications in each of the bins above and an average of 242 citations per article. This leads to a vastly improbable value of $r = 188.4$. With a total of 61,062 citations, Author C accounts for more than 1.5% of all citations in the data set. There are also indications of less favorable correlations. Author D has a total of 41 publications with 18, 23, 0, 0, 0, and 0 in each of the bins above and an average of < 1 citation per article. The resulting value of $r = 4.43$ underscores the fact that an improbable citation record is not necessarily a 'good' one. This is a problem that can be remedied. If information regarding all author publication records is available, we can construct a much more precise measure of scientific excellence that takes this effect into account. This is done in Chapter 5.

Given the total population of authors in SPIRES, these numbers offer an objective indication of the extreme improbability that the citation records of Authors A, B, and C were drawn at random. These examples are far from exceptional. There are strong correlations in the citation data and these merit quantitative study. The differences between the Authors A

⁷Note the interesting connection to information theory; the quantity r can be identified as the relative entropy d (also known as the Kullback-Liebler distance). The relative entropy is defined for two probability distributions, $\{p_k\}$ relative to $\{q_k\}$, as $d = \sum_k p_k \log_2(p_k/q_k)$. To make the connection to r , simply insert Equation (2.1) in Equation (2.2), and set $q_k = p_k^\xi$, where $\xi = (m_k/p_k + 1/(M-1))$ (the change of base of the log results in a constant factor $\log_2(10)^{-1}$). This identification ensures us that r has many important mathematical properties, for example, it is positive and equals zero if and only if $p_k = q_k$, and furthermore, it is a convex function of p_k . More on this interesting subject can be found in [62,63].

and B can appear surprising at first glance and they emphasize the importance of *a priori* criteria.

Although Author B has an average citation rate almost twice that of Author A, his citation record is *more* probable by a factor of 10^8 . This is a natural consequence of the power-law distribution which makes it far more improbable to have 10 papers with 100 citations each, than one paper with 1000 citations. The question of which of these options is ‘better’ requires a subjective answer, and it is unlikely that any single quantitative measure will satisfy everyone. Therefore, although the interpretation of non-statistical fluctuations in individual citation records is subjective, the likely presence of such fluctuations can be identified with ease and objectivity. The method developed in Chapter 5, will take the structure of author citation records into account in a much more sophisticated fashion.

Improbable Departments

Calculating r for entire departments is also possible. Physics Department Δ , which includes Author C, published a total of 1309 papers from 1980 to 2000, distributed with 81, 324, 474, 175, 216, 39 papers in the citation summary bins. This results in a $r = 285$. Physics Department Γ , which includes Authors A and B, published a total of 1309 papers during the same period with 81, 388, 378, 77, 28, 3. This yields the somewhat smaller value of $r = 65.9$. Such information can be of practical value since it seems likely that the ‘most improbable’ departments will have the greatest success in attracting the ‘most improbable’ author.

2.5 Summary

In this section, the homogeneity of the theory, experiment, and phenomenology subfields of the SPIRES database has been demonstrated. The data is well described by two power-laws; one for the minimally cited papers and one for the highly cited regime.

Striking features of the data set are the extremely large number of minimally cited papers, and that a small fraction of the papers account for most of the citations accumulated by the entire data set—4% of all papers account for 50% of the citations. The well known fact, that true progress in physics is driven by a few great minds, is documented in the SPIRES database to an extent that is almost unsettling. The picture that emerges is, thus, a small number of interesting and significant papers swimming in a sea of ‘dead’ papers. This has the practical consequence that any study seeking to understand the dynamics of interesting papers will be forced to discard most papers and accept the greatly increased statistical uncertainties. In the case of the SPIRES data set, this would amount to roughly 10,000 papers.

It has also been noted that merely considering the distribution of *papers* is not sufficient for an in depth understanding of SPIRES. To truly understand the network of scientific publications and the dynamics of excellence, we need to understand the SPIRES database in terms of individual authors or, more precisely, their citation records. This point of view makes the situation described above even more dramatic. Recall that Author C above accounts for 1.5% of all citations in the database, another example is that 7 authors (not necessarily the highest cited) account for 6% of all citations.

As a preliminary measure of these correlations, the measure of ‘unlikelihood’, r , has been introduced. Further, this measure offers a tool for comparing citation records in different fields with a known and controllable bias. We have seen that in spite of its virtues, r also has weaknesses. Some of these, such as the fact that comparisons between fields cannot but

involve unsupported assumptions and biases; cannot be eliminated. They should be made visible and discussed. Other problems—for example that ‘improbably’ bad authors, such as author D in the above, have a high r -value—can be remedied; this is surely a task for the rest of this thesis. Again, studying the correlations in the database, has a two fold purpose: To investigate the theoretical properties of a network of *authors* where a ‘node’ is an author citation record; and to *utilize* this knowledge to learn about scientific excellence.

Including knowledge about the author of every publication adds a new level of complexity to the network we have been considering in Chapter 2. In the present chapter, we shall study the structure that emerges when we include this new level in the investigation of the citation network. We will also discuss the basic statistics and properties of the author network. From now on, let us use the distinction that the *author network* refers to the entire network, including the author information, whereas the *paper network* refers to the network of papers discussed in the previous chapter. This augmentation of our knowledge, will set the stage for a more comprehensive analysis of the database, where we utilize this new knowledge to understand the longitudinal correlations that authors impose on the paper level.

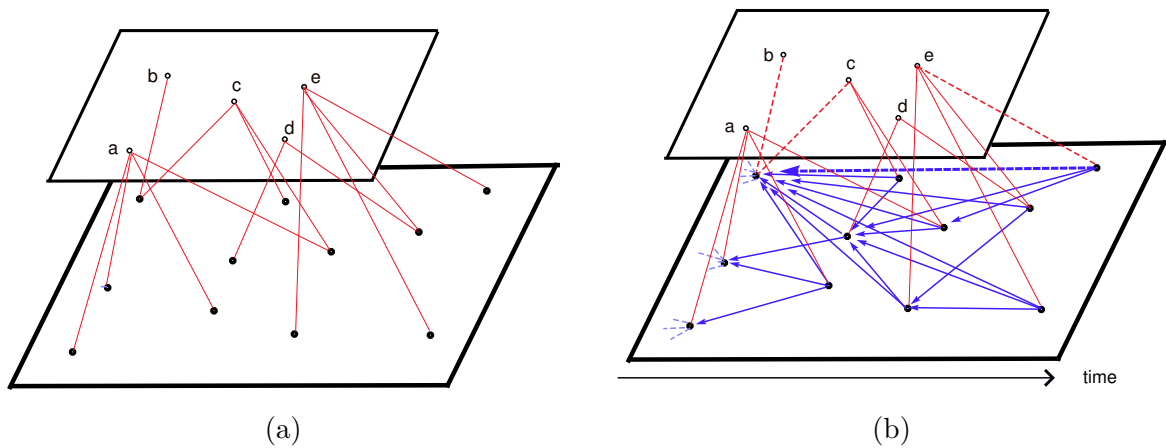


Figure 3.1: A visualization of the Author network. (a) Displays (a small portion of) the author level connecting to (an even smaller portion of) the paper level; each author is represented by a ‘o’ and a letter from a to e, each paper by a ‘•’. (b) Here, the paper-network from Figure 2.1 has been added to the picture. This representation provides an excellent illustration of how the author-level induces correlations on the level of papers; see main text for further details.

3.1 A New Level of Complexity

One way to visually represent the author network is to use two levels. To gain a deeper understanding of it, consider Figure 3.1. The (upper) level of authors in Figure 3.1 (a) connects to the (lower) paper level by means of their publications. Each of the 5 authors (a – e) have authored a number of publications represented by ‘•’s on the paper level. In Figure 3.1 (b) the directed network of citations and references between papers that was discussed in the previous chapter, has been added to the lower level, *cf.* Figure 2.1.

In analogy to the definition of citations and references from the previous chapter, we define an *author’s* references as the papers listed at the end of *all* of his papers and correspondingly, his citation count is the cumulated sum of citations in this entire citation record. This definition and the two level representation underscores that *references and citations between authors run via the paper network*. As an example of this, consider Figure 3.1 (b), where the dashed line illustrates how author e cites authors b and c via a reference from one of e’s papers to a (highly cited) paper co-authored by b and c. Sometimes, it is convenient to disregard the fact that citations between authors run via the network of papers, and collapse the two levels into one single level of authors citing other authors.

For example, collapsing the network in Figure 3.1 (b) so that only the links between authors are visible, results in the structure seen in Figure 3.2. This figure, however, represses

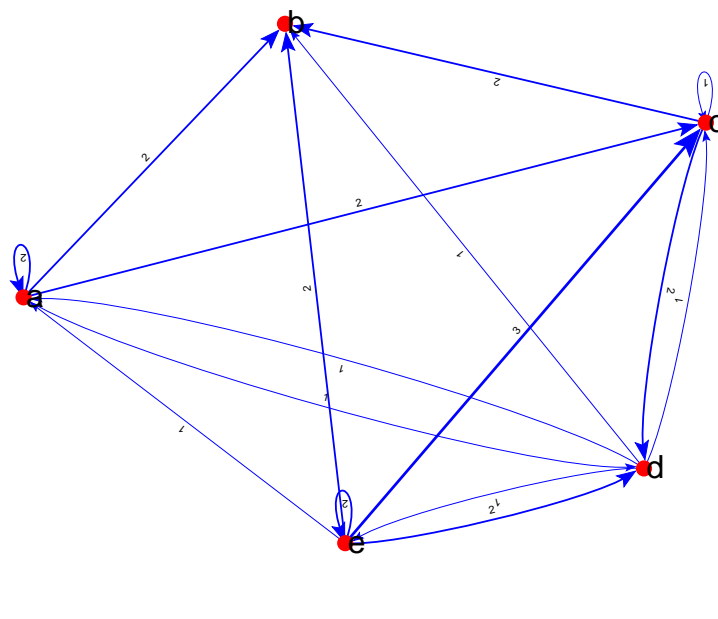


Figure 3.2: A visualization of the author network from Figure 3.1 (b) collapsed into one level of authors citing authors; each line is labelled by the number of citations it represents, also the thickness of a line is proportional to the number of citations it represents. Loops signify self citations. This network was generated using Pajek network visualization software [2].

so much structure that it becomes confusing: Here, a number is affiliated with each edge, the graph contains loops, *etc.*—this representation of the network also makes the time-line

from Figure 3.1 (b) an impossibility. Generally it holds true that the two-level representation strengthens our intuitions about the structure of the network; it will prove indispensable for modelling the network structure in Chapters 6 and 7.

3.1.1 Disconnecting the World Wide Web

Previously, we have discussed the similarities between the paper network and the www. The inclusion of the author level in the considerations, however, reveals that the two networks are in fact radically different; the internet does not possess any structural property analogous to the strong correlations that the author level imposes on the network of papers. This fact has been almost completely ignored in the literature, where the citation network is usually considered a much simpler network than the www, because of the impossibility of new connections arising between old vertices, that is, papers in the database suddenly adding new papers to their list of references [13, 14].

3.1.2 Data Structure and Notation

It is clear from the above that access to all reference lists in SPIRES would allow a complete reconstruction of the author network, and would allow us to calculate any imaginable property of the web of science. Unfortunately, the data that is publicly available directly from SPIRES, is not complete. For each paper, only *author information*, *publication year*, and *total number of citations* is available¹, and therefore a complete recreation of the network is impossible at this point. In this section, let us discuss what data *is* available, and how it is organized. In doing this, it is convenient to introduce a bit of notation; this notation will prove its worth when we begin analyzing more complicated aspects of the data structure.

After the parsing², the papers written by each author is collected in a list. In this publication record, each paper is labelled by citation count and year of publication. Accordingly, let us denote the set of authors

$$\mathcal{K} = \{\mathbf{K}_1, \mathbf{K}_2, \dots, \mathbf{K}_i, \dots, \mathbf{K}_{\mathcal{N}}\}, \quad (3.1)$$

where \mathcal{N} is the total number of authors. Each author \mathbf{K}_i is a vector

$$\mathbf{K}_i = \begin{pmatrix} i(1) \\ i(2) \\ \vdots \\ i(n_i) \end{pmatrix}. \quad (3.2)$$

The elements of \mathbf{K}_i , the $i(j)$'s, are records of the i th author's individual papers. The $i(j)$'s are ordered in time, such that $i(1)$ is author \mathbf{K}_i 's first publication, *etc.* The reason that these vectors only contain the relative time information, is that we are interested primarily in exposing the properties of *authors*; this means that the relative time ordering is more interesting for our present purposes. Later, we will include the year of publication in the analysis of the paper distribution, labelling these records with a time-stamp $i(j, t)$ —because

¹Naturally, details of where a given paper is published and its title are also included, but that is not of importance to this investigation.

²The raw data outputted from SPIRES was parsed using a PERL script included in Appendix A.4. For details on the original data structure and sources of error, *cf.* also Appendices A.1 and A.2.

it is natural to expect that, since older papers have had more time to accumulate citations than newer papers, older papers are generally more cited. This topic is treated explicitly in Chapter 7, where it is also concluded that this effect is negligible³.

To access the citation counts of the individual papers, $i(j)$, I use the terminology $k_{i(j)}$ (note that these $k_{i(j)}$'s are equivalent to the k 's of the previous chapter, the only difference being the subscript; the subscript indicates that we know which author has published the paper in question and the time of publication). These primary definitions allow us to express a number of quantities in a precise form. The total number of citations of the i th author is the sum of the citation counts, $\sum_j k_{i(j)} \equiv m_i$. Clearly, the dimension, $\dim(\mathbf{K}_i) \equiv n_i$, is the number of papers published by the i th author. The reader should note that the fact that the \mathbf{K}_i 's have different dimensions (this simply means that each author has not necessarily published the same number of papers), means that \mathcal{K} is *not* a matrix, as one would naturally suspect from equation (3.1), and from the notation in general.

3.2 The Data

With this notation to clarify our thinking, we are equipped to address the actual SPIRES data. First, however, we need to examine a few complications regarding the data. The first problem can be easily explained using a familiar example; the co-author network.

3.2.1 Problems

Paper Weights

The author network is closely related to the co-authorship network studied by Newman in [48, 49, 50]. The co-author network is the network, where a link between two authors arises, when they have co-authored a paper. Therefore each '•', where two or more lines meet in Figure 3.1 (a), constitutes a link in this network. If we, for example, begin with author b, author c is one step away from b in the co-author network, and via c, author a has b-number 2, *etc.* We know from Section 1.3.1 that the co-author network is a small-world network of considerable size and complexity, with power-law distributions of the number of authors per paper. Ordering papers according to authors, as in Equation (3.1), creates an unavoidable problem with the statistical weight of each paper.

In the paper network, each link between papers is unique: One reference results in one citation. This is not the case for the author-by-author distribution. The example of a link between authors in Figure 3.1 (b) shows that each paper is necessarily weighed by the number of authors. One reference from author e results in *two* citations—one for both authors b and c. More generally, *one* paper with α authors citing *one* paper with β authors results in $\alpha\beta$ links between authors, and the existence of the co-author network demonstrates the scope of this problem: If an experiment paper with 1500 authors cites another paper with 1000 authors, we end up with 1,500,000 links between authors stemming from one citation of a paper⁴.

³As mentioned earlier, the information on *when* a given paper $i(j)$ received each of its citations is not available with the present data—this is unfortunate, since the citation history for a paper is quite an interesting subject [64]. We will, however, spend a little time on this in Chapter 7, more specifically in Section 7.3.

⁴This example is exaggerated for more than one reason: For the experimental cooperations, the large groups of authors are never mentioned explicitly in the SPIRES entry, the names are usually attached in a separate

To minimize this effect, in the remainder of the paper, when the author level is included in the description, *we shall solely consider the theory subset*. As was pointed out earlier, the theory subset typically has fewer authors per paper than the other SPIRES subfields that have a substantially higher number of authors, *cf.* Newman's results from Section 1.3.1. Later in this chapter we will investigate the subject in more detail.

The 'David Gross'-Effect

The next data-related problem, is what we shall call the 'David Gross'-effect. The name is due to problems that arose when I tried to find a list of the papers written by this particular author (who is the director of the Kavli Institute for Theoretical Physics in Santa Barbara). Intuitively, one would expect the name David Gross⁵ to be relatively unique.

The first realization in this respect is that searching for 'D. Gross' in the database, is too broad of a search; this is illustrated in Table 3.1. A search for 'D. Gross' finds a total of 271 papers. Since we are searching manually, we have an option that can help us refine our search

| Number | Occurrences | Name |
|--------|-------------|-------------------|
| 1 | 34 | Gross, D |
| 2 | 1 | Gross,D |
| 3 | 14 | Gross, D A |
| 4 | 55 | Gross, D H E |
| 5 | 65 | Gross, D J |
| 6 | 1 | Gross,D J |
| 7 | 2 | Gross, D L |
| 8 | 1 | Gross, Dan A |
| 9 | 1 | Gross, David |
| 10 | 97 | Gross, David J |
| | | Total: 271 papers |

Table 3.1: Names that pop up when searching SPIRES for D. Gross.

(this is not available when doing automated searches): We can do a so-called *name search* of SPIRES, searching for D. Gross in the *names* section of SPIRES. This yields the response:

- (a) Gross, D. H. E. (Hahn-Meitner Inst. & Freie U., Berlin),
- (b) Gross, Dan (General Electric),
- (c) Gross, David J. (ITP, Santa Barbara),
- (d) Gross, Klaus-Dieter (GSI, Darmstadt).

file that is not included directly in the database and, therefore, not parsed by my PERL script; hence these large cooperations are listed under the name of the group, *e.g.* 'Higgs Particle Search Group'. Generally, in papers with more than 3 authors, only the first author is listed, and the rest are subsumed under the Latin abbreviation '*et al.*', meaning 'and others'. When cleaning up the SPIRES data it is necessary to take this into consideration, otherwise '*et al.*' would be categorized as one of the most productive authors of all time.

⁵Obviously, there is nothing special about David Gross in this respect; any other inconspicuous name could have been used to the same effect.

Clearly, *our* D. Gross, is (c) in this list above; his middle initial allows us to eliminate no. 3, 4, 7, and 8 from the list in Table 3.1, but it is still possible to attribute the remaining 199 papers to *our* David Gross. If we, however, want to be completely certain that we are talking about *the* David J. Gross, we can only include no. 10 from Table 3.1, since the *names* section of SPIRES is not complete—there may be other D. J. Gross’ out there. Choosing this strategy assigns only 97 papers to the director of the KITP. The most realistic bid, when including all of the available information, is including no. 5, 6, and 10 in the search for papers, leaving us with a list of 163 articles.

In summary, the maximum number of papers we came across for this search is 271 papers, whereas the minimum is 97 papers. The difference is 174 papers! Including all of the information available, we find that the true answer lies somewhere approximately in between these two numbers at 163 papers. The point of this exercise is to illustrate the simple point that a person’s last name and first initial is not enough to identify a person uniquely. On the other hand, including *all* initials, or even spelling out someone’s first name may be too specific and exclude from the search a large number of papers. This concrete example should also demonstrate to the reader how profound this effect is, and why taking it into consideration is important.

Dealing with the ‘David Gross’-Effect

To remedy this problem, we will simply parse the theory subfield data twice. Once where *last name and first initial* are used as the criterion for identifying a person as a single author, and once where *last name and all initials* are used to distinguish authors. The first method will—as we have seen in the example above—underestimate the number of authors, and the latter method will most likely overestimate the number of distinct authors; we can think of this as an upper and lower bound on the number of authors in the database. When stating results, the convention ‘*First Initial*’ result (*‘All Initials’ result*)’ will be used. Regarding plots, the convention is to plot the results for the lower bound; this is not a problem, since for all practical purposes, the differences between the upper and lower bound are indistinguishable on log log scales. Furthermore, it is assumed that the mechanisms behind a given author including his middle initial or not, *etc*, are random; therefore, it is assumed that ‘All Initial’ parsing removes *random* papers from authors’ citation records, whereas the ‘First Initial’ parsing unites citation records of random authors.

3.2.2 Quantitative Comparisons

The data set for the theory subfield consists of a total 44,397(52,139) authors. After removing papers for which no citation data is available and papers with publication dates that fall outside the time interval 1945–2001, we are left with $\mathcal{N} = 34,434(39,921)$ authors. A partial reason for the dramatic drop in the number of authors, when papers with no available citation information are removed, is the large fraction of authors—some 38(41)%, corresponding to 16,890(21,573) people—present in the database, with only one publication before the cleaning up. The approximately 21% of the total papers with no available citation information are distributed almost evenly among authors, and this means that for a large number of authors no citation information is available for their single paper. Accordingly, not only the N/A-papers, but also their authors are removed from the database. A similar, although not as pronounced, mechanism is a play for other ‘minimally cited’ authors.

| | First Initial | All Initials |
|---|---------------|--------------|
| Total # of Authors | 34,434 | 39,921 |
| Total # of Papers | 281,816 | |
| Total # of Citations | 4,571,192 | |
| Ave. # of papers per author, $\langle n_i \rangle$ | 8.2 | 7.2 |
| Ave. # of citations per author, $\langle m_i \rangle$ | 133 | 115 |
| Ave. # of citations per paper, $\langle k_{i(j)} \rangle$ | 16.2 | |

Table 3.2: A summary of the basic statistics for the author network.

The remaining authors have published a total of $\sum_i n_i = 281,816$ papers (the number of papers and citations is naturally the same, independently of how they are distributed among authors, so there is only one result here). We know from Chapter 2 that the number of distinct papers in the theory subset is 159,946; this number is approximately doubled because of the co-author effect; in other words, the average number of authors per paper is a little under two for the theory subfield. As promised, this number is small compared with the average number taken for SPIRES as a whole. This amounts to an average of 9.0 authors per paper—the largest collaboration in SPIRES is 1,681(!) authors [49].

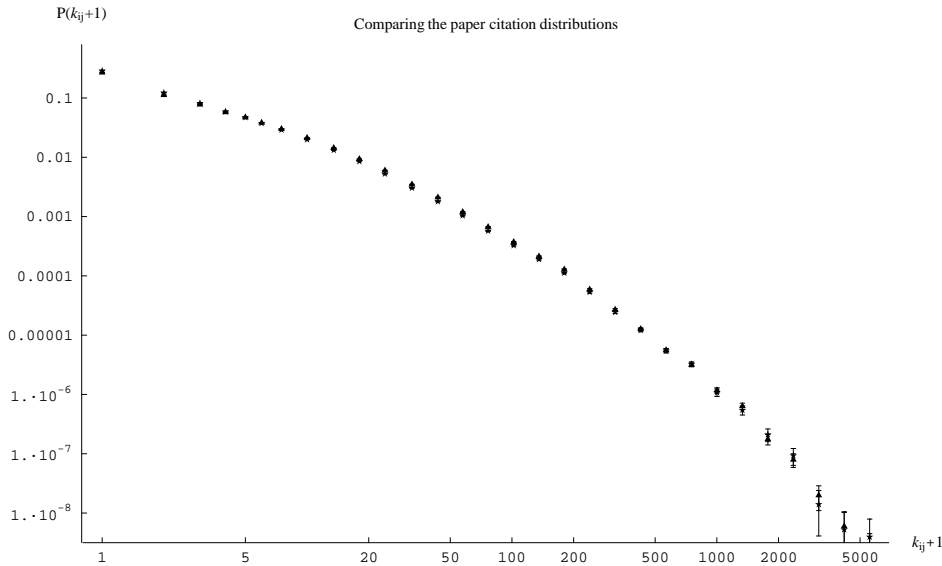


Figure 3.3: A comparison of the papers in the theory subfield, collected paper-by-paper (★) as was discussed in great detail in Chapter 2 and author-by-author (▲). The normalized distributions are virtually indistinguishable.

SPIRES theory has an average of $\langle n_i \rangle \approx 8.2(7.2)$ publications per author and the corresponding mean is $\tilde{n}_i = 2(2)$ papers. The data set has received $\sum_i m_i = 4,571,192$ citations from authors in the entire SPIRES database, resulting in an average of $\langle m_i \rangle \approx 133(115)$ citations per author, with a considerably lower mean of $\tilde{m}_i = 8(7)$ total papers. The average number of citations per paper $\langle k_{i(j)} \rangle \approx 16.2$. This value of $\langle k_{i(j)} \rangle$ corresponds well with the corresponding value for the data set counted paper-by-paper. This set of data yielded a value

of $\langle k \rangle_{old} = 14.77$. The basic statistics are summarized in Tables 3.2 and 3.4. As in the previous chapter, the large factor differences between the means and the medians are indicative of fat-tailed distributions.

Even though this average number of citations per paper stays relatively constant, regardless of the weighing of papers by the number of co-authors, the question remains, however, whether or not the counting of citations author by author has changed the shape of the *distribution* of paper citations. This question is answered in Figure 3.3, where the normalized distribution of citations, collected paper-by-paper from Chapter 2 (★) is plotted alongside the corresponding distribution collected author-by-author (▲), both are normalized to one. It turns out that the normalized distributions are virtually indistinguishable—counting papers author-by-author does not alter the shape of the distribution. This is remarkable and once again demonstrates the remarkable homogeneity of the SPIRES database; the collaborations in SPIRES span the entire range of citations so evenly that including them is virtually equivalent to including a multiplicative factor. There is no ‘typical’ co-authored paper.

3.3 The Frequency Distributions

Now that we have established the fact that the distribution of citation remains remarkably constant, in spite of the weighing of papers by number of authors, it is time to reap the benefits of counting the papers author-by-author and discover some of the properties of the *authors* in SPIRES, taken as a whole. As we will learn in this section, the two (independent) power-law structure is a common property of all of the distributions we will be considering; this supports the idea that different dynamics rule the highly- and minimally cited regimes.

3.3.1 Total Citations per Author

Let us begin by discussing the distribution total number of citations per author. This is the distribution that Laherrere and Sornette investigated for the 1120 most cited physicists in the interval 1981-1997 [53]. Laherrere and Sornette found this distribution to be described by a stretched exponential, $N(m_i) \sim \exp[(m_i/m_0)^\beta]$, with $\beta \approx 0.3$. It is evident from Figure 3.4 (a)

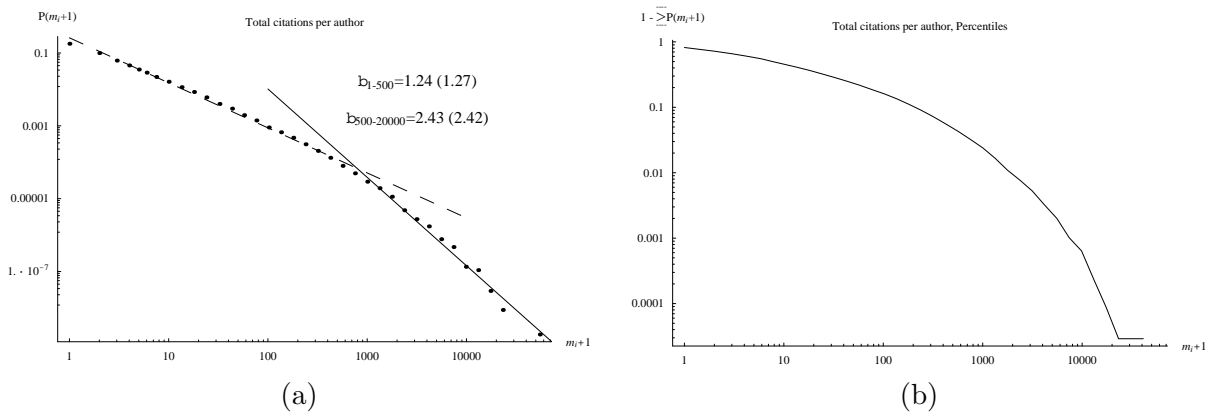


Figure 3.4: (a) The normalized, binned distribution of total citations per author, $m_i + 1$ vs. $P(m_i + 1)$. The two straight lines have slopes of $\beta_{1-500} = 1.24(1.27)$ and $\beta_{500-20000} = 2.43(2.42)$, where $P(m_i + 1) \sim (m_i + 1)^{-\beta}$. (b) Displays the distribution of percentiles, $m_i + 1$ vs. $1 - \sum P(m_i + 1)$. The most cited author is cited 61,062 times in the database.

that the $m_i + 1$ vs. $P(m_i + 1)$ distribution is well described by two power-laws, $P(m_i + 1) \sim (m_i + 1)^{-\beta}$, $\beta_{1-500} = 1.24(1.27)$ and $\beta_{500-20000} = 2.43(2.42)$. At the phenomenological level (which is the level we are primarily working on in this chapter), we are, however, not in a position to challenge the stretched exponential description suggested by Laherrere and Sornette; the data *is* well approximated by a stretched exponential⁶. In Chapters 6 and 7, Sections 6.3.2 and 7.1, the subject of the functional form of the probability distributions is discussed on the basis of a theoretical model for SPIRES. For now, let us settle with establishing the fact that, although there is nothing sacrosanct about the two power-law shape, the data is reproduced rather faithfully by the two independent power-laws. Furthermore, this structure is even clearer in the following two plots, and it is—as was the case for the citation distribution in Chapter 2—natural to expect that the different power-laws are caused by different dynamics in the two regimes.

Figure 3.4 (b) is included for more utility-minded reasons. The ‘total number of citations’-distribution allows an author to compare himself directly to the distribution of all authors in SPIRES; in Figure 3.4 (b), it is easy to read off which percentile one belongs to. For example, we find that only 8(7) career citations are needed to be in the top 50% of theoretical high energy physics authors of all time—by definition, this number is also the median of the total citation distribution. To make the top 10 percent, a career total of 238(195) citations is needed; making the 99th percentile requires your career total to exceed 2170(1876) citations, *etc.*

3.3.2 Papers per Author

The number of papers per author was considered by Newman in connection with his investigation of the co-author network; these results were summarized in the Introduction. In his first paper on this subject, Newman hypothesized that this distribution was described by a power-law with an exponential cut-off that he attributed to the limited time window under consideration (1995 – 1999) [48]. However, in the next paper on the same subject, in which the time window was expanded to include the years 1974 – 1999, the following conclusion was reached: With regard to the SPIRES distribution of papers per author, “neither pure nor truncated power law fits the data well” [49]. Keeping in mind that the data Newman considered in his second paper is identical to the data considered here (except that I have access to the data from before 1974 and up to 2001), it seems clear that Newman is hesitant, exactly because the data is clearly described by a two power-law structure. This structure is *very* clear from Figure 3.5; this is why the data cannot be fitted with neither the pure nor the truncated power-law.

The slopes of the power-laws $P(n_i) \sim n_i^{-\gamma}$ for the papers per author distribution are $\gamma_{1-50} = 1.47(1.55)$, and $\gamma_{50-300} = 3.86(4.23)$. The reason for the relatively large discrepancy between the slope of the highly publishing authors, when counting ‘first initial’ and ‘all initials’, is that the ‘first initial’-counting creates unnaturally long citation records which result in less steep slopes, whereas the ‘all initials’-counting divides publication records of single authors into separate pieces, which in turn result in steeper slopes. The fundamental two power-law structure remains unaffected, however. Again, it is remarkable that the two power-law structure seems to remain a common feature of every single aspect of the SPIRES

⁶Another option that would fit the data well, would be to make a smooth interpolation between the two power-laws. This, however, would bring nothing new to the analysis; it would only make the analysis a little less transparent.

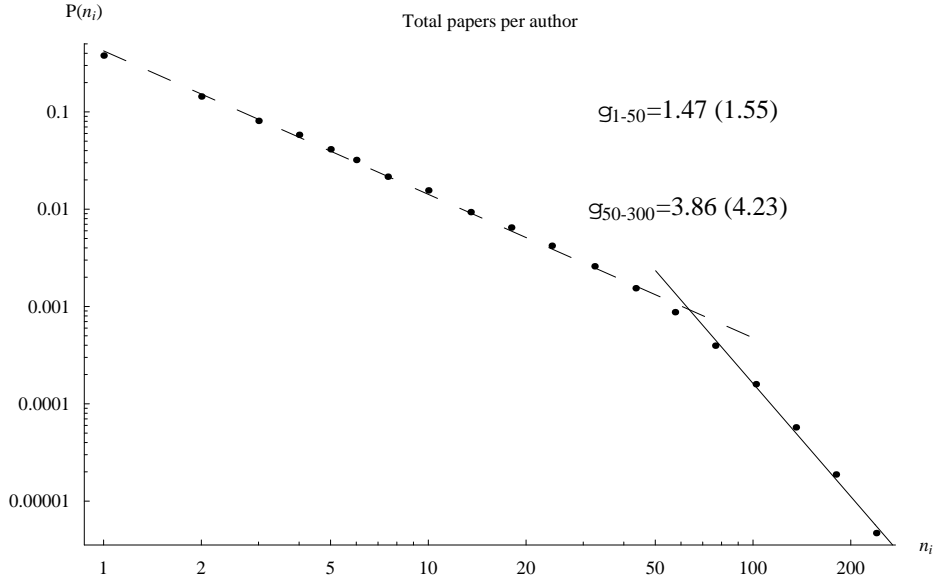


Figure 3.5: The normalized frequency distribution of total papers per author, n_i vs. $P(n_i)$. The two power-laws plotted along with the data points have slopes of $\gamma_{1-50} = 1.47(1.55)$, and $\gamma_{50-300} = 3.86(4.23)$, where $P(n_i) \sim n_i^{-\gamma}$.

database. The latter power-law must necessarily be steep, since it is only possible for a scientist to produce a finite number of papers in the span of a career. The average number of papers published per person per year is slightly increasing for SPIRES year by year, and therefore the $50 < m_i < 300$ distribution may very well consist mainly of authors that have already retired.

3.3.3 Average Number of Citations per Paper per Author

The final plot we will be considering in this section, is the distribution of the average number of citations per paper per author, m_i/n_i vs. $P(m_i/n_i)$. This plot is interesting primarily for its utility. If you are a young author and you want to compare yourself to the rest of the distribution, the total number of citations is not a good measure, simply because young authors have not written a lot of papers. Therefore, the distribution of author paper averages can be useful. This distribution allows any author at any stage in his or her career to compare himself to the entire population of authors, simply because this distribution connects the citation-count of that same author directly to the number of papers published by the i th author. Connecting m_i and n_i for individual authors, causes the distribution, in Figure 3.6, to actually contribute information that is not contained in the two previous plots.

This distribution is also described by a double power-law structure, $P(\frac{m_i+1}{n_i}) \sim (\frac{m_i+1}{n_i})^\delta$, with $\delta_{1-50} = 1.44(1.45)$ and $\delta_{50-200} = 3.32(3.15)$ in the low and high average regime, respectively. The reason the slope of the high average regime is steeper for the ‘first initial’ data is that the ‘David Gross’-effect causes us to cut random papers out of highly cited author’s citation records. Most of the time, this results in adding another author to the group of authors with low averages, because cutting out random papers is equivalent to drawing a paper from the paper citation distribution at random: We know that, because of its power-law structure, it is much more likely to draw a minimally cited paper. This is barely visible on the

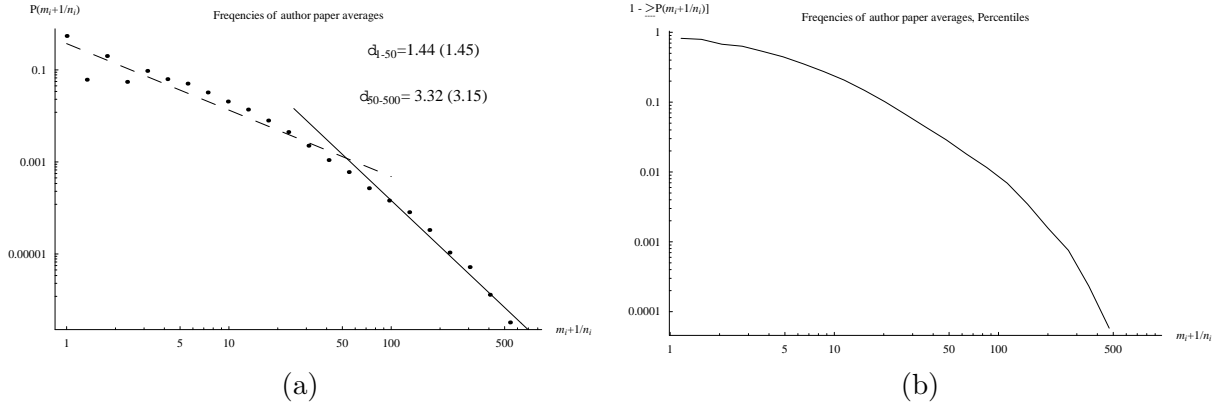


Figure 3.6: (a) The normalized frequency distribution of the average number of citations per paper per author, $(m_i + 1)/n_i$ vs. $P([m_i + 1]/n_i)$. (b) The corresponding distribution of percentiles, $(m_i + 1)/n_i$ vs. $1 - \sum P([m_i + 1]/n_i)$. The slopes of the two power-laws are $\delta_{1-50} = 1.44(1.45)$ and $\delta_{50-200} = 3.32(3.15)$

log log plot. Once in a while, however, the same effect can result in extremely high average author paper citations, because a highly cited paper is drawn at random. For example, the highest average for the ‘all initials’ data is 982.5 due to the fact that Erick *J.* Weinberg left out his middle initial in two of his 81 publications. One of these papers is his most cited paper with 1963 citations. The other paper is uncited. The average of these two papers is the before-mentioned 982.5 citations per paper; E. J. Weinberg’s true average is 69 citations per paper. Because of this effect, Figure 3.6 should be used exhibiting great caution. Many of the author averages of over 100 citations per paper are due to citation records that are artificially shortened, like in the example above. Edward Witten, who is the most cited author in the database has an average of 241 citations per paper.

The dips for non-integer values in the low-average part of the distribution in Figure 3.6 (a) is due to the large number of authors with only one publication—these comprise 38(41)% of the papers—and the discrete nature of both the paper- and citation-count. Clearly, these authors can only have an integer valued average number of citations per paper: Since most of the authors with only one publication are minimally cited, this boosts the integer valued averages of 0 and 1 citation and suppresses the corresponding 1/2-integer valued averages. For the 14(15)% of authors with two publications, we expect the probability of receiving an average of 1/2 or 3/2 citations per paper to be a little less likely than receiving 0 or 1 citation on average, due to the power-law distribution of total author citations. As the number of published papers increases, the discretization effects are suppressed and the average number of papers per author becomes a ‘continuous’ variable.

3.4 The Scientific Staff

From Figure 3.5, we know that the number of papers per author decays as a power-law $P(n_i) \sim n_i^{-\gamma}$ with $\gamma_{1-50} = 1.47(1.55)$ for 1 – 50 publications and $\gamma_{50-300} = 3.86(4.23)$ for 50 – 300 publications. This means that the vast majority of authors in SPIRES have only published a small number of papers before leaving high energy physics. It is only natural that many authors leave SPIRES; some move on to other branches of physics, or leave academic

| Minimum # of author publications | # of authors | | % of total population | | % of total citations | |
|-------------------------------------|---------------|--------------|-----------------------|--------------|----------------------|--------------|
| | First Initial | All Initials | First Initial | All Initials | First Initial | All Initials |
| 1 | 34,434 | 39,921 | 100 | 100 | 100 | 100 |
| 5 | 11,633 | 12,010 | 34 | 30 | 94 | 92 |
| 10 | 7,015 | 6,637 | 20 | 17 | 88 | 86 |
| 15 | 4,990 | 4,832 | 14 | 11 | 83 | 80 |
| 20 | 3,799 | 3,643 | 11 | 9 | 77 | 74 |
| 25 | 2,983 | 2,809 | 9 | 7 | 72 | 69 |
| 50 | 1,063 | 966 | 3 | 2 | 49 | 46 |

Table 3.3: The number of authors remaining in SPIRES as the author’s minimum number of papers per author increases. The percentage of total citations generated by this population is also included; a minimal fraction of all authors generate a majority of the citations.

physics altogether to pursue other careers. This is due to the simple fact that every professor has many Ph.D. students, but (on average) only one of these students get to fill his position. We are interested in scientific excellence, so it is interesting to find out what happens to the distributions of citations, when authors with just a few publications are removed. In Table 3.3 we explicitly see the sizes of the remaining author populations, as authors with less than 5, 10, 15, 20, 25, and 50 papers, are removed from the network. These numbers should remind us that we are dealing with a power-law distribution that has a *very* steep slope from 50 through 300 papers.

3.4.1 The Drop in Total Citations

First, let us consider the effect on the distribution of the total number of citations per author. The resulting distribution is displayed in Figure 3.7 and explicit percentages are displayed in Table 3.3. The entire distribution (the red dots) is normalized to one; the remaining distributions are normalized by the same factor, so that it is easy to see exactly from which part of the distribution, authors are removed—these, however, can no longer be regarded as probability distributions.

In removing the minimally publishing authors, we ‘hollow out’ the minimally cited end of the author spectrum, in such a way that the remaining distributions develop peaks that are different from $m_i = 0$. The exact location of these peaks depend upon how many papers are removed, and are clearly visible from the graph. On the more qualitative side, the mean climbs from 133(115) for $n_i \geq 1$ to 1108(1123) for $n_i \geq 25$, and the median rises from 8(7) to 518(528) for these populations. To understand why the ‘First Initial’-results are higher than the ‘All Initial’-results when including the entire population and the other way around after ‘pruning’ the distribution so that it contains only the Scientific Staff, we once again have to turn to the ‘David Gross’-effect for an explanation. The ‘First Initial’ counting finds fewer authors, and since the number of citations is constant, this results in a higher average value. How does this change when authors are removed? Well, since the ‘All Initials’ counting finds a higher number of authors, and because these authors have shorter publication records, a greater number of these authors are removed; this mechanism is clear from Table 3.3. Many of these discarded authors are based on random clippings from the remaining authors’ (long) publication records—but since these clippings correspond to a random draw from the distribution of blue dots in Figure 3.8 (b), a majority these papers are minimally cited. Accordingly, the remaining authors have a higher average number of citations than the ‘First Initial’-group that has no ‘clippings’.

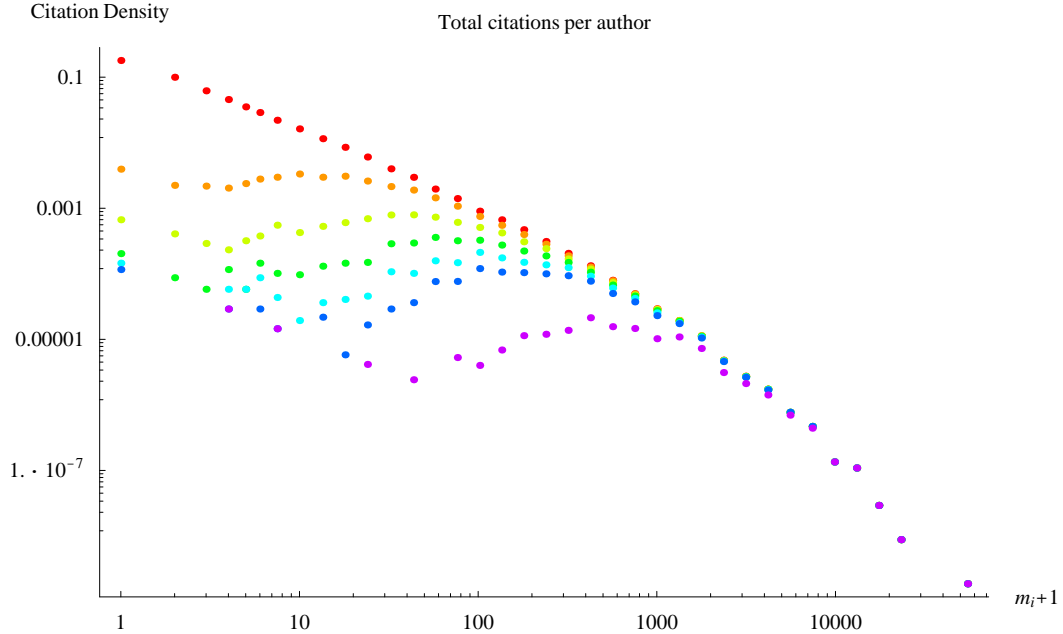


Figure 3.7: The change of the distribution of total author citations (red dots), as authors with less than 5 (orange), 10 (yellow), 15 (green), 20 (turquoise) 25 (blue), and 50 (purple) citations are removed from the database. The total distribution has global normalization 1.

Getting back to the distribution of total citations, there is nothing unexpected about the minimally cited part of the distribution, being hollowed out; authors with only a few publications are less likely to receive many total citations. Naively, we would expect that not being cited is an attributing reason for authors to leave academic physics, but even if the average number of citations for the ‘minimally publishing’ authors were the same as for the rest of the population, an author with 4 publications would have a hard time competing with an author with 40 publications. More qualitative statistics regarding this distribution are summarized in Table 3.4, in the Summary.

The only vaguely surprising fact about Figure 3.7 is that it reveals that SPIRES actually contains authors with 25 or more publications and *a total of 0 citations*. One author, \mathcal{U} , with 52 publications has a career total of only 4 citations. In the context of the Power of Excellence defined in Section 2.4, drawing these people at random on the citation distribution is immensely improbable. For author \mathcal{U} , we find $r = 20.6$, with r defined as in Equation (2.2). This again emphasizes the need to refine the concept of the Power of Excellence, if we do not want the name of this measure to contrast the content.

3.4.2 The Paper Citation Distribution

The changes in the distribution of citations of publications when ‘minimally publishing’ authors are removed, offer more of a surprise. In Figure 3.8, the changes in the distribution of citations per paper (red) are plotted as authors with 5 (orange), 10 (yellow), 15 (green), 20 (turquoise), 25 (blue), and 50 (purple) publications are removed from the database. Figure 3.8 (a) shows the change of the distribution of paper citation, as larger and larger parts of the author distribution are removed; here, only the total distribution is normalized

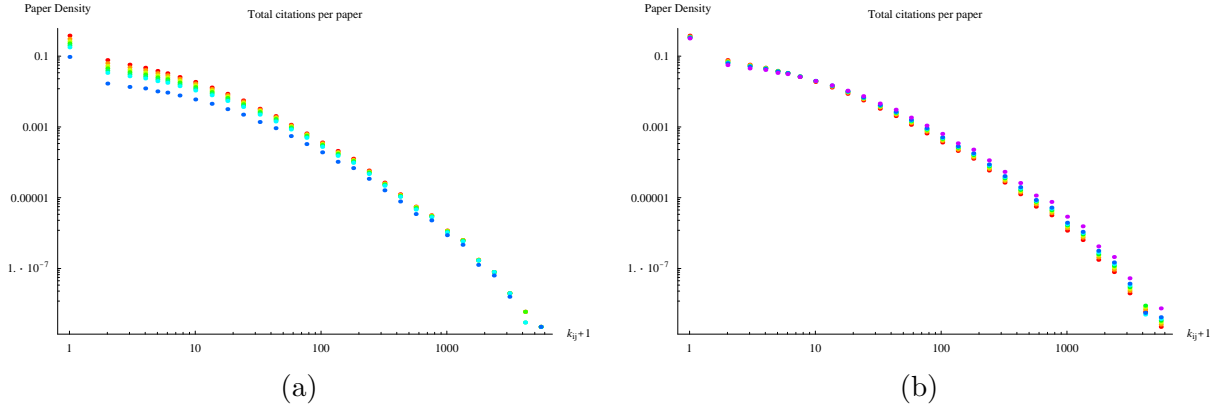


Figure 3.8: The change of the distribution of paper citations (red dots), as authors with less than 5 (orange), 10 (yellow), 15 (green), 20 (turquoise), 25 (blue), and 50 (purple) citations are removed from the database. In (a) the distribution including all the authors is normalized to one, the other distributions are normalized by the same factor. In (b) the global normalization of each distribution is 1.

to one—the same strategy as in Figure 3.7. This figure is interesting because the distribution of paper citations is remarkably constant when compared to the significant drop—the ‘hollowing out’—of the low-cited part of the distribution of total citations, as the minimally publishing authors are removed. There is a little more of a ‘drop-off’ for the minimally cited papers, than for the highly cited papers, but it is clear to see that papers are removed from the entire range of the distribution of paper citations.

The distributions displayed in Figure 3.8 (b) are the same, but in this figure, each sub-distribution is normalized to one. This version of the plot demonstrates that the probability distribution ‘tips over’, when the ‘minimally publishing’ authors are removed: While the probability remains constant at around 7.5 citations per paper, the probability of drawing minimally cited papers falls off, and the probability of finding papers with many citations grows proportionally higher. The change for this distribution is far less drastic than the change for the Total citations distribution; this is reflected in the changes in the mean and median, that grow from $\langle k_{i(j)}(n_i \geq 1) \rangle = 16.2$ to $\langle k_{i(j)}(n_i \geq 25) \rangle = 21.9$, and $\tilde{k}_{i(j)}(n_i \geq 1) = 3$ to $\tilde{k}_{i(j)}(n_i \geq 25) = 5$. Further quantitative results can be found in Table 3.4. The main conclusion we can draw from the above, is that authors who are an integral part of academic physics, *i.e.* authors with long publication records, still publish a majority of un- and minimally-cited papers. By extension, we can also conclude that the authors with only a few publications publish highly cited papers. It is interesting to investigate this topic in further detail.

In Figure 3.9, the distributions of minimally publishing authors that are *removed* from Figure 3.8, are displayed. The power-law behavior is present once again, and we notice quite a few highly cited papers are amongst these author’s publications. How is this possible? It seems counter-intuitive that it is possible for a person, who only publishes *once* in SPIRES, to write a ‘best-selling’ paper. The most convincing answer to this question points to the fact that most authors who publish in SPIRES do not write their first papers alone; they collaborate with their advisors or other experienced scientists. In a collaboration between an experienced scientist and a young Ph.D. student, it is likely that the ‘properties’ of the experienced scientist is reflected in their joint paper, since the more experienced scientist is

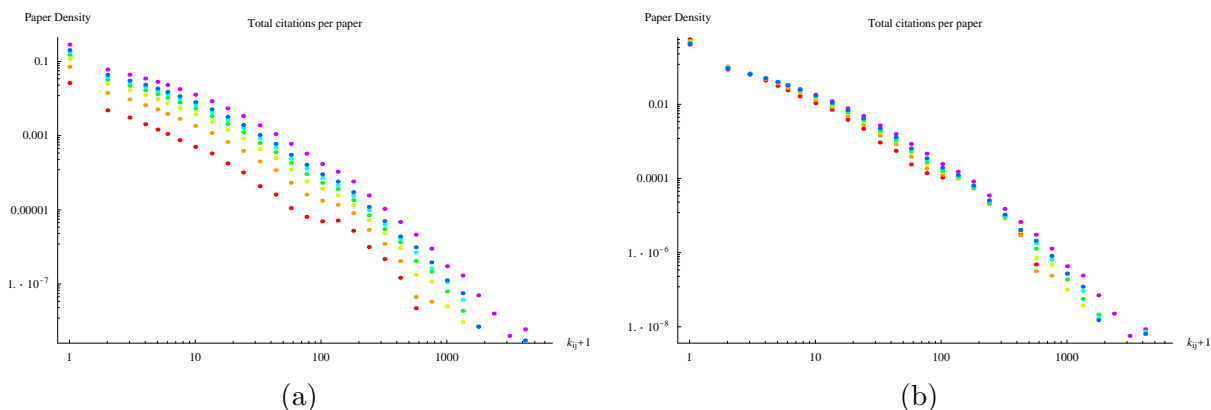


Figure 3.9: The distribution of the authors that are *removed* in Figure 3.8. In this figure, the color coding scheme is a little different. The distribution of citations in papers from authors with only one publication is plotted in red. The less than 5 publications distribution is in orange, the less than 10 is red, *etc.* In (a) the normalization is set to constant so that the global normalization of the entire population is 1. As usual, (b) displays the same distributions, but this time the normalization of each distribution is set to 1.

not interested in jeopardizing his reputation by letting a young (potentially insane) Ph.D. student publish a paper in his name, without thorough supervision, and therefore has a profound influence on final outcome of the paper.

No matter what influences the distribution of ‘minimally citing’ authors, it is interesting that the quality of papers published by this group of authors is relatively high. Let us make the distinction that people with more than 25 publications are considered the ‘Scientific Staff’. The number 25 seems reasonable, since it is possible to generate a citation record of approximately this size as a Ph.D. and *post. doc.* without ever finding a job in high energy physics. On the other hand, people with more than this number of publications most probably have acquired some sort of some permanent job in the world of high energy physics. Thus, an extremely interesting question for the rest of this thesis is: ‘What qualities⁷ sets the Scientific Staff apart from the rest of the database?’.

A first stab at answering this question has already been taken. Figure 3.7 is not very interesting in this respect; it is clear that all other things being even, authors with many papers will receive more citations than authors with few citations. However, Figure 3.8, is worth noticing because in this figure, an explicit connection is created between the number of papers per author and the distribution of paper citations; we explicitly get to see the distribution of citations of papers written by the Scientific Staff compared to the total distribution. In the following—Figure 3.9—we also get an explicit look at the paper citation distribution of the authors with only a few publications. The lesson we can learn from these two figures is that although the Scientific Staff is clearly more cited than the ‘minimally publishing’ authors, the latter population still does remarkably well, and it includes the author of several highly cited papers. A more quantitative comparison of the two populations can be found in the summary, *cf.* Table 3.4.

⁷Obviously, one quality that sets these authors apart from the rest of the population of authors is that they publish more papers, but the question remains: Why do they publish more papers? Are they more persistent, are they luckier, or do they simply *write better papers*?

3.4.3 Average Number of Citations per Paper per Author

The properties of authors as a function of their citation record, are made even more explicit in Figure 3.10, where the changes in the distribution of author's average number of citations (m_i/n_i) is displayed as 'minimally publishing' authors are removed. Here the tendency is clear.

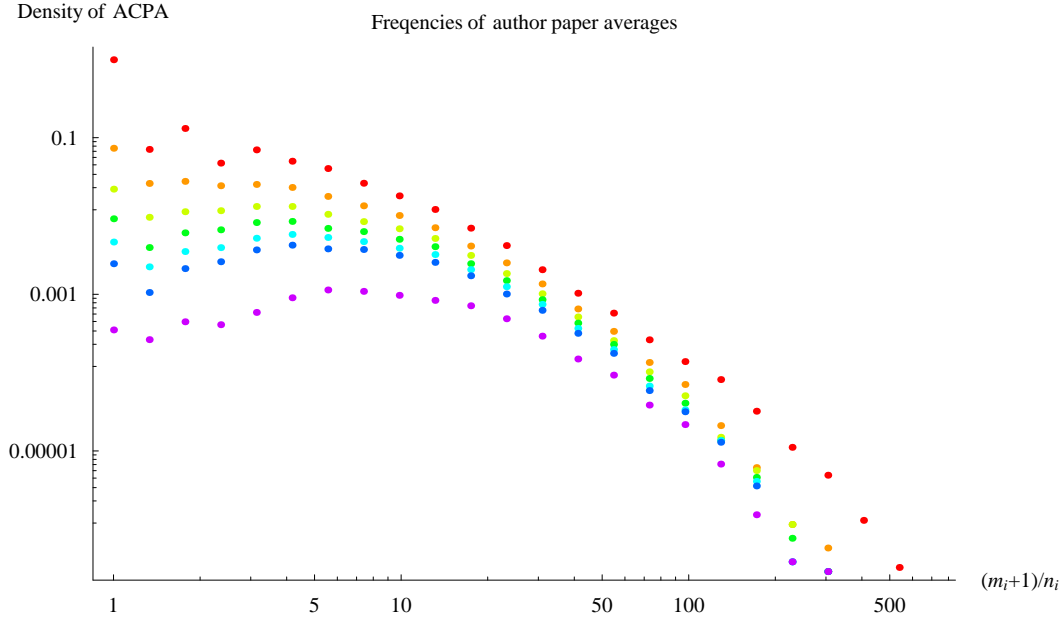


Figure 3.10: The change of the distribution of the average number of citations per paper per author (ACPA; red dots), as authors with less than 5 (orange), 10 (yellow), 15 (green), 20 (turquoise) 25 (blue), and 50 (purple) citations are removed from the database. The total distribution has global normalization one.

A large fraction of the authors with low average paper citations are also 'minimally publishing': When the length of the publication record *and* the total number of citations for individual authors is taken into consideration, we explicitly see that the minimally publishing authors are mainly removed from the low-average part of the distribution. Again, this is confirmed by the explicit numbers for the mean and median, *cf.* Table 3.4. For the distributions of authors with 10 or more publications, we even see the population developing non-zero peaks.

In this figure, another effect is also at play; namely that many of the authors with 'unnaturally' high averages are removed. This is due to the fact that the problem with accidentally cutting out short bits of highly cited authors' citation records, as a result of the 'David Gross'-effect, is (partially) removed. Because we are removing the shorter publication records. For example the two papers in which Erick J. Weinberg forgot to sign his middle initial, have been removed from the distribution. This effect actually makes the impact on the mean number of the average number of citations per paper per author a little less pronounced. It grows from $\langle \frac{m_i}{n_i} (n_i \geq 1) \rangle = 8.5(8.6)$ to $\langle \frac{m_i}{n_i} (n_i \geq 25) \rangle = 18.8$. The impact on the median is less sensitive to this effect, climbing from $\widetilde{m_i/n_i}(n_{\geq 1}) = 3.0(3.0)$ to $\widetilde{m_i/n_i}(n_{\geq 25}) = 12.2(12.5)$.

3.5 Summary

The first important result of this chapter is that to understand the structure of the author network, it is convenient to think about it in terms of two interacting levels: A level of authors and a level of papers, where the paper level is exactly the network, we discussed in Chapter 2. We saw how the citations and references between authors run via their papers.

We then proceeded to find that there are two intrinsic problems in counting papers author by author: One is that counting papers this way, weighs papers by their number of co-authors, both for the incoming and outbound distributions. To limit this effect, we only consider the theory subset of SPIRES, because this set typically has fewer authors (on average 1.8, compared to 9 for SPIRES as a whole) per paper than the rest of the database. When comparing the distributions of citations of the papers in the theory subset, counted as in the previous chapter and counted author by author, we found that these were—when each normalized to 1—indistinguishable, for all practical purposes. In other words, there is no ‘typical collaboration’ that tweaks the distribution; SPIRES is so homogeneous that weighing the papers by number of co-authors does not pose a problem. The other problem is what we called the ‘David Gross’-effect. This name is used to point to the problem that counting authors by last name and first initial alone, tends to underestimate the number of authors in the database, while counting authors by last name and all initials tends to overestimate the number of distinct authors in the database. We solved this latter problem by simply including both counts, using them as a lower and upper bound.

With the information on authors in SPIRES in hand, we began utilizing this new knowledge to plot the distributions of total author citations, total author publications, and average number of citations per paper per author. All three of these distributions were described by the double power-law structure that we already know from the distribution of paper citations, albeit, here, with different slopes. It is interesting in itself, to know the shape of these distributions, but these are also useful tools for determining the quality of authors in the database. Among the remarkable results from this part of the chapter, we found that 8(7) career publications is sufficient to reach the top 50 of all time authors in SPIRES. A depressing 18(19)% of all authors in SPIRES have got zero citations; clearly these kinds of statistics are due to the large number of authors with only a few publications: The mean number of publications in the theory subfield is 8.2(7.2) papers per author, but the median number of papers is mere 2(2) papers per author.

Considerations along these lines led us to introduce the concept of the ‘Scientific Staff’. This group consists of the people who have published more than 25 papers; we began investigating what happens to the network, when these authors are removed from the database. The main conclusion, with regard to this data, is that even though the authors with only a few publications are not at all as highly cited as the Scientific Staff, on average, their paper citations still follow the independent power-laws, $P(k_{i(j)}) \sim k_{i(j)}^{-\alpha_{sci}}$, with $\alpha_{sci,low} = 1.16(1.15)$ and $\alpha_{sci,high} = 2.88(2.87)$, and that there are highly cited authors amongst them—this is underscored by the flatter slopes of the power-laws. Conversely, we find that even though the Scientific Staff do better than the ‘minimally publishing’⁸ authors on average, the distribution of their publications still follow a power-law $P(k_{i(j)}) \sim k_{i(j)}^{-\alpha_{min}}$, with $\alpha_{min,low} = 1.49(1.48)$ and $\alpha_{min,high} = 3.06(2.99)$; this steeper slope corroborates with the evidence from the mean and median values in Table 3.4. In conclusion, *even the Scientific Staff publish a (vast) majority*

⁸In this instance, by ‘minimally publishing’, the authors with less than 25 papers are used as an example.

| | | The Scientific Staff | | |
|--------------|-----------------------------------|--|---------------|---------------|
| | | $n_i \geq 1$ | $n_i \geq 25$ | $n_i \geq 50$ |
| Total Mean | $\langle m_i \rangle$ | 133 (115) | 1108 (1123) | 2106 (2178) |
| Total Median | \tilde{m}_i | 8 (7) | 518 (528) | 1236 (1238) |
| PC Mean | $\langle k_{i(j)} \rangle$ | 16.2 | 21.9 | 26.4 |
| PC Median | $\tilde{k}_{i(j)}$ | 3 | 5 | 6 |
| ACPA Mean | $\langle \frac{m_i}{n_i} \rangle$ | 8.5 (8.6) | 18.8 (19.4) | 24.4 (25.4) |
| ACPA Median | \tilde{m}_i/n_i | 3.0 (3.0) | 12.2 (12.5) | 16.4 (17.2) |
| | | The ‘minimally publishing’ distributions | | |
| | | $n_i = 1$ | $n_i < 25$ | $n_i < 50$ |
| Total Mean | $\langle m_i \rangle$ | 6 (6) | 40 (38) | 70 (63) |
| Total Median | \tilde{m}_i | 1 (1) | 6 (5) | 7 (6) |
| PC Mean | $\langle k_{i(j)} \rangle$ | 6.0 | 9.7 | 11.8 |
| PC Median | $\tilde{k}_{i(j)}$ | 1 | 2 | 3 |
| ACPA Mean | $\langle \frac{m_i}{n_i} \rangle$ | 6.1 (6.3) | 7.5 (7.8) | 8.0 (8.2) |
| ACPA Median | \tilde{m}_i/n_i | 1.0 (1.0) | 2.5 (2.5) | 3.0 (2.9) |

Table 3.4: Reducing the author population: Overview of important quantitative results, *i.e.* the mean and median for the Total, the Paper Citation (PC), and the Average Citations per Paper per Author (ACPA) Distribution.

of *minimally cited papers*.

Thus, the (depressing) conclusion of the previous chapter is amply confirmed: the progress of science is truly driven by the work of a few excellent authors. Three (two) percent of the authors in the database are responsible for generating 49(46)% of the citations. Note that this result is radically stronger, than the corresponding conclusion from the previous chapter: That 4 percent of *papers* in the database produce 50 percent of the citations, is the weaker conclusion because we now know that even the members of Scientific Staff publish a majority of unknown papers! These conclusions are highly interesting and in the following Chapter, we will proceed to illuminate the same problem from a slightly different angle, *viz.* by considering the *author citation histories*.

In the sociological literature concerning citations¹, the way in which a paper accumulates citations is denoted a paper’s *citation history*. This information is not available for the papers in SPIRES, but moving up one level in the hierarchy of the network, we find something even better. The citation records of each author, the \mathbf{K}_i ’s, are—by construction—*author citation histories*; the chronologically ordered record of a given author’s publications and their citations.

Our investigation in Chapter 3 led us to single out two populations of authors in SPIRES: The Scientific Staff and the ‘minimally publishing’ population. In the theoretical investigation focusing on ‘*SPIRES: the complex network*’, these two populations are equally interesting and should both be included in the analysis; recall that what makes scale-free networks scale free, is the overwhelming number of nodes with a low number of in- and outbound links, and the small number of highly connected network ‘hubs’. But when we turn to the more utility-minded use of SPIRES—the investigation of scientific excellence—the difference between these two populations becomes much more interesting. When a physics department is out to employ a new researcher, it is not interesting to compare him to the entire population of physicists. This is perhaps the most important realization from the previous two chapters: The Scientific Staff is a small, dynamical part of a network swamped by an overwhelming majority of exanimate authors and papers. It is far more interesting to compare the prospective co-workers to the Scientific Staff of physics faculties all over the world. This chapter will focus on the more utility-minded aspects of the analysis of SPIRES. We will find that the concept of the author citation history can be used to argue that a minimum of 25 published papers is a good criterion for separating the Scientific Staff from the remaining authors in SPIRES.

4.1 Average Citation Histories

The author citation history offers a novel incision in the citation distribution. Let us take a look at what happens, when we compare first publications to first publications, second

¹ Cf. Vlachý for an introduction to ‘*scientometrics*’ [64]. This review paper also contains a comprehensive list of references.

publications to second publications, *etc*; we are going to investigate the properties of the $k_{i(1)}, k_{i(2)}, \dots$ distributions. In the process of studying these paper-by-paper distributions, we will come across several surprising conclusions.

4.1.1 Plotting the Different Averages

The first step in this direction is to take a look at the average number of citations, paper by paper; $\langle k_{i(1)} \rangle, \langle k_{i(2)} \rangle, \dots$, where these averages are defined as²:

$$\langle k_{i(\lambda)} \rangle \equiv \frac{\sum_j k_{j(\lambda)}}{N(\lambda)}, \quad \lambda = 1, 2, 3, \dots, \max(n_i). \quad (4.1)$$

The notation $N^{(\lambda)}$ is simply shorthand for the number of authors with citation records of length λ or greater, $N(n_i \geq \lambda)$. Note that, by this definition, authors with only one publication are not included in the $\lambda \geq 2$ averages, *etc*. We know from Chapter 3 that $N^{(\lambda)}$ is a rapidly decaying function of λ : The corresponding un-cumulated probability distribution is displayed in Figure 3.5.

Getting back to the subject at hand, the $\langle k_{i(\lambda)} \rangle$'s plotted against λ are displayed in Figure 4.1. Recall that the distribution of publications per author, is characterized by a two power-law structure. The presence, in Figure 3.5, of the second—and very steep—power-law that sets in at around $\lambda = 60$ is also visible in Figure 4.1, where fluctuations begin to dominate the average values for $\lambda \geq 60$. Beyond this point, the average value of the λ th paper is highly sensitive to whether or not a single author's λ th paper is highly cited; this is simply because of the relatively small number of papers involved in each average.

The primary aim of this chapter is to illuminate the differences between the Scientific Staff and the 'minimally publishing' authors. So far, we have used the convention that the database should be divided around authors with 25 publications. It seems clear that authors with 25+ publications should be considered full-fledged members of the Scientific Staff—with this number of publications, an author has had some sort of career in high energy physics and should be included in the considerations. However, deciding exactly who else to include under the label, 'the Scientific Staff'—and who to exclude is a difficult matter: There are authors in the minimally publishing sub-population who publish highly cited papers; this distribution also follows a power-law, *etc*. In the minimally cited part of the distribution there are, of course, also talented young authors, making their way through their first publications. To

²A few remarks on the choice of notation and language are necessary here. The word 'population' and 'sample' have different uses in the literature, *cf.* [62]. In this thesis the word 'population' is used to describe the total population of authors in SPIRES, and *not* a more abstract space. Because the statistical quantities mean, median, variance, *etc*, of a sample by definition are *unbiased estimators* of the population ditto, the word 'sub-population' will be used rather than the word 'sample', when we discriminate, in the following, between authors with different numbers of publications—because, as we shall see, these quantities for the sub-population of authors with more than 25 publications, are not unbiased estimators of the corresponding population measures. Whether or not we are discussing properties of a population or of a sample, also makes a difference notation-wise. The average of a population quantity is sometimes called the expectation value of that quantity and denoted $\langle \cdot \rangle$, whereas the notation for the sample average is a bar over the quantity in question; for example, the average of the random variable x , is denoted \bar{x} . Usually the mean, variance, *etc*, for the entire population are set in Greek letters, *e.g.* μ, σ , *etc*, and the corresponding quantities for a sample are typeset using Latin letters: m, s , *etc*. In this thesis, we will primarily be considering populations and sub-populations and therefore use the 'population'-convention of Greek letters everywhere. For a further discussion, the reader is referred to [62].

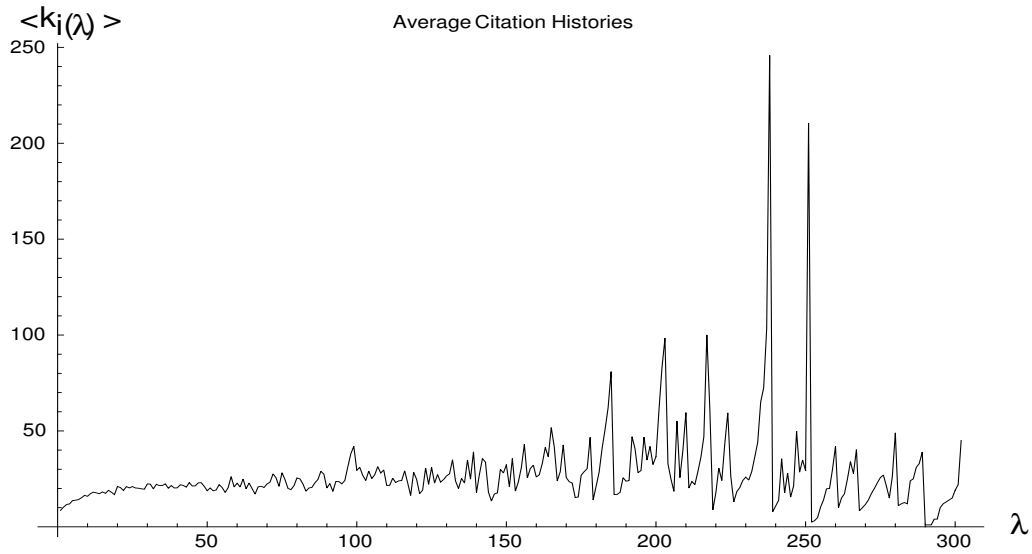


Figure 4.1: This plot is the average-number-of-citations history, that is, the average number of citations for the λ th publication, $\langle k_i(\lambda) \rangle$, plotted against λ . The longest list of publications, $\max(n_i)$, is 302 papers.

determine whether or not more authors should be included in the Scientific Staff, we will focus the analysis on the authors with $0 \leq \lambda \leq 25$ papers.

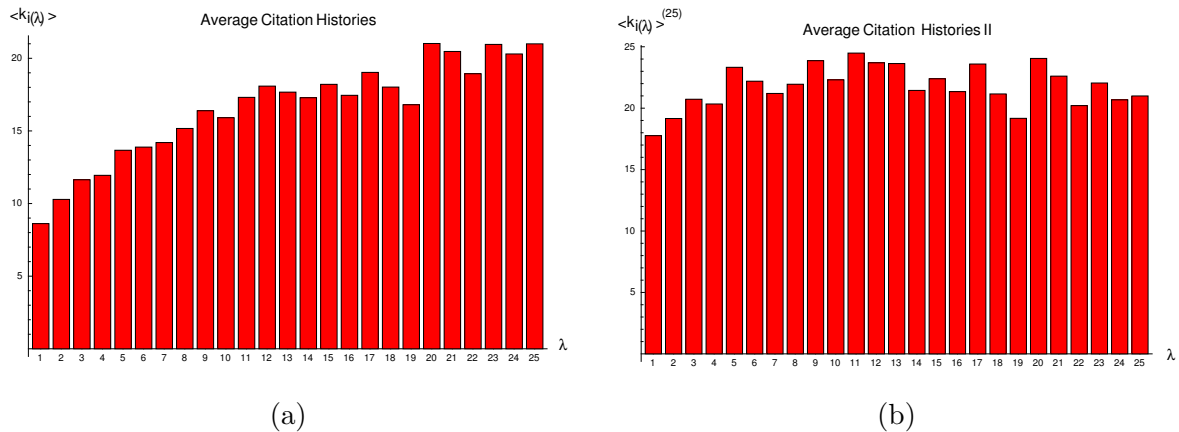


Figure 4.2: Clearly, the interesting part of Figure 4.1 is the first 25 publications. (a) This figure displays the first 25 elements of Figure 4.1, and (b) shows the corresponding distribution for authors with more than 25 publications. This second distribution is remarkably constant! Also: the scales on the y -axes of the two figures are different.

Because of the power-law nature of the paper-per-author distribution, and seeing as we are looking at average values, discussing this group of authors results in the added bonus of working with statistically significant material; again, *cf.* the fluctuations in Figure 4.1. The

sizes of the groups of scientists with $\lambda \geq 1, 5, 10, 15, 20, 25, 50$ publications is available in Table 3.3.

In Figure 4.2 (a), the first 25 averages from Figure 4.1 are plotted as a bar chart. There is nothing surprising about this figure. When we take all of the available authors into consideration, we find that the average number of citations for first publications is $\langle k_{i(1)} \rangle = 8.6(8.7)$, and that the average of the λ th paper grows steadily throughout the first 25 publications, where it seems to level out at around 22-23 citations per publication. This suspicion is confirmed in Figure 4.1, where it is clear that this trend continues up to around 60 publications. After this point fluctuations begin to dominate—although the average is still roughly centered around 20 citations.

The big surprise, however, comes from Figure 4.2 (b). Here, I have plotted the average number of citation histories, *for authors with more than 25 publications*. In other words, the modified averages:

$$\langle k_{i(\lambda)} \rangle^{(25)} \equiv \frac{\sum_j k_{j(\lambda)}^{(25)}}{N^{(25)}}; \quad \lambda = 1, 2, 3, \dots, 25, \quad (4.2)$$

where the superscripted ‘(25)’ signals that we are only considering authors with 25 or more publications. For $\lambda > 25$, we simply define $\langle k_{i(\lambda > 25)} \rangle^{(25)} \equiv \langle k_{i(\lambda)} \rangle$. These averages are extremely interesting because, for the Scientific Staff, they are remarkably *constant* over their first publications. The average of the columns in Figure 4.1 (b) ($\sum_{\lambda=1}^{25} \langle k_{i(\lambda)} \rangle^{(25)} / 25$) is 21.8(22.5) citations—with standard deviation 1.7 (2.0). This number corresponds extraordinarily well with the average number of citations per paper for authors with more than 25 publications $\langle k_{i(j)} \rangle^{(25)} \approx 21.9(22.8)$. The correspondence is remarkable. It appears that the average number of citations per paper, is constant throughout publication histories for authors with long publication records. In other words, we have discovered one important aspect in which the Scientific Staff stands apart from the remaining distribution. On average, their first papers are as good as any paper they will ever publish; when we take Figure 4.1 into consideration, we can state this more generally: Each paper published by a member of the Scientific Staff (on average) has a high *and constant* level of quality. In this respect, the approximately 8.7 (7.0)% the authors in the database, who publish more than or equal to 25 papers, have properties that differ radically from the properties of the database as a whole (34,434(39,921) authors).

4.1.2 Drawing the Line

Thus, a hypothesis is beginning to take form: The people we have denoted the Scientific Staff are exactly the people who publish papers of high quality *from day one*, and who continue to publish papers of constant, high quality. If this hypothesis is at least roughly correct (which is what will be argued in the following), then an indication that 25 papers is a sensible cut, stems from the fact that the average of the columns in Figure 4.2 (b) is virtually identical to the average number of citations per paper for the *total* distribution of papers, written by authors with more than 25 career publications.

Generalizing this idea, we can present a more quantitative argument for why the cut should not be 10, 15, or 20 publications. If we assume that the Scientific Staff produces papers of a constant quality, then the difference

$$h(\Lambda) \equiv \langle k_{i(j)} \rangle^{(\Lambda)} - \Lambda^{-1} \sum_{\lambda=1}^{\Lambda} \langle k_{i(\lambda)} \rangle^{(\Lambda)}, \quad (4.3)$$

is an informative quantity. The two terms on the rhs. express properties of the $\Lambda+$ sub-populations, and compares the average of the total sub-population to the average of the first Λ distributions. Thus Equation (4.3) should approach zero, as Λ approaches the ‘correct’ size for the Scientific Staff. Of course this criterion is merely a rule of thumb, and it only makes sense under the assumption that the average number of citations of the λ th publication, is constant for the Scientific Staff.

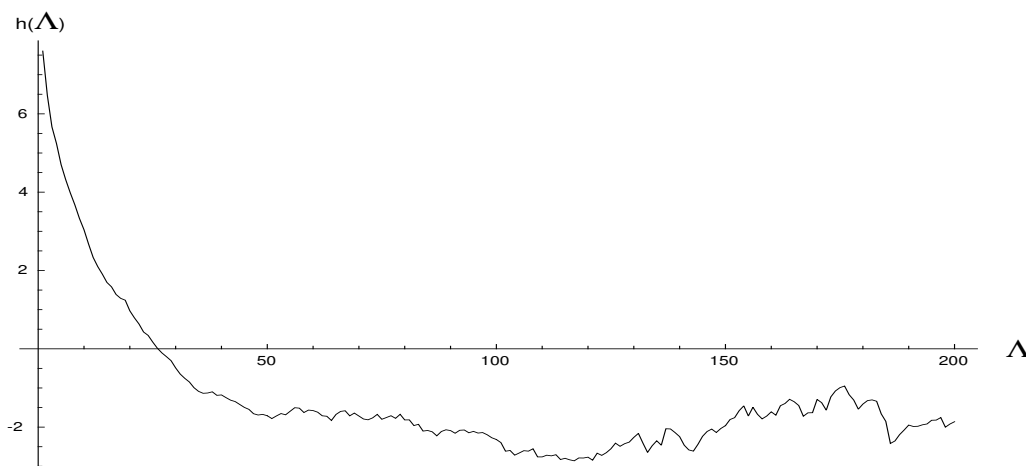


Figure 4.3: The difference $h(\Lambda)$ —defined in Equation (4.3)—plotted against Λ , in the range $1 \leq \Lambda \leq 200$. This difference is 0 for $\Lambda = 26(27)$, after which it continues to drop to -2, where it remains roughly constant.

In Figure 4.3 the difference $h(\Lambda)$, defined in Equation (4.3), is plotted for $1 \leq \Lambda \leq 200$. For $\Lambda = \max(n_i)$ we have that $h(\Lambda) = 0$, since in this case, the two terms on the rhs. in Equation (4.3) become identical. This, however, is of periphery importance since the criterion begins to lose its meaning for large values of λ : The average number of citations per publication, for authors with publication records of more than 200 publications, is highly dependent upon properties of the individual authors—simply because there are only 15(11) authors with more than 200 publications in the database. The difference $h(\Lambda)$ vanishes for $\Lambda = 26(27)$, and then continues to drop off (!), assuming a roughly constant value of -2 . This means that our assumption regarding constant quality is not entirely correct; the balance between the two terms shifts in the interval $25 \lesssim \Lambda \lesssim 50$. That the balance between these two terms ‘tips over’, indicates that many authors with more than 50 publications actually publish some of their best work early in their career. Although not entirely correct, the hypothesis of a relatively constant quality work by the Scientific Staff is, however, still supported by Figure 4.3—a difference of merely two papers on average, still reflects to a rather constant average level of quality.

In summary, 25 publications is a reasonable place to draw the line between the Scientific Staff and the remaining sub-population of authors. The interval from 25 to 50 in Figure 4.3 is rather interesting: If the cut was made at an even higher number of publications, the average number of citations for the λ th paper would *not* be as constant; we would see higher averages for the early work by this remaining, elite group. Placing a cut at around 50 citations, is not in our best interest. The authors that are removed from the database, if this interval is not included, may not have exactly the same citation patterns as the 50+ authors, but they are

full-fledged members of SPIRES nonetheless, and should *not* merely be regarded as ‘noise’, swamping the Scientific Staff.

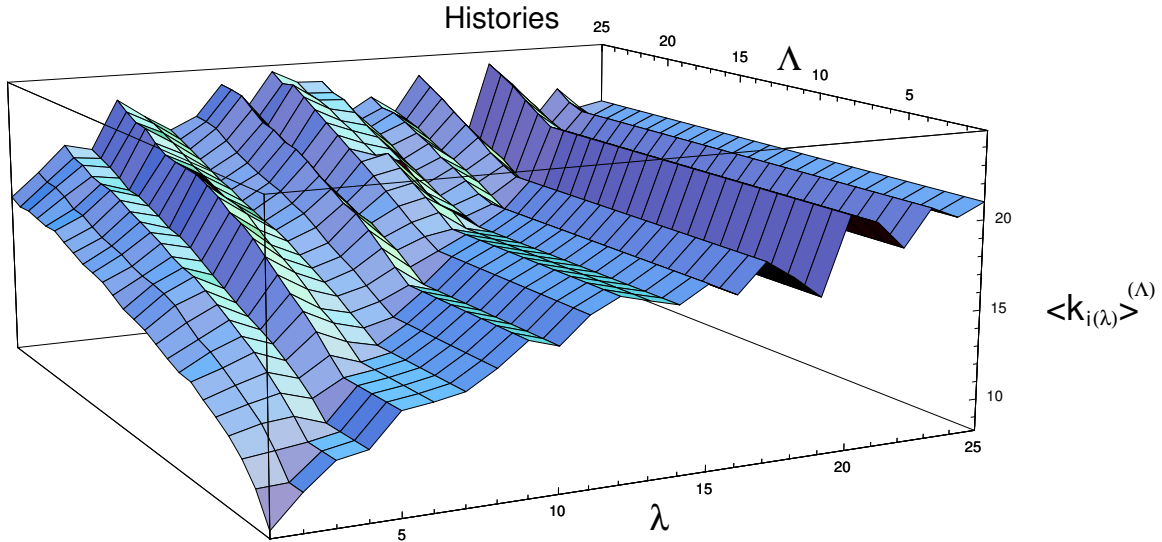


Figure 4.4: The evolution from Figure 4.2 (a) to (b). The z -axis is the average number of citations, paper-by-paper, $\langle k_{i(\lambda)} \rangle^{(\Lambda)}$. Recall that Λ signals the minimum length of the publication record of the authors involved in the average. The x -axis keeps track of Λ ; and the y -axis displays λ , the publication number in question.

Another way to think about what happens as the population of authors is restricted, is displayed in Figure 4.4. This figure illustrates the evolution from Figure 4.2 (a) to (b), as Λ grows from 1 to 25. The considerations from above are confirmed: As Λ grows, the average number of publications grows steadily until reaching the constant level from Figure 4.2 (b). There are no surprises here.

4.2 Total Distribution Histories

So far all that we can say is, of course, that the Scientific Staff have a constant, and markedly higher *average* number of citations per publication for their first publications, than the total population of the database. We would like to say something stronger than this; we have no *a priori* reasons to suspect that the $k_{i(1)}$ -distribution looks anything like the *distribution* of second- or 25th papers; the investigation in Chapter 3, however, leads us to expect to see the seemingly ubiquitous power-laws.

4.2.1 Unfolding the Averages

The most intuitive thing to do, in order to get started, is to simply take a look at the distributions of citations for the λ th paper—since these are precursory to the averages. These

are plotted in Figure 4.5.³

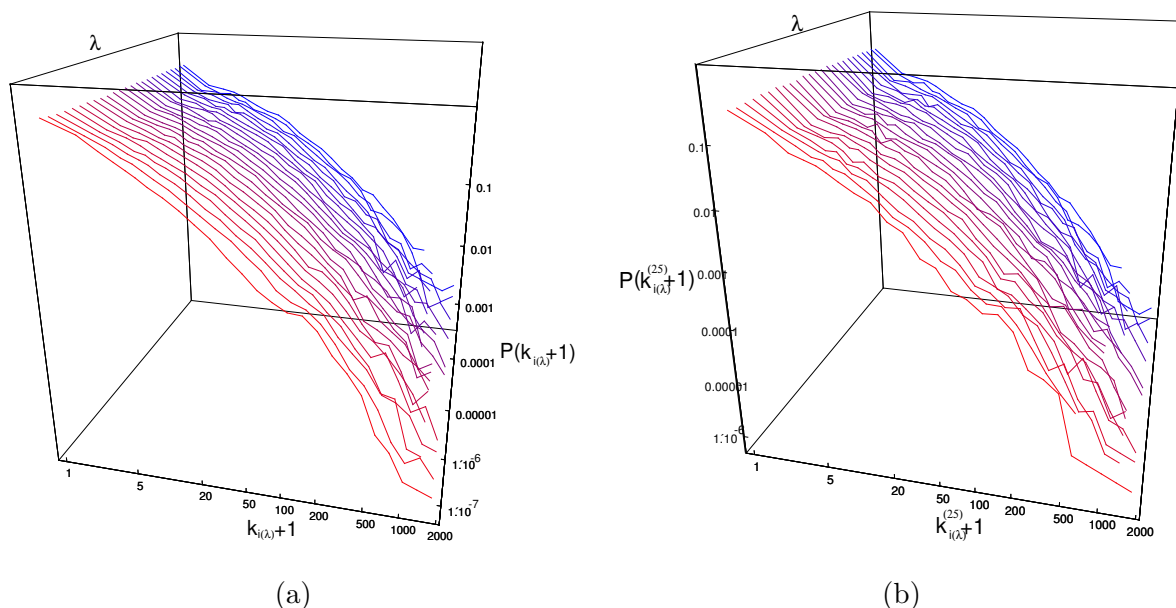


Figure 4.5: The averages from Figure 4.2 ‘unfolded’ so that the entire distribution that resulted in each average is visible. The red distributions are small λ ’s and the blue distributions are for λ closer to 25. Thus, in (a) we see the distribution of first, second and so forth papers for the entire distributions. In (b) the corresponding distributions are plotted for the 25+ data. Each distribution is normalized to one and plotted on log log scales.

Unfortunately, this figure is too complex to extract any precise information from. It is clear that ‘unfolding’ the averages from Figure 4.2 results in 25 reasonably similar distributions. But, when it comes to pin-pointing any specific differences amongst these, we are at a loss. Each and every distribution looks the same plotted on the double-log scales. Displaying these 3D structures on the 2D surface of a piece of paper, is problematic. Again, it is a testament of the enormous homogeneity of SPIRES that the same two power-law structure is visible in the author distribution histories, both for the database as a whole *and* for the Scientific Staff. The fact that most publications are quickly forgotten, is truly an integral part of the dynamics of science: Even the best of scientists publish a majority of papers that are virtually forgotten the minute they are published.

In Figure 4.6 (a), we explicitly see the difference between the citation probability distributions for the first paper of every author in the database and the same distribution for the twenty-fifth paper by the 25+ authors. The probability of receiving 1000+ citations is about an order of magnitude higher for a member of the Scientific Staff, than for the database taken as a whole, but the differences between these two distributions are not dramatic. Although we have not previously discussed the author distribution histories, last chapter’s discussion of the distribution of paper citations, as authors were removed (*cf.* Figures 3.8 and 3.9) should have prepared us for what we are seeing here.

Thus, Figure 4.6 is one possible way of clarifying what is going on in Figure 4.5. In this particular 2D representation, the 1st and 25th paper distributions have been plotted on the

³To do these two plots, the distributions are placed in a 25×25 matrix—this means that the data point corresponding to the most cited paper in the 25th paper distribution, is not visible in Figure 4.5.

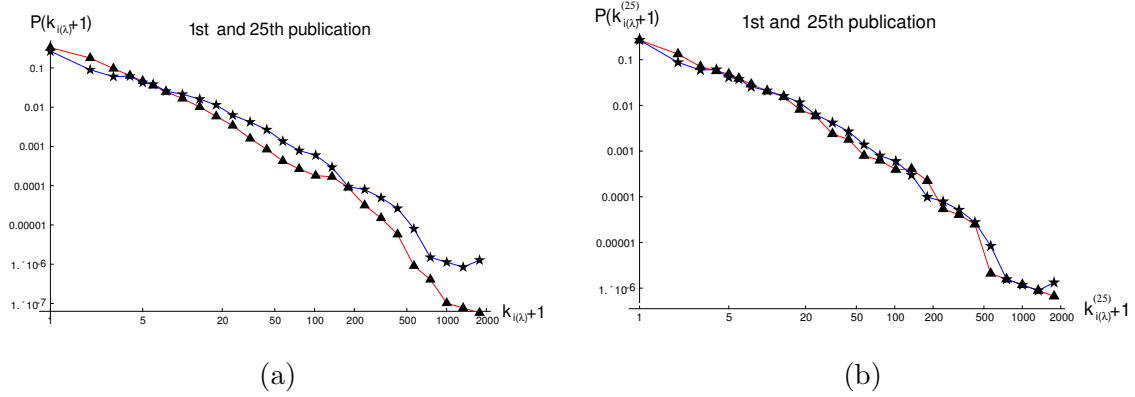


Figure 4.6: The first (red, \blacktriangle) and twenty-fifth (blue, \star) distribution from Figure 4.5. All four distributions are normalized to one. In (b), the similarity between the distributions of first and twenty-fifth publication for authors with more than 25 publications, is remarkable. The graphs are plotted on double log scales.

same graph, since they are the two distributions in each of the subplots in Figure 4.5, that intuitively are most different. In the analysis above, we have already taken advantage of the fact that it is far easier to locate differences between the distributions in this, much simpler, representation of the data.

Again, it is the (b)-part of the figure in question that is the most interesting. The first and twenty-fifth paper distribution are amazingly alike for the Scientific Staff. The likeness between the two distributions in Figure 4.6 (b) is uncanny. This figure also gives us a direct look at just *how* remarkably homogeneous Scientific Staff is. The data collapse is massive: When it comes to the distribution of citations of individual papers for the Scientific Staff, it appears that it is not a problem to lump all of the data into one distribution, simply because of the tremendous homogeneity of this group of authors.

4.2.2 Percentiles

We know from Figure 4.5 that all of the histories of distributions of citation look amazingly alike, no matter which sub-population of authors we discuss. Therefore, Figure 4.6 is a reasonable way to begin to form an impression of the change of the distribution of authors, restricted to the Scientific Staff. There is, however, another way of projecting the 3D information from Figure 4.5 onto a 2D plot that allows us observe the structure of each paper distribution much more clearly and explicitly. This is done in Figure 4.7.

Here, instead of plotting the entire distributions, notable percentiles for each λ have been used to reduce the 3D info. The number of publications needed for an author's λ th publication to be in the 50th (purple, \blacktriangle), 90th (green, \blacksquare), 95th (blue, \star), and 99th (red, \blacklozenge) percentile is plotted versus λ . Again, the familiar pattern emerges: In the (a)-part of the figure, there is a constant growth as λ increases from 1 to 25. The number of publications needed to be part of each percentile, are much lower than for the Scientific Staff (the scales on the y -axes are different in these two sub-figures). For $\lambda = 25$, the two plots merge (by definition). Naturally the figures for the 99th percentile are somewhat fluctuating for both groups of authors.

Our hypothesis that, for the Scientific Staff, the distributions are remarkably uniform

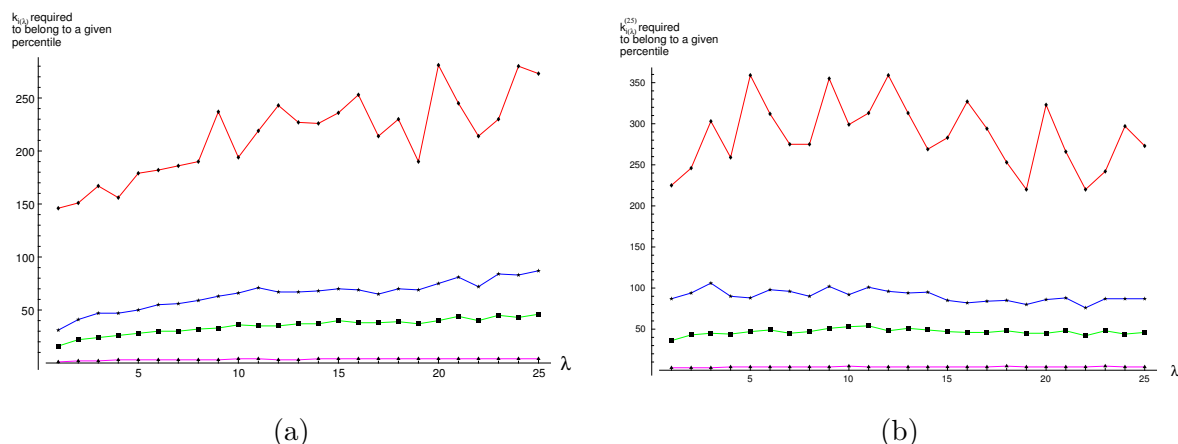


Figure 4.7: The number of citations that an author's λ th publication needs in order to be part of the 50th (purple, ▲), 90th (green, ■), 95th (blue, ★), and 99th (red, ◆) percentile. The (a)-part of the figure corresponds to the other (a)-parts in this chapter and displays the data for the first 25 papers, where all possible authors are included. Figure 4.7 (b), shows the information for the 25+ authors. The reader should note that the scales on the y -axes are different.

throughout the first 25 publications is confirmed in Figure 4.7 (b). This figure also strengthens the credibility of the hypothesis of an almost constant quality of publications for the Scientific Staff. However, there should sound a *caveat lector* here: We know that $h(\Lambda) \rightsquigarrow -2$ for long publication records. This means that authors with more than 50 publications, on average have received more citations for their earlier work. The reason we cannot see this imbalance directly in the data, is that the work of authors with fewer (25-50) papers is included in the paper-by-paper averages for these values of λ . In essence, this 'waters out' the highly successful early/middle years of the authors with 50+ publications. This results in the remarkably constant averages we see for the $1 < \lambda < 60$ part of Figure 4.1; and in the similar probability distribution for each of these λ s. That the speculation on the 'watering out' is true, is verified in the last plot, Figure 4.8, of this chapter. This figure is similar to Figure 4.7 (b), except here the data for the 50+ sub-population is displayed. From this figure, it is clear that this group of established authors actually were (somewhat) more cited in their early careers.

4.3 Summary

In this chapter, the concept of author citation histories was introduced. We used this concept to argue two things. Firstly, that the average quality (ability to attract citations) of publications by the Scientific Staff, is roughly constant starting with their very first publications by and throughout their careers. Secondly, using this fact as a criterion, we found that 25 publications was actually a reasonable place to draw the line between the 'minimally publishing' population and the Scientific Staff.

More concretely, we made this distinction on the basis of the *average* author citation histories, that is, the average value of all first papers, the average value of all second papers, *etc.*, and comparing these to the average of the total distribution of authors with less than a given number of publications, *cf.* Equation (4.3). We then proceeded to expand our attention to the history of entire distributions, *i.e.* we still compared first publications to first publications,

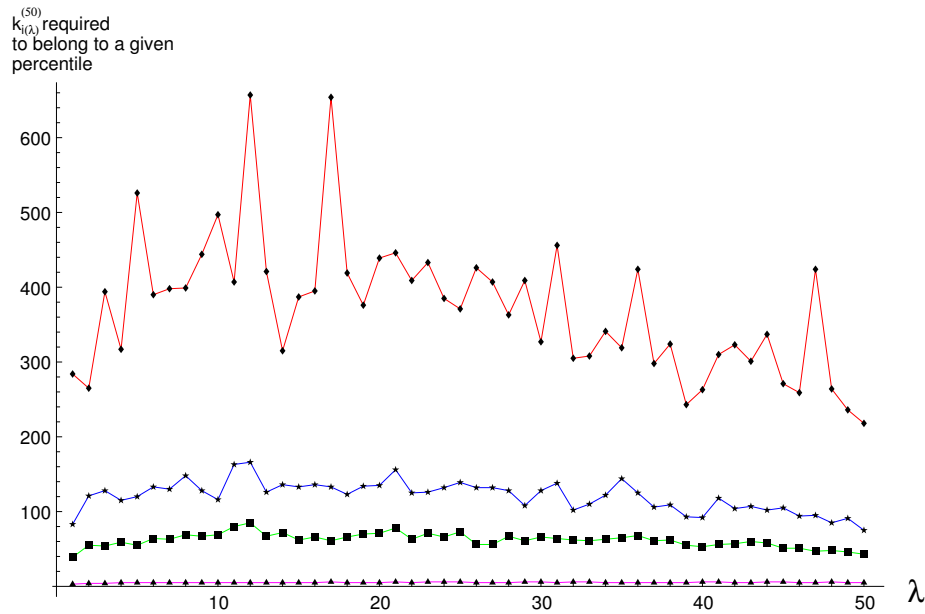


Figure 4.8: The percentiles for the 50+ author sub-population. The legend is identical to the one in Figure 4.7. This group of authors have high citation rates for their early papers ($5 \leq \lambda \leq 30$ papers). This higher rate of publication is also reflected in higher averages, $\langle k_{i(\lambda)} \rangle^{(50)}$.

etc. As expected, these distributions of paper citations were all described by the familiar two power-law structure. The distribution history approach re-verified the homogeneous nature of the group of authors with 25+ publications. Especially the impressive resemblance between the distributions of 1st and 25th publications for this group of authors, in Figure 4.6 (b), is convincing. The next point of action was to discuss the same subject once again but, this time, on the basis of the number of papers needed to belong to selected percentiles. Comparing the ‘histories’ of the percentiles in Figure 4.7 reconfirmed the homogeneity of the Scientific Staff.

However, further investigations resulted in discovery a slight inhomogeneity—hidden as a ‘longitudinal’ correlation within the Scientific Staff. With Figure 4.8, we realized that the citation records of the 50+ authors were not entirely as homogeneous (although still homogeneous) as we expected: It turns out that the authors with 50+ publications actually did a little better⁴, citation wise, in their early publications than in their later years. This inhomogeneity does not alter the fact that the Scientific Staff, *i.e.* the authors with more than 25 publications, are a remarkably homogeneous group within SPIRES.

That the authors of the Scientific Staff share the remarkable property that the citation distributions of their first publications are virtually identical to the distributions of citations in their later work, allows us to draw a very interesting conclusion. Using statistical methods, we can utilize this property of the Scientific Staff, taken as a whole, to make predictions as to how well young authors, with only a few publications, are likely to do in the future!

⁴Or rather, they did extremely well, *cf.* the spikes for $\lambda = 12, 17$, in Figure 4.8, and the number of citations needed to belong to some of the more exclusive percentiles (95th and 99th) in this figure, compared to the same numbers in Figure 4.7.

Principal Component Analysis

In the previous chapters, we have spent a great deal of time and energy on isolating the group of authors in SPIRES that we have named ‘the Scientific Staff’. We know now that these authors are the backbone of SPIRES—2,983(2,809) people who are responsible for around half of the citations generated in theoretical high energy physics since 1945. These are the people whom the universities have granted a desk, pen and paper, and maybe even access to a personal computer; in other words, these are the people who occupy positions at physics departments around the world. Therefore, the Scientific Staff is the group of people we want to compare ourselves and new authors to.

In the previous chapters, we have slowly introduced the notion of so-called longitudinal correlations in SPIRES. For instance, these correlations are the reason that we did not use χ^2 -tests, when comparing the subfield data in Chapter 2. The assumption for the χ^2 -test to be meaningful is that the data placed in each bin is statically independent. However, we know that it is not—there are correlations among the bins, simply because some authors are more cited than others. These correlations across the bins—hence the expression ‘longitudinal’—are also the reason that defining the Power of Excellence, in Section 2.4, was a meaningful enterprise.

In the context of the longitudinal correlations, it is clear that the Scientific Staff is much more interesting than the ‘minimally publishing’ authors—because, considering the extreme case, there cannot be any correlations associated with authors with only one paper. Furthermore, because the ‘minimally publishing’ population vastly outnumber the Scientific Staff in the database, their inclusion in any investigation of the correlations in the data, would ‘swamp’ any interesting discoveries that we could make about the Scientific Staff. Therefore, we will only be concerned with the Scientific Staff in this chapter.

We are going to use the multivariate statistical method of *principal component analysis* (PCA) to examine the longitudinal correlations of the Scientific Staff. Put differently, we will investigate SPIRES to find out what ‘types’ of authors are to be found within the ranks of the Scientific Staff. This will enable us to solve some of the problems with the Power of Excellence; recall, for example, that some authors who are ‘aggressively inept’, such as the author \mathfrak{U} , from Section 3.4, manages to be more ‘improbable’ than many accomplished authors. This

particular author has published more than 50 papers with only 4 total citations, which has resulted in a higher Power of Excellence than most authors in the population. Obviously, an analysis based on a more complicated concept, such as PCA, is not as intuitively appealing as the simple probabilistic measure of the ‘improbability’ of authors from Chapter 1, but it gets the job done.

5.1 Preliminaries

First, let us review a few key concepts necessary to understand the method of principal components.

5.1.1 Multivariate Statistics

The first point of order is a quick brush-up on some key concepts from multivariate statistics. I assume that the reader is familiar with univariate statistics. We start out with a set of p measurements on each of \mathcal{N} distinct objects (in our case authors). If we let y_{ij} denote the original measurements, such that y_{1k} is the measurement of the k th author’s first parameter, and the the same author’s citation count in the next bin is y_{2k} citations, *etc*; we can organize our data in a ‘measurement matrix’, $\mathbf{Y} = (y_{ij})_{p \times \mathcal{N}}$, where the information about the k th author is listed in the column-vector,

$$\mathbf{y}_k = \begin{pmatrix} y_{1k} \\ y_{2k} \\ \vdots \\ y_{pk} \end{pmatrix}. \quad (5.1)$$

Now, let us define y_j to be a stochastic variable defined on the ensemble of the measurements in the j th row-vector $(y_{j1}, y_{j2}, \dots, y_{j\mathcal{N}})$. The average of y_j is denoted μ_j (mean of citation counts in the j th bin). To keep things compact, we can write these population¹ means in a column-vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)'$. Similarly, the variance of the measurements in the j th bin is given by $\sigma_j^2 = \langle (y_j - \mu_j)^2 \rangle$, where the average runs over all elements of y_j .

However, if we consider the fact that the different y_j ’s come together as a group in a stochastic *vector* variable, \mathbf{y} , we must consider the fact that there might be relationships among the y_j ’s described by joint probability distributions. A measure of how two stochastic variables vary together is the *covariance*. The covariance between y_j and y_k is defined as

$$\sigma_{jk} = \langle (y_j - \mu_j)(y_k - \mu_k) \rangle, \quad (5.2)$$

where this average runs over all possible pairs of values that y_j and y_k may take on together. Let us inspect Equation (5.2) to gain a little intuition about the concept of covariance. Since the covariance is defined as the average of the product of the deviations of y_j and y_k from their respective means, we have that if ‘large’ (as in larger than the mean) values of y_j and y_k tend to happen together (or similarly for ‘small’ values, *i.e.* smaller than the mean), the two deviations $(y_j - \mu_j)$ and $(y_k - \mu_k)$ will be positive (and in the ‘smaller’ case, both negative) at the same time, so that the product is positive, which will make the average in Equation

¹For conventions regarding the use of the terms ‘population’ and ‘sample’—and on notation related to the statistical analysis in general, the reader is referred to Footnote 2 in Chapter 4.

(5.2) positive. A similar argument can convince us that if ‘large’ values of y_j tend to coincide with ‘small’ values of y_k , then the covariance will be negative. Finally, if the two variables in question are truly unrelated, the products in Equation (5.2) will tend to cancel out when averaged over the entire population (obviously $\sigma_{jk} = 0$ is a necessary condition for the two variables to be uncorrelated, not a sufficient one). In summary, the quantity ‘covariance’ can intuitively be considered a measure of how associated the values of y_j are with y_k ’s.

The information from these individual covariances is usually gathered in a symmetric $p \times p$ matrix called the *covariance matrix*:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{pmatrix}. \quad (5.3)$$

It is sometimes informative to separate the ‘spread’ contained in the variances from the ‘association’. To accomplish this, we can define a special measure of association that takes into account that different elements of \mathbf{y} may vary differently on their own. This *population correlation coefficient* is defined as $\rho_{jk} = \sigma_{jk} / (\sqrt{\sigma_j^2} \sqrt{\sigma_k^2})$. In this definition, note that $\sqrt{\sigma_j^2} = \sigma_j$ is the standard deviation of measurements in the j th bin, and correspondingly for y_k . Thus, ρ_{jk} scales the information on association in the covariance in accordance with the magnitude of variation in each variable. With the correlation coefficient, we can think about associations among variables measured on different scales. This information is usually summarized in a matrix, called the correlation matrix, or $\mathbf{\Gamma}$, that is organized analogously to the covariance matrix. Finally, note that knowledge of the variances *and* $\mathbf{\Gamma}$, is equivalent to knowledge of $\mathbf{\Sigma}$, the covariance matrix. The subject of multivariate statistics is reviewed comprehensively in [65, 66].

5.1.2 Matrix Algebra

The next point of order stems from matrix algebra, and is due to the fact that the method of principal components is based on a classic result from this field that a $(p \times p)$ square, symmetric, nonsingular matrix (*e.g.* the covariance matrix), can be reduced to a diagonal matrix $\mathbf{\Lambda}$ by pre- and postmultiplying it with the orthonormal matrix \mathbf{U} , such that

$$\mathbf{U}' \mathbf{\Sigma} \mathbf{U} = \mathbf{\Lambda}, \quad (5.4)$$

where the diagonal elements of $\mathbf{\Lambda}$, $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigenvalues of $\mathbf{\Sigma}$, and the columns of \mathbf{U} , $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ are the eigenvectors of $\mathbf{\Sigma}$. Proof and a more mathematically pleasing formulation of this theorem can be found in any book on matrix algebra, or in [62].

Geometrically, one can think of the procedure of diagonalizing a matrix (in this case $\mathbf{\Sigma}$), as a *principal axis* rotation of the original coordinate axes y_j about their means μ_j . The elements of the eigenvectors are precisely the direction cosines of the new axes related to the old.

5.2 The Method of Principal Components

With these preliminaries out of the way, we are now ready to discuss the theory of PCA. For a more comprehensive review, the reader is referred to the literature, *cf.* [65, 66, 67]. In

all simplicity, PCA is a multivariate technique in which a number of correlated variables are transformed (linearly) to a set of uncorrelated variables whose variances are as large as possible.

The starting point is the covariance matrix, Σ . We use the principal axis transformation described in Section 5.1.2, to transform our p correlated variables y_1, y_2, \dots, y_p into p new uncorrelated variables z_1, z_2, \dots, z_p . Now, it can be shown that the axis along which the variance is maximal, is the eigenvector \mathbf{u}_1 of the the matrix equation

$$\Sigma \mathbf{u}_1 = \lambda_1 \mathbf{u}_1, \quad (5.5)$$

where λ_1 is the largest eigenvalue of Σ , and the variance along the new axis. The second eigenvector is defined as the one with the second largest eigenvalue (and thus the second largest variance); sorted after descending size of eigenvalue, the other principal axes and eigenvectors obey similar equations. We find that the matrix \mathbf{U} of all eigenvectors forms a new set of orthogonal axes that are ideally suited for a description of our data set. We can think of the process of finding the new variables, as translating the origin of the original coordinate system to $\boldsymbol{\mu}$ and then rotating the coordinate axes until they pass through the directions of maximum variance. We end up with the transformation:

$$\mathbf{z} = \mathbf{U}'[\mathbf{y} - \boldsymbol{\mu}], \quad (5.6)$$

where \mathbf{y} is the measurement variable and $\boldsymbol{\mu}$ is the means; both are $p \times 1$ vectors. We call the transformed variables the *principal components* (PC's) of \mathbf{y} . Clearly, the i th principal component is given by

$$z_i = \mathbf{u}_i'[\mathbf{y} - \boldsymbol{\mu}], \quad (5.7)$$

and has mean zero and its variance is the i th eigenvalue, λ_i . Making a final distinction, we shall call the transformed *variables* for PC's and the individual transformed *observations* of authors, the \mathbf{z}_k 's for z -scores².

5.2.1 Conservation of Variability

One important goal of multivariate analysis is to be able to summarize results about the entire multivariate distribution with a few generalized measures. One way of generalizing the variance into include only one number, is to consider the sum of the variances of variables,

$$\text{Tr}(\Sigma) = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_p^2. \quad (5.8)$$

Notice that since we know from linear algebra that the trace is invariant under change of basis, we have that $\text{Tr}(\Sigma) = \text{Tr}(\Lambda)$, which means that the sum in Equation (5.8) is a preserved quantity under the principal axis rotation.

This is an important result because it shows that the eigenvalues (that are the variances of the principal components) may be treated as variance components. Further, the ratio of each characteristic root to the total *will indicate the proportion of the total variability accounted for by the individual PC's*. This result will prove its worth in the following: If the main variance of the data set lies in a small dimensional space (the first few eigenvalues), then one can gain a great deal of understanding of the data from the projection onto the first

²The use of the word 'score' has its genesis in psychology and education where PCA analysis is often applied.

few eigenvectors alone. This property is a primary reason why the principal components are interesting variables, and why the set of uncorrelated variables can be considered ‘smaller’. This subject will be discussed in further detail in Section 5.3.4. Now, let us stop ‘beating around the bush’ and use the SPIRES data to illustrate and exemplify these and further properties of PCA.

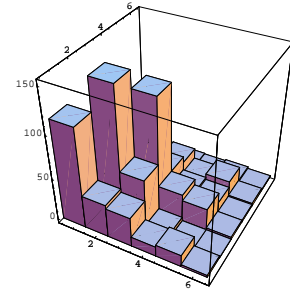
5.3 The SPIRES Covariance Matrix

5.3.1 Finding the Right Bins

Before we can get started with the PCA, we need to set up a covariance matrix based on the SPIRES data. With regard to the SPIRES database, we are interested in the longitudinal correlations in the paper citation distribution. We want to know how the papers of individual authors are correlated across the distribution of citations. To use a concrete example, take bins from the ‘search summary’ in Section 2.4 and distribute each author’s papers in these to produce the \mathbf{y}_k ’s. Gathering this information for the entire Scientific Staff, allows us to compute a 6×6 covariance matrix³.

$$\Sigma_{test} = \begin{pmatrix} 117. & 38.9 & 34.7 & 11.3 & 13.7 & 2.19 \\ 38.9 & 153. & 59.1 & 4.12 & -1.32 & -0.495 \\ 34.7 & 59.1 & 134. & 32.7 & 25.8 & 2.31 \\ 11.3 & 4.12 & 32.7 & 16.0 & 13.5 & 1.58 \\ 13.7 & -1.32 & 25.8 & 13.5 & 20.7 & 3.02 \\ 2.19 & -0.495 & 2.31 & 1.58 & 3.02 & 0.886 \end{pmatrix}$$

(a)



(b)

Figure 5.1: The covariance matrix for the Scientific Staff with bins defined by the ‘search summary’ option on the SPIRES webpage. In (a) the covariance matrix is written out explicitly, and in (b), it is visualized as a bar chart, where the height of the bars signify the size of the σ_{jk} ’s. In this visual representation, the sizes of the covariances ‘jump out’ of the page. To optimize the amount of significant information conveyed to the reader, only the ‘First Initial (FI)’-data is used in this figure.

This 6×6 matrix is written out in Figure 5.1. As an added bonus, Figure 5.1 illustrates the problem with using the SPIRES bins for the PCA analysis. The more devoted readers probably recall that the reason we used the SPIRES ‘search summary’ bins, was that this information is readily available for every author in SPIRES. From the SPIRES website, this information can be obtained by the touch of a button. Hence, given the probabilities for having papers in each bin, revealed by the analysis in Chapter 2, Table 2.3 (or, in the case of

³The ‘search summary’ categories and corresponding probabilities for the Scientific Staff, are listed here in the format: ‘Category (bin-interval): ‘First Initial’-probability [‘All Initials’-probability]’. Unknown papers (0): 250 [0.248], Less known papers (1–9): 0.385 [0.378], Known papers (10–49): 0.269 [0.274], Well-known papers (50–99): 0.025 [0.0544], Famous papers (100–499): 0.043 [0.0420], Renowned papers (500+): 0.00359 [0.00383]. The probabilities listed in Table 2.3 are for the entire distribution, without the weighing of papers due to counting author-by-author. Therefore, they are different from the corresponding probabilities for the Scientific Staff.

the Scientific Staff, in Footnote 3), anyone can calculate their own—or anyone else’s—Power of Excellence.

The problem with the SPIRES bins is that the 2nd and 3rd bins are too large. These two bins span the interval 1 – 49 papers, containing some 65.4(65.2)% of papers produced by the Scientific Staff. Having only two bins in this highly interesting middle interval—interesting because it contains papers of high quality, papers that are not exactly classics, but surely papers that are useful to other scientists—is surely too little. As promised, this is clear from Figure 5.1, where the variances of these two bins completely dominate the covariance matrix. To solve this problem, we are going to have to give up the advantage that came from everyone being able to calculate their own Power of Excellence⁴, simply by using the ‘search summary’ option at the SPIRES website and define a set of *new* bins. Before we move on, observe that another point is being made in Figure 5.1, *viz.* that the visual representation of the covariance matrix, along with the considerations in Section 5.1.1, gives us a very intuitive grasp of the correlations in the data.

Getting back to the subject of choosing the right bins: We want to find just the right number, so that no structure eludes us—all the while, we do not want too many bins, since choosing too many bins tends to chop up the picture; resulting in unwanted noise. A promising line of attack derives from the investigation of the Scientific Staff in Chapter 4, since we know that the number of citations needed for a paper by the Scientific Staff, to belong to selected percentiles is remarkably constant throughout their publication records. With these bins, we also have complete control of how much data goes in each bin. However, ‘complete control’ is putting it too strongly—we have to accept the fact that citations come in whole numbers—thus, bins containing some of the prevalent citation-numbers ($k < 20$) cannot be expected to make up ‘nice’ fractions of the database. Therefore, in Table 5.1 both the intended and actual percentiles are listed.

| Bin number | Within this percentile (intended) | Citations in each bin | | Actual percentile | |
|------------|-----------------------------------|-----------------------|--------|-------------------|---------------|
| | | FI | AI | FI | AI |
| 1 | 0th–25th | 0 | 0 | 0.0th–25.0th | 0.0th–24.8th |
| 2 | 25th–50th | 1–4 | 1–4 | 25.0th–48.9th | 24.8th–48.0th |
| 3 | 50th–75th | 5–16 | 5–17 | 48.9th–74.3rd | 48.0th–74.6th |
| 4 | 75th–90th | 17–47 | 18–49 | 74.3rd–90.0th | 74.6th–90.0th |
| 5 | 90th–95th | 48–88 | 50–91 | 90.0th–95.0th | 90.0th–95.0th |
| 6 | 95th–99th | 89–271 | 92–279 | 95.0th–99.0th | 95.0th–99.0th |
| 7 | 99th–100th | 272– | 280– | 99.0th–100th | 99.0th–100th |

Table 5.1: The list of the number of citations needed for a paper to lie within each percentile range, both intended percentiles and actual percentiles. The binsizes of both ‘First Initial (FI)’ and ‘All Initials (AI)’ are listed.

With these new bins, we are ready to calculate the final covariance matrix for SPIRES; we simply take the publication record of each author in the Scientific Staff and distribute his papers in the bins defined in Table 5.1. This results in the measurement matrix \mathbf{Y}_{final} , and

⁴The terribly interested reader can, of course, still diagonalize the matrix in Figure 5.1 and use the ‘citation summary’ option from the SPIRES website to find his own citation information, and go about determining his own z -scores in this fashion.

using Equation (5.2), we find

$$\Sigma_f = \begin{pmatrix} 117. & 25.2 & 24.8 & 22.8 & 10.5 & 12.4 & 5.10 \\ 25.2 & 73.0 & 41.9 & 9.57 & -1.17 & -3.03 & -1.23 \\ 24.8 & 41.9 & 76.1 & 43.5 & 13.1 & 8.35 & 1.51 \\ 22.8 & 9.57 & 43.5 & 60.2 & 22.1 & 19.4 & 5.20 \\ 10.5 & -1.17 & 13.1 & 22.1 & 14.1 & 11.8 & 3.69 \\ 12.4 & -3.03 & 8.35 & 19.4 & 11.8 & 17.1 & 5.94 \\ 5.10 & -1.23 & 1.51 & 5.20 & 3.69 & 5.94 & 3.83 \end{pmatrix}; \quad \mu_f = \begin{pmatrix} 12.6 \\ 12.1 \\ 12.8 \\ 7.90 \\ 2.54 \\ 2.02 \\ .506 \end{pmatrix}, \quad (5.9)$$

which is the covariance matrix and the means that we will work with in the following. Note that the sum of the elements in μ_f add up to the average number of publications per author, $\langle n_i \rangle^{(25)} \approx 50.5$, and that these are distributed according to the probabilities given in Table 5.1; this is as it should be. In the interpretation of this matrix, only the numbers for the ‘First Initial’ parsing will be used explicitly as well as in the plot. The reason for this is to keep things as simple as possible; the PCA analysis demands display of quite a few tables of numbers as it is, and the ‘All Initials’-counting does not add anything qualitative (*cf.* Table 5.1) to the picture that is being drawn from the FI data; hence this extra data set is not included in the remainder of this chapter.

5.3.2 Interpreting Σ_f

First of all, let us consider the information available from simply inspecting the covariance matrix. The covariance matrix is the starting point for many multivariate statistical methods and contains a great deal of information about the system in question. In Figure 5.2 (a), the covariance matrix from Equation (5.9) is plotted, and in Figure 5.2 (b), the corresponding correlation matrix, Γ_f . The covariance matrix verifies that the data truly does contain longitudinal correlations, since most off-diagonal elements are non-zero. Apart from this naïve observation, the first thing one notices is that the variances of the first 3 bins dominate the picture. Because all authors publish a majority of papers that end up in these bins, it is natural to expect the variances of these bins to be rather large. In the correlation matrix, each entry is scaled in accordance with the magnitude of variation in each variable, so the diagonal of this matrix $\rho_{jj} = 1$.

When we turn to the covariances between the different bins, we notice that the 0-citations bin has a high, positive correlation with every other bin; the correlation matrix shows that the size of the association is also noticeably constant. This indicates that all authors in SPIRES, no matter how excellent they are, write a remarkable fraction of un-cited papers. The covariance matrix contains almost no negative entries, but the second bin, containing papers with 1 – 4 citations, is actually negatively correlated with bin number 5, 6, and 7. This means that authors producing papers that end up in the second bin, generally do poorly, when it comes to writing papers in the higher percentiles—and the other way around.

Aside from these two exceptions, the general rule is that bins are correlated most with the bins immediately next to them. This simply means that authors that tend to publish many papers in one bin also tend to publish quite a few papers that end up in the neighboring bins, and conversely that if authors’ contributions in a certain bin is below the mean, this *also* tends to be the case in the neighboring bin. This mechanism is, of course, also intuitively appealing. An author who exclusively publishes papers that end up in bin number one and four is not likely.

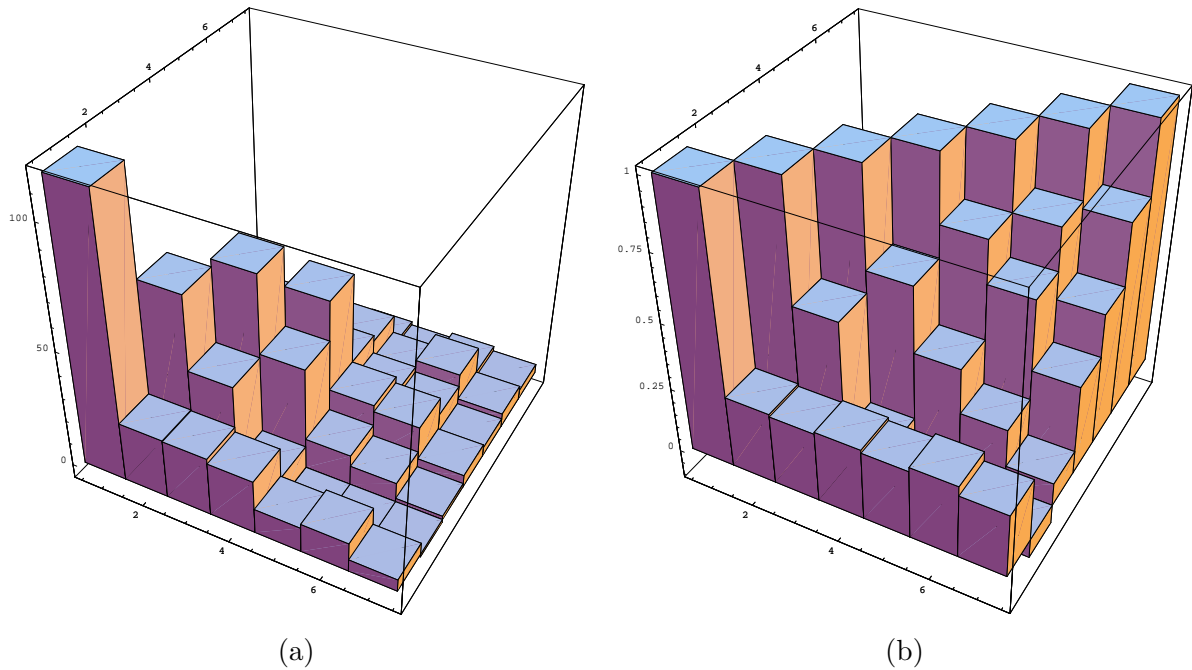


Figure 5.2: The covariance matrix for the bins defined in Table 5.1 and the data for the Scientific Staff.

| u_1 | u_2 | u_3 | u_4 | u_5 | u_6 | u_7 | Eigenvalue | Explained by λ_i | Cumulative |
|--------|---------|---------|--------|--------|---------|---------|--------------------|--------------------------|------------|
| 0.605 | 0.785 | 0.0490 | 0.120 | 0.0370 | 0.00416 | 0.00615 | $\lambda_1 = 175$ | 48.5% | 48.5% |
| 0.396 | -0.273 | 0.703 | -0.518 | 0.0702 | 0.0291 | 0.00243 | $\lambda_2 = 84.5$ | 23.4% | 72.0% |
| 0.526 | -0.485 | 0.00372 | 0.618 | -0.327 | -0.0154 | 0.00357 | $\lambda_3 = 69.4$ | 19.2% | 91.2% |
| 0.402 | -0.267 | -0.554 | -0.232 | 0.606 | -0.193 | -0.0334 | $\lambda_4 = 17.8$ | 4.94% | 96.2% |
| 0.145 | -0.0482 | -0.298 | -0.250 | -0.214 | 0.875 | -0.118 | $\lambda_5 = 9.18$ | 2.54% | 98.7% |
| 0.129 | 0.0135 | -0.312 | -0.432 | -0.613 | -0.339 | 0.456 | $\lambda_6 = 3.21$ | 0.892% | 99.6% |
| 0.0396 | 0.0264 | -0.0982 | -0.180 | -0.313 | -0.286 | -0.881 | $\lambda_7 = 1.41$ | 0.391% | 100% |

Table 5.2: The eigenvectors and eigenvalues of Σ_f . The u_i are plotted in Figure 5.3. These vectors compose an orthonormal set; the first element of each eigenvector is positive by (my) convention.

5.3.3 Diagonalizing the Covariance Matrix

The picture formed from these very general remarks becomes a good deal clearer, when we diagonalize the covariance matrix. We know that the eigenvectors of Σ_f are the uncorrelated axes of rotation that maximize the variance. As such, we can give them a very intuitive meaning. These eigenvectors can be interpreted as the ‘archetypes’ of authors in the database, and are listed in Table 5.2. In Figure 5.3, a graphical representation of the eigenvectors of Σ_f is plotted.

The first eigenvector (or *eigenauthor*, if one were so inclined) u_1 , has only positive coefficients that reflect the shape of the distribution of paper citations—the percentage of papers in each bin; there is about 25% of the data in each of the first 3 bins, and 15% in bin number four. In the last three bins there are about 5%, 4%, and 1% of the data. This vector explains 48.5% of the total variability. The next eigenvector u_2 , explaining some 23.4% of the total variability, is an excited mode—separating the 0-cited bin from the rest of the database. An author with a large positive load on this vector has a large fraction of zero-cited papers,

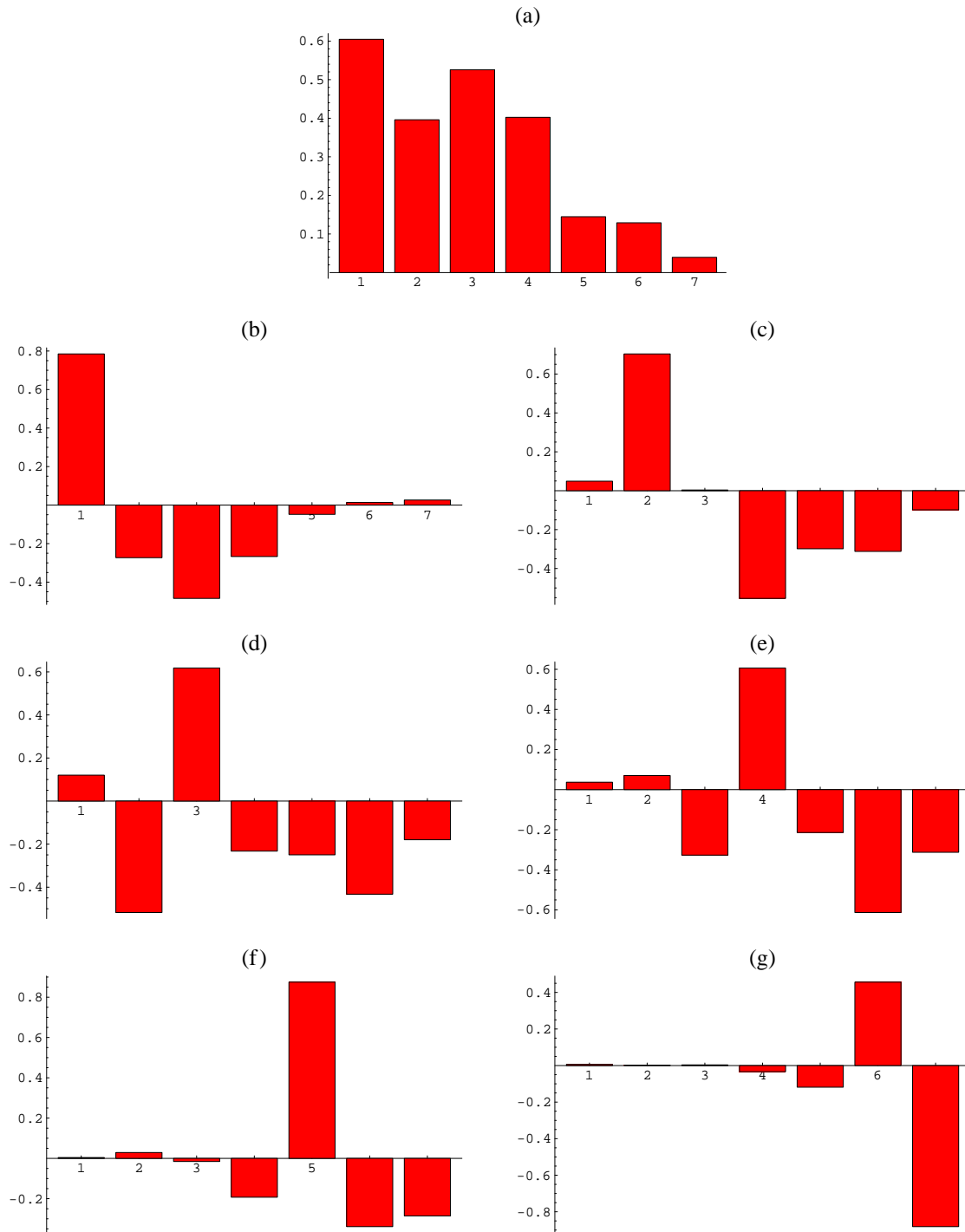


Figure 5.3: The eigenvectors of Σ_f . In (a), u_1 is plotted with the x -axis enumerating the entries 1 – 7 in the vector, and the y -axis displays the numerical value of the u_{1i} -entry. In (b) u_2 is plotted, and so forth for the remaining eigenvectors. These vectors are normalized to one, and by convention, the value of the first entry is chosen to be positive.

while a negative load here would reflect doing rather nicely in the mid-bins. The third PC (19.2%) represents another contrast, dividing the database around the third bin (containing the 5 – 16 citation papers)—low-cited papers (0 – 4 citations) on one side and highly cited papers ($17 - k_{max}$) on the other. The weight of the bins closest to the dividing one is greater. The pattern in the following eigenvectors is more unclear, except for \mathbf{u}_7 ; there is no doubt that a negative load on this vector signals that the author in question is doing very well indeed. That the latter bins generally are a little confusing (would you want a positive or negative load on eigenvector number 4,5,6?) and seem to provide very little information, is to be expected; why this is so will be explained in the following section.

5.3.4 Knowing When to Quit

In Section 5.2.1 we touched on the subject of PCA's ability to reduce the number of variables in a data set from p to $k < p$ dimensions. Considering the extreme case, if one had a 25-variable problem, and the first three PC's accounted for 96% of the variability of the data, it would be tempting to use just those three and ignore the remaining 22 that account for the remaining 4%. However, what is k ? The larger k , the better the fit of the PCA model, and the smaller, the simpler the model will be. Connecting to our own data the question becomes, how many of the $p = 7$ eigenvectors are needed to determine the character of an author in SPIRES? A large number of criteria have been designed to isolate the right value of k , *cf.* [67]. Here, we will keep it simple.

The most primitive rule is called the 90% rule and it simply states that you should keep enough eigenvalues (and corresponding PC's) to explain between approximately 90% of the variability of the data. Turning to Table 5.2, it seems we should keep $k = 3$ eigenvalues, since these 3 eigenvalues explain a little more than 91% of the variability in the data. Another test, is a widely used graphical technique called the 'scree test'. The scree test consists of plotting the size of the eigenvalues against their number.

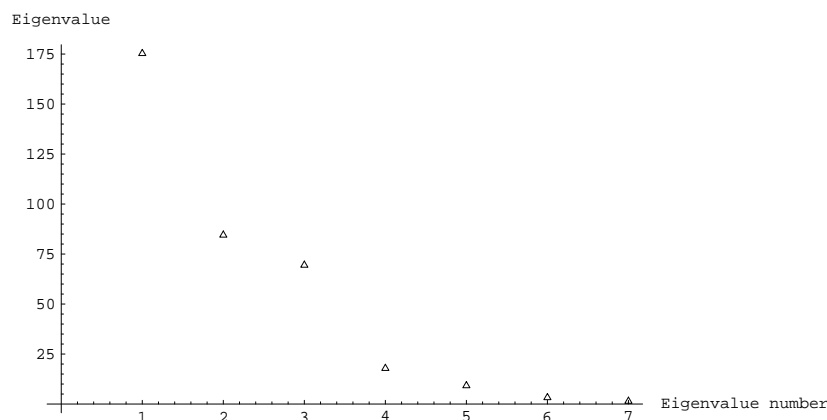


Figure 5.4: A Scree plot; the values of the eigenvalues of the covariance matrix are plotted on the ordinate and the number of the eigenvalue is plotted on the abscissa. Note that the last four roots are much smaller than the rest and lie nearly on a straight line.

Scree is the rubble at the bottom of a cliff, and in this case it refers to the eigenvalues

that may be discarded. In Figure 5.4, we see that the sizes of the eigenvalues drop off rather rapidly at first, and then flatten out. This image looks a little like a cliff. The scree test tells us to keep the cliff and discard the scree, except for the first in the 'flat' part of the plot. In this case the pronounced cliff is the first three eigenvalues, and the scree are the last four ditto. Thus, this test recommends that we use $k = 4$ eigenvalues in our further analysis.

A last criterion says that we should keep the components with sound subject matter interpretations. This very sensible advice will be followed closely in the following. We will focus our attention on the first three or four eigenvalues, but keep an eye out for sound subject matter interpretations. Take for example the \mathbf{u}_7 eigenvector—as was mentioned earlier, there is no doubt that a large negative load on this vector is positive for an author, since this reflects highly cited papers in that author's citation record.

5.3.5 Residual Analysis

To illuminate this line of thought a little further, a few remarks on residual analysis are in order. One of the primary uses of PCA is quality control. In the context quality control, the word 'quality' means keeping within some well-defined norm. Therefore, PCA is helpful when it comes to detecting the so called *outliers*—objects that lie outside the defined norm. In the following I will outline the *idea* behind residual analysis, but not go into any details on how the actual quantities involved are calculated, since these tend to become technical and are basically uninteresting in this context. The interested reader is referred to Appendix B.1 for a review of the concrete *modus operandi*.

Two different types of outliers are usually distinguished between, *viz.* Type A and Type B. Type A are the 'boring outliers' that generally stick out from the distribution one wishes to assume; these outliers are interesting in their own right because they are the 'extreme' authors in SPIRES, but they are boring in the sense that they would be identified as outliers whether or not PCA had been employed; they could have been picked up by other multivariate techniques. These outliers are, however, also picked up by the PCA analysis. The Type B outlier, is a wholly different type of outlier that is not detected by other multivariate techniques; *these outliers are the authors whose observation vectors cannot be appropriately described by the subset of PC's one chooses to use.* The residual analysis is based on the following considerations.

If we retain all the PC's, we can express our original variables in terms of the PC's by inverting Equation (5.6). We find that

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{U}\mathbf{z}, \quad (5.10)$$

but clearly, this is only an exact representation of \mathbf{y} if all the PC's are used. If $k < p$ PC's are used, only an estimate $\hat{\mathbf{y}}$ of \mathbf{y} is produced, namely

$$\hat{\mathbf{y}} = \boldsymbol{\mu} + \tilde{\mathbf{U}}\tilde{\mathbf{z}}, \quad (5.11)$$

where $\tilde{\mathbf{U}}$ is now a $p \times k$ matrix and $\tilde{\mathbf{z}}$ is a $k \times 1$ vector. We can rewrite Equation (5.11) as

$$\mathbf{y} = \boldsymbol{\mu} + \tilde{\mathbf{U}}\tilde{\mathbf{z}} + (\mathbf{y} - \hat{\mathbf{y}}). \quad (5.12)$$

Here, the first term on the rhs. represents the contribution of the multivariate mean, the second term the contribution from the PC's, and the final term represents the amount that is unexplained by the model; *the residual*. This third term gives us an estimate of whether

or not a particular observation vector is adequately described by the subset of PC's that we have chosen to use. Residual analysis is based on analyzing the Q -statistic, given by the sum of squares of the residuals:

$$Q = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}). \quad (5.13)$$

This represents the sum of squares of the distance of $\mathbf{y} - \hat{\mathbf{y}}$, from the k -dimensional space, that the PCA model defines. The Q -statistic is the starting point for the residual analysis, where the idea more or less is testing whether or not the Q -value is larger or smaller than a given measure; this is not the place for the details of the concrete procedure, since these tend to become rather technical, and the interested reader is once again referred to Appendix B.1, where the actual method is described.

The interesting conclusion of this argument, is a rather unusual one. In analyzing excellence in SPIRES, we are uninterested in the norm. We are actually interested in the exact opposite: the truly excellent authors are the outliers, and thus a 'reverse' quality control of the database will help isolate the best authors with ease. Thus in the continued analysis of SPIRES, the lesson to bring along is that even though the first three eigenvectors are enough for most purposes, we cannot ignore the latter ones. Screening authors from a sample of applicants, using residual analysis, can be used to locate interesting authors for employment. Further, screening is *necessary* in order to isolate the authors, for whom the description provided by the first few eigenvectors simply is not sufficient. Now, let us leave this tangent and return to the main investigation.

5.4 An Example

Let us put all of the above together in a concrete example. This example will illustrate the use of PCA, focusing its use as an augmentation of the measure of excellence, r , from Section 2.4, Equation (2.2). Thus, we dig up the authors A, B, C, from Chapter 2, and author \mathcal{U} , whom we found in Section 3.4.1 (this author takes the place of author D). These authors have totals of 201, 178, 252, and 52 publications, respectively. In Table 5.3 the measurements and z -scores for these four authors are listed. The z -scores are found using Equation (5.6).

| Citation counts | | | | z -scores | | | |
|-----------------|----------------|----------------|----------------------------|----------------|----------------|----------------|----------------------------|
| \mathbf{y}_A | \mathbf{y}_B | \mathbf{y}_C | $\mathbf{y}_{\mathcal{U}}$ | \mathbf{z}_A | \mathbf{z}_B | \mathbf{z}_C | $\mathbf{z}_{\mathcal{U}}$ |
| 18 | 19 | 5 | 48 | 60.5 | 51.9 | 27.9 | 7.64 |
| 35 | 51 | 13 | 4 | -38.8 | -31.0 | -16.1 | 38.4 |
| 60 | 49 | 19 | 0 | -18.4 | 7.46 | -55.4 | 1.81 |
| 57 | 37 | 38 | 0 | -1.18 | -8.78 | -56.2 | 3.97 |
| 21 | 9 | 35 | 0 | 7.79 | 2.34 | -56.4 | 2.08 |
| 9 | 8 | 75 | 0 | 4.14 | -2.67 | -21.2 | 0.235 |
| 1 | 5 | 67 | 0 | -0.819 | -2.71 | -30.2 | 0.241 |

Table 5.3: Measurement vectors and z -scores for the test-authors from Chapter 2.

We found that the authors A, B, C, and \mathcal{U} had values of $r = 17.8, 9.8, 188.4$, and 20.6 respectively⁵. These authors were chosen carefully to illuminate some problems with the

⁵These values change to $r = 12.8, 6.1, 151.5, 22.0$, if they are drawn on the distribution of citations of papers by the Scientific Staff and the bins given in Table 5.1, cf. also the probabilities given in Footnote 3 in this

Power of Excellence.

The first problem concerns the difference between authors A and B. Author A has a much higher value of r than author B—author B is more probable by a factor of 10^8 —even though, inspecting their citation counts in Table 5.3, can convince us that author B has written more top-cited papers, has a higher average number of citations per paper, and has more total career citations than author A. This is simply a consequence of the power-law nature of the paper citation distribution, a distribution where 10 papers with 100 citations is vastly less probable than one paper with 1000 citations. Using PCA, this difference between the two authors is illustrated by the different sign on their respective loads, on the third characteristic vector. Both authors have a relatively large load on \mathbf{u}_1 , and similarly a negative load on \mathbf{u}_2 , the eigenvector that separates the 0-cited bin from the middle ones. Of course, it is still a subjective judgment whether one prefers one or the other type of author; some would say the greater accomplishment was writing the one paper with 1000 citations, whereas others would appreciate the author that steadily keeps publishing papers that accumulates high, but not exceptionally high, numbers of citations. However, with the PCA analysis, we can distinguish between these two kinds of authors with ease.

The next problem with the r , defined in Section 2.4, concerns the ‘improbably bad’ authors whose r -values in some cases are higher than for ‘good’ authors. The extreme example is author \mathcal{U} , who is *more* ‘improbable’ than both authors A and B, whom we know are accomplished authors in SPIRES. A mere glance at $\mathbf{z}_{\mathcal{U}}$, however, can convince us that this author cannot hide from the conclusions of the principal component analysis. The most revealing factor is \mathcal{U} ’s large positive load on the zero-citation vector, \mathbf{u}_2 , that instantaneously reveals that although this author is very improbable, he is not someone anyone would want as a part of their research group (... unless he had an *extremely* pleasant personality).

Author C is the most remarkable author in SPIRES. His merits are listed in Chapter 2 and they are truly amazing. Had we only retained the first 3 PC’s, we might not have noticed anything unusual about this author; his z -scores for these first PC’s are not unusual compared to authors A’s and B’s. This is a good example of where the residual analysis from Section 5.3.5 enters the picture. Screening the authors A through \mathcal{U} immediately alerts us that we should pay special attention to author C, and not only rely on the information contained in the first eigenvectors. In Summary, the PCA is an ideal tool for studying authors in SPIRES and combined with the Power of Excellence, we are equipped to handle and competently analyze scientific excellence in the SPIRES database.

5.5 Summary

The method of PCA dates back to the 1930’s, *cf.* the pioneering paper by Hotelling [68]. In the time between then and now a myriad of refinements, conventions, new measures, *etc.*, have sprung to life. These are interesting in and of themselves, and the interested reader is referred to Jackson’s book for a view of the entire subject [67]. In this chapter, the method of PCA has been cut down to the bare essentials. This has been done deliberately to spare the reader from the tedious details of the concrete evaluation, locking the focus onto the ideas

chapter. That these authors are *less* ‘improbable’ compared to the members of the Scientific Staff, is not surprising. The reason that author \mathcal{U} is *more* improbable here, is that it is of course even more improbable receive this few citations in comparison to the scientific staff. Furthermore, the change of the bin sizes does not help either; the second bin has been subdivided, which makes this author even more improbably unsuccessful.

and mechanics of the method.

After going over the preliminary theory and reviewing the concepts behind the method, a couple of problems with the choice of bin sizes was cleared out of the way, so that we were able to establish the final covariance matrix and diagonalize it. The eigenvectors showed us the archetypical authors in SPIRES. The primary point of our interest in PCA, however, is to apply the method as an aid in the investigation of excellence; in this respect there were two uses for the analysis.

The first one was as a ‘reverse’ quality control to help us pinpoint interesting authors; the authors that are outliers from the norm, and whom we want to pay special attention to. The second—and in the context of this thesis as a whole—a very important one, was the use of PCA for classifying excellence. We saw (using four concrete examples) that PCA can be used as an augmentation of the Power of Excellence; illuminating some of the dark points that a one-number representation, such as r , of a complex entity, such as scientific excellence, must necessarily have. Finally, the residual analysis helped us isolate the authors that are too unique for a wholesale explanation.

It is now time to change course. We have taken the statistical analysis of the SPIRES data far, and gained an enormous amount of knowledge about the database. While focusing on the data in the investigation of scientific excellence, it has become apparent that the distinction made by Feynman in the Introduction, does not form a water tight barrier—in the process of generating ‘useful’ results, we have made quite a few new discoveries by applying a host of tools from statistical physics.

At this point, it is time to approach the data from a different angle, namely by modelling the citation network and investigating whether we can create a model of the citation process that can illuminate something about the internal dynamics of the SPIRES database; this is the method often used in the physics of complex system, *cf.* the analysis of the paper by Bak, Tang, and Wiesenfeld [5]. In the Introduction, we have discussed some of the earliest attempts to model the structure of complex networks, the random graph and the Watts-Strogatz model (Section 1.1.2). In this chapter, we shall proceed to study in some detail the *Growing Network* model, first proposed by Barabási and Albert. After an introduction to this model, analyzing its properties from different standpoints, a modified version designed specifically for the SPIRES database is presented. But let us not get ahead of ourselves, so without further ado...

6.1 Introduction to the Growing Network Model

At this point, the mindful reader should be well aware that many complex networks and, in particular, the distribution of citations in the SPIRES database, have power-law degree distributions. Both the random graph and the WS model have Poissonian degree distributions and even though the WS model seems to mimic the structure of human social relations fairly well the model suffers one major drawback; we need a model-network that mimics the scale free power-law behavior seen in many real-world networks.

To obtain this result, we have to change our focus a little bit, and this is exactly what Barabási and Albert did [4]. Instead of focusing on recreating network topology these authors focus on modelling network dynamics, anticipating that if the correct dynamics were

pinpointed, the topology would follow. In doing this, they (independently) rediscovered a special case of model proposed by Simon [69, 70].

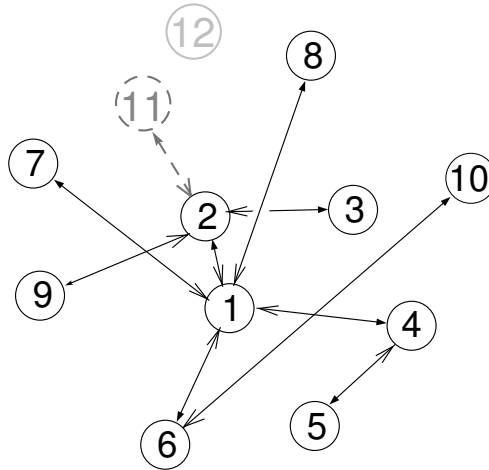


Figure 6.1: An illustration of the evolution of the simplest case of the Growing Network. Nodes are added one at a time and each newly added node links to a node already present in the system. In this representation each node is labelled by the time s it was added to the network. Node 11 is being added and 12 is waiting to come into existence. This is signified by their color. The degrees are the number of arrows impinging on each node. Here, the notion of a directed network has also been introduced. The arrows point both ways, but solid arrowheads signify outgoing links, whereas the drawn arrows are the incoming edges. In this simple version every node has only one outgoing edge (except the very first node), node 1 has in-degree 5, node 2 has in-degree 3, *etc.* Note also that 1 is the ancestor of 6, and 9 is the descendant of 2.

The Growing networks (GN) model that Barabási and Albert proposed is based on two fundamental mechanisms, growth and preferential attachment.

- *Growth.* Setting up the model consists of starting out with an initial number of nodes m_0 . At each time step a new node with $m \leq m_0$ edges are added that link to nodes already present in the system.
- *Preferential attachment.* The probability $\Pi[k(s, t)]$ that, at time step t , a new node attaches to a node at site s already present in the system is proportional to the number of edges of that node, $k(s, t)$, at time t ; in other words $\Pi[k(s, t)] = k(s, t) / \int_0^t du k(u, t)$. The nodes are named according to the moment they were added to the system, thus the first node is $k(1, t)$ the second $k(2, t)$ ¹, *etc.*

Clearly, after t time steps this model has $t + m_0$ vertices, and mt edges. So far the model is *undirected*, that is, an edge connects two nodes and contributes to the degree of both nodes, thus there are $2mt$ node endpoints. The GN model is easy to implement and numerical runs shows that it evolves into a scale free state where the distribution of node degrees follows a power-law of slope $\gamma_{\text{model}} = 3$, *cf.* Figure 6.2. All in all, this looks like a good starting point for creating a model for the complex network of scientific publications.

¹I am using a notation that is slightly different from Barabási and Albert's.

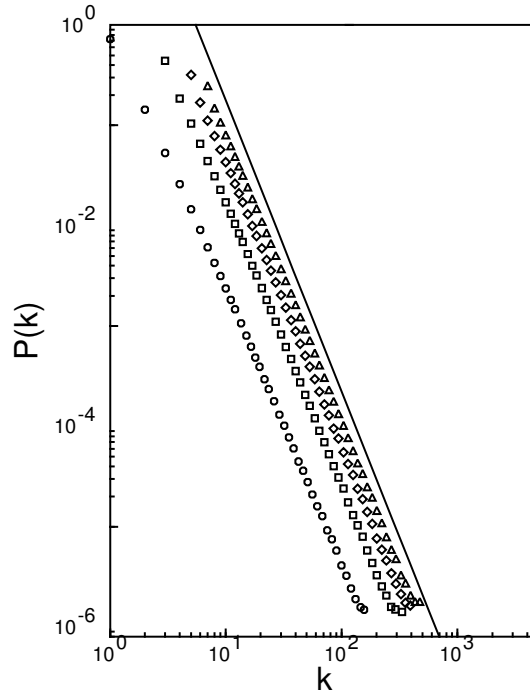


Figure 6.2: Degree distribution of the scale-free model, with $N = m_0 + t = 300,000$ and $m_0 = m = 1$ (circles), $m_0 = m = 3$ (squares), $m_0 = m = 5$ (diamonds) and $m_0 = m = 7$ (triangles). The slope of the dashed line is $= 2.9$.

Several analytical approaches can be utilized to find an analytical solution for this model. As each of these have strengths and weaknesses, I will briefly outline them in the following.

6.1.1 Continuum Solution

Perhaps the most intuitive analytical solution for the GN model was proposed by Barabási and Albert [4, 35], and was first coined the *mean field approach* by its inventors. I will briefly run through it here. The continuum approach focuses on the time dependence of the degree $k(s, t)$ of a given node at site s in our network. This approach is only ‘mean field’ in the sense that k is treated as a continuous real variable, which means that $\Pi(k(s))$ can be interpreted as a continuous rate of change. This approximation also means that the continuum solution is only valid in the $k \gg 1$ regime².

The degree $k(s, t)$ changes when a new node is added and links to a node at the s th site. The probability of this process is clearly $\Pi(k(s))$, thus

$$\frac{\partial k(s, t)}{\partial t} = m\Pi(k) = m \frac{k(s, t)}{\int_0^t du k(u, t)}. \quad (6.1)$$

²To keep this approach very intuitive, I am following Barabási and Albert rather closely. In Section 6.1.3, the master equation approach is employed to illuminate this line of attack further.

The integral $\int_0^t du k(u, t)$ is simply the number of edge endpoints and it runs over all nodes in the system, thus it is equal to $2mt$. The interesting limit is $t \rightarrow \infty$ where the initial conditions are unimportant, so we can simply write

$$\frac{\partial k(s, t)}{\partial t} = \frac{k(s, t)}{2t}. \quad (6.2)$$

Equation (6.2) is a first order linear homogeneous DE that can be trivially solved using separation of variables, using the initial condition that $k(t, t) = m$ (each node has m edges at its introduction) to find

$$k(s, t) = m \left(\frac{t}{s} \right)^{\frac{1}{2}}, \quad (6.3)$$

which allows us to see that the $k(s, t)$'s all evolve in the same way following a power-law with slope $1/2$. This means that the vertices that have the most connections are those that have been added at the early stages at the network.

We want to find an explicit expression for the probability distribution, $P(k)$. One way to reach this goal is to begin with the probability $P(k(s, t) < k)$ that $k(s, t)$ is smaller than a given k at time t . Inserting Equation (6.3) in this probability, and solving for $s = s(k, t)$, we see that this is $P(s > tm^2/k^2)$. Since the new nodes are added at equal time intervals, the probability density of s is flat, $P(s) = 1/(m_0 + t)$. Using these results we find

$$P(k(s, t) > k) = 1 - P(s \leq tm^2/k^2) = 1 - \frac{1}{t + m_0} \frac{tm^2}{k^2} \quad (6.4)$$

Equation (6.4) is precisely the cumulated distribution, so the well known probability distribution $P(k)$ can be found by differentiating,

$$P(k) = \frac{\partial P(k(s, t) < k)}{\partial k} = \frac{2m^2t}{t + m_0} k^{-3}, \quad (6.5)$$

and in the $t \rightarrow \infty$ (large network) limit that we are interested in, we find the very simple result

$$P(k) \sim 2m^2 k^{-3}, \quad (6.6)$$

where $\gamma = 3$ as predicted by the numerical runs of the model. Note that in this simple version of the GN model, the slope of the power-law, γ , is independent of m , as was also indicated by our initial runs, m only causes a displacement of the y intercept. This solution is useful for 'quick n' dirty' results because of its intuitive appeal and ease of calculations. The continuum solution suffers from the drawback that much of the interesting structure in the citation distribution is found for small k (the sea of dead papers), where this solution is not necessarily valid. This problem is remedied in the next analytical approach.

6.1.2 The Rate Equation Approach

This approach was proposed by Krapivsky, Redner, and Leyvraz [42] and uses the concept of a *rate equation*. The utility of the rate equation in non-equilibrium statistical physics has been demonstrated for a diverse range of phenomena, such as aggregation [71], coarsening [72] and epitaxial surface growth [73].

Here, the focus is on the *average* number of nodes, $N_k(t)$ with k edges at time t . In the simplest case ($m = m_0 = 1$)³, the change of $N_k(t)$ when a new node is added, can be expressed as

$$\frac{dN_k}{dt} = \frac{(k-1)N_{k-1} - kN_k}{M_1} + \delta_{1k}, \quad M_1 = \sum_k kN_k. \quad (6.7)$$

The first term on the right hand side accounts for processes where a node with $(k-1)$ edges is connected to the new node, increasing N_k by one. By definition there are N_{k-1} nodes with $(k-1)$ edges, and according to the model the rate at which these processes occur is proportional to $(k-1)N_{k-1}$. The M_1 ensures the correct normalization. The second term on the right plays a similar role, accounting for the loss that occurs when a node with k edges is connected to the new node, causing a loss in N_k with probability kN_k/M_1 . Third term on the right is a kronecker δ -function that accounts for the addition of new nodes.

The terminology M_1 is used because the sum $\sum_k kN_k$ is (obviously) identical to the first moment of the $N_k(t)$ distribution, which can be found using the general identity

$$\dot{M}_n = \frac{d}{dt} \sum_k k^n N_k = \sum_k k^n \frac{dN_k}{dt}, \quad (6.8)$$

where the last derivative is given directly by equation (6.7). This sum can easily be written out and be evaluated explicitly, yielding $\dot{M}_1 = 2$ that can be integrated to find $M_1(t) = M_1(0) + 2t$. We should have anticipated this result, since the ‘physical’ interpretation of the first moment is exactly that it corresponds to the number of link endpoints; it is the same sum we encountered in equation (6.1). Again, we are interested in the asymptotic ($t \rightarrow \infty$) regime, where the initial conditions are irrelevant. Hence, we can insert $M_1 = 2t$ into Equations (6.7) and solve the first order non-homogeneous differential equation that arises. This can be mechanically solved to yield $N_1 = 2t/3$. Inserting this result N_2 equation, produces another DE that can be solved to find $N_2 = t/6$, the structure of which reveals that all N_k ’s depend linearly on time. Thus we can substitute $N_k(t) = P(k)t$ into (6.7), and with a minimum of algebraic manipulations, we find

$$P(k) = P(k-1) \frac{k-1}{k+2}. \quad (6.9)$$

This equation can be solved for $P(k)$ by realizing that

$$P(k) = P(1) \prod_{i=2}^k \frac{i-1}{i+2} = n_1 \frac{3!}{k(k+1)(k+2)} = \frac{4}{k(k+1)(k+2)}, \quad (6.10)$$

where most of the terms in the product cancel out, resulting in a very simple expression for the degree distribution. This result is valid for all k in the large t limit, and should be compared to Equation (6.6), where we have an agreement. It is straight forward to show that this power is also independent of the number of links added.

6.1.3 Master Equation

Yet another approach is the master equation approach suggested by Dorogovtsev, Mendes, and Samukhin [43]. Here, the probability $p(k, s, t)$ that at time t , a node introduced at time

³Which is what is solved in [42].

s has degree k . The master equation for the simple version of the GN model becomes.

$$p(k, s, t + 1) = \frac{k-1}{2t} p(k-1, s, t) + \left(1 - \frac{k}{2t}\right) p(k, s, t). \quad (6.11)$$

This should remind the reader of equation (6.7) and indeed, solving the master equation results in the recursion relation (6.9). The master equation and the rate equation approaches are equivalent in this respect [13], and for calculating the scaling behavior of the degree distributions, they can be used interchangeably.

Obviously, the master equation contains more information than the corresponding rate equation. The master equation resolves nodes, not only—as it is the case for the rate equation approach—by their degrees k , but also by the time s they were added to the system. In other words, the master equation also gives us information on the *age distribution*. This trait, however, makes many exact calculations using the master equation more involved and less transparent than the corresponding calculation using the rate equation, and the reader will be spared the complicated solution of the GN model here (the solution can be found in [43]).

The master equation has other virtues, as it turns out; it elegantly illuminates many aspects of the continuum approach [74], and in this section, we will use some of the results found here to understand exactly the mechanics of the continuous approach. We will also discuss an important (and general) relation between the exponents of the degree distribution (γ) and the average degree (β), *cf.* equation (6.22) [14, 74].

In equation (6.11), we have the same conventions for t and s as in the continuous approach, that is, $t = 1, 2, 3, \dots$ and $s = 0, 1, 2, \dots, t$. Thus in the simplest imaginable version, at time $t = 1$, we have a pair of connected nodes at $s = 0$ and 1 , so that our initial condition is $p(k, s = 0, t = 1) = \delta_{k1}$ and the boundary condition is accordingly $p(k, s = t, t = 1) = \delta_{k1}$. Now, we can rewrite equation (6.11) to yield

$$2t[p(k, s, t + 1) - p(k, s, t)] = (k-1)p(k-1, s, t) - kp(k, s, t). \quad (6.12)$$

Going to the continuous limit in k and t , and transforming the difference equation (6.12) into a differential equation gives us

$$2t \frac{\partial p(k, s, t)}{\partial t} = - \frac{\partial [kp(k, s, t)]}{\partial k}. \quad (6.13)$$

The next step is to introduce the average degree of an individual node, $\bar{k}(s, t)$ as

$$\bar{k}(s, t) = \sum_{k=1}^{\infty} kp(k, s, t) = \int_0^{\infty} dk kp(k, s, t), \quad (6.14)$$

which is the logical definition, since $p(k, s, t)$ is the probability that the individual node has k edges. Now, if we apply $\int_0^{\infty} dk k$ to equation (6.13), it is easy to find

$$\frac{\partial \bar{k}(s, t)}{\partial t} = \frac{\bar{k}(s, t)}{2t} = \frac{\bar{k}(s, t)}{\int_0^t du \bar{k}(u, t)} \quad (6.15)$$

where the lhs. is trivial and the rhs. is found simply by integrating by parts. This equation should remind the reader of equation (6.2), since they are identical. The only difference is that we now possess a new understanding of the content of considering $k(s, t)$ a ‘continuous’

variable, as it was loosely stated above; the definition is given in equation (6.14). The bar over the $k(s, t)$ reminds us of our newfound knowledge. The reader should note that the recognition that equation (6.15) stems from the continuum limit of the master equation is exactly why solving this equation is now called the continuum solution and not the ‘mean field’ solution as it was first labelled by Barabási and Albert. As we know, equation (6.15) can be solved to yield the result that $P(k) \sim k^{-\gamma}$, $\gamma = 3$.

Having reached this result, let us back up a little, and rewrite equation (6.13) in a more ‘agreeable’ form

$$\frac{\partial[kp(k, s, t)]}{\partial \ln \sqrt{t}} + \frac{\partial[kp(k, s, t)]}{\partial \ln k} = 0. \quad (6.16)$$

This differential equation is solved by any function $h(k, s, t)$ of the form $h(\ln k - \ln \sqrt{t} + K(s))$, where $K(s)$ is some function of s . Because of the discreteness of the model *etc*, we choose the solution $kp(k, s, t) = \delta(\ln k - \ln \sqrt{t/s} + C)$, where C is a constant. Using well-known properties of the δ -function and combining this solution with the boundary condition, $p(k, t, t) = \delta_{k,1}$, yields

$$p(k, s, t) = \delta\left(k - \sqrt{\frac{t}{s}}\right) \quad (6.17)$$

that the transition to the continuous limit leads to a δ -function form of the degree distributions of the individual nodes⁴. More generally, passing to the continuous limits of k and t , only demanding that $\bar{k}(s, t)$ is a solution to an equation ‘similar’ to (6.15), the general result is [14, 74]

$$p(k, s, t) = \delta(k - \bar{k}(s, t)). \quad (6.18)$$

From now on, we can think of this result as *the* continuum approximation, since equation (6.18) allows us to derive any result for any model in the continuum solution context. For instance, knowledge of $p(k, s, t)$ allows one to find the total degree distribution for the entire network $P(k, t) = 1/(t+1) \sum_{s=0}^t P(k, s, t)$, where the stationary distribution is found by once again taking the continuous limit

$$P(k) = P(k, t \rightarrow \infty) = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t ds p(k, s, t), \quad (6.19)$$

and, letting $t \rightarrow \infty$, assuming that a stationary distribution exists (we know that it does). This again provides us with the familiar result for $P(k)$.

At this point, we are ready to formulate a more general result for the continuous approach, namely that a knowledge of the average degree of nodes, $\bar{k}(s, t)$ allows us to find the total degree distribution $P(k, t)$. The general relation is motivated by the following considerations. Let us think of $\bar{k}(s, t)$ as a solution to an equation ‘similar’ to equation (6.15), we have that

$$P(k, t) = \frac{1}{t} \int_0^t ds \delta(k - \bar{k}(s, t)) \quad (6.20)$$

$$= \left[\frac{\partial \bar{k}(s, t)}{\partial s} \right]^{-1} \bigg|_{s=s(k, t)} \quad (6.21)$$

⁴Note that this result has a quite different form compared to the result obtained using the master equation *without* going to the continuous limit. Using the exact master equation, we would have found that $p(k, s, t) = \sqrt{s/t} \exp(-k\sqrt{s/t})$, in the scaling limit (k, s, t large, and $s \ll t$, and $k\sqrt{s/t}$ fixed). This exponential cutoff is interesting in itself, as it reveals information about the finite size effects of these models that will be discussed in some detail later. Here, it is sufficient to note that the δ -function form works effectively for both scale free and exponentially decaying networks *cf.* [14, 74]

where the integral is performed via a change of variables that yields the second equality, and where $s = s(k, t)$ is a solution of the equation $k = \bar{k}(s, t)$. Assuming that $P(k) \sim k^{-\gamma}$ and $\bar{k}(s, t) \sim s^{-\beta}$, we find that $s \sim k^{-(1/\beta)}$. Using equation (6.21), we find that $k^{-\gamma} = \partial k / \partial s \sim k^{-(1+1/\beta)}$. Thus we see that the relationship between β and γ can be described by

$$1 = \beta(\gamma - 1). \quad (6.22)$$

In general, for linearly growing networks, we know that if we assume that $P(k)$ is stationary and follows a power-law, $P(k) \sim k^{-\gamma}$, and that $\bar{k}(s, t) \sim s^{-\beta}$ *i.e.* that these functions exhibit scaling behavior (in the scaling regime, that is), then equation 6.22 is valid: It follows from equation (6.19) that for the distribution to be stationary, we must have $p(k, s, t) = \rho(k, s/t)$. Further, we have that the single node probabilities must be normalized, such that $\int_0^\infty dk p(k, s, t) = 1$. Inserting $\rho(k, s/t)$ into this equation, we get $\int_0^\infty dk \rho(k, x) = 1$. For this definite integral over k to be independent of x , we find that $\rho(k, x)$ must be of the form $\rho(k, x) = f(x)g(kf(x))$, where $f(x)$ and $g(x)$ are arbitrary functions.

Next, we employ the scaling assumption for $\bar{k}(s, t)$. From this and the definition of the average degree in equation (6.14), we get $\int_0^\infty dk k \rho(k, x) \sim x^{-\beta}$; and substituting our newly found expression for $\rho(k, x)$ into this relation, we learn $f(x) \sim x^\beta$, using a variation of the argument above (for the definite integral over k to scale as $x^{-\beta}$, this constraint is impinged on $f(x)$). Naturally, we can set $f(x) = x^\beta$, absorbing constants in the scaling function, $g(x)$. This yields the scaling form of degree distributions of individual vertices

$$p(k, s, t) = \left(\frac{s}{t}\right)^\beta g\left[k \left(\frac{s}{t}\right)^\beta\right]. \quad (6.23)$$

The scaling function depends on the model that is under consideration; in footnote 4 the exact $p(k, s, t)$ is stated for the simplest formulation of the GN model.

Taking the final step, we now use the assumed scaling of $P(k)$ that $\int_0^\infty dx \rho(k, x) = k^{-\gamma}$, using the result in equation (6.23), and the rapid convergence⁵ of $\rho(k, x)$ for large x to yield the relation that $\gamma = 1 + 1/\beta$, which is exactly the content of equation (6.22). Deriving this last result no approximations have been made, and within the assumptions, the relation (6.22) is universal.

6.2 Limiting Cases

Since we are going to work with the GN model, it is instructive to take the model apart to study what happens when we remove either element of the model, *i.e.* the preferential attachment or the growth characteristics. What are the minimum requirements for the model to display scale free characteristics. I will work in the continuum framework due to the transparency of the calculations—it is not difficult to find the following results using the rate or master equation.

6.2.1 No Preferential Attachment

First, let us discuss what happens when there is growth but no preferential attachment. A new node connects to the nodes already present in the system with equal probability, that

⁵The general form of $g(x)$ for a generalized version of the GN model, derived using the master equation, can be found in [43]. For an example, *cf.* footnote 4.

is, $\Pi(\bar{k}(s)) = 1/(m_0 + t - 1)$, (with all variables defined as in Section 6.1.1) thus $\Pi(\bar{k}(s))$ is independent of $\bar{k}(s)$. We can insert this result in equation (6.1) to find the differential equation

$$\frac{\partial \bar{k}(s)}{\partial t} = m\Pi(\bar{k}(s)) = \frac{1}{m_0 + t - 1}. \quad (6.24)$$

Solving for $\bar{k}(s)$ with the initial condition that every node has m edges when introduced, $\bar{k}(s = t, t) = m$, we find that $\bar{k}(s, t)$ follows a logarithmic time dependence

$$\bar{k}(s, t) = m(\ln \left(\frac{m_0 + t - 1}{m_0 + s - 1} \right) + 1). \quad (6.25)$$

Using the familiar technique, we find that in the $t \rightarrow \infty$ limit

$$P(k) = \frac{e}{m} \exp \left(-\frac{k}{m} \right), \quad (6.26)$$

and hereby concluding that removing the preferential attachment from the model eliminates the scale free feature of the GN model. That this is indeed the case is also clear from numerical runs.

6.2.2 No Growth

Now, what happens if we eliminate the growth feature and keep the preferential attachment. Our diminished model begins with N nodes and no edges. At each time step we randomly select a vertex and connect it with probability $\Pi(\bar{k}(s)) = \bar{k}(s) / \int_0^t du \bar{k}(u)$ to vertex s in the system. We use the same method above. This time, the rate of change of the time evolution of the node degrees has two terms. Firstly, we have the random selection of a link $\Pi_{rand}(\bar{k}(s)) = 1/N$ (N is the system size), and secondly the probability that the randomly chosen node connects to a given node $\Pi(\bar{k}(s)) = \bar{k}(s) / \int_0^t du \bar{k}(u)$. Since every edge links to two nodes, we have that $\int_0^t du \bar{k}(u) = 2t$. Thus,

$$\frac{\partial \bar{k}(s)}{\partial t} = \Pi_{rand}(\bar{k}(s)) + c\Pi(\bar{k}(s)) = \frac{1}{N} + \frac{N}{N-1} \frac{k_i}{2t}, \quad (6.27)$$

where $c = N/(N-1)$ originates from the fact that we exclude from the summation the possibility that edges can originate and terminate in the same node. Equation (6.27) is a differential equation that is solved using standard methods to have the form:

$$\bar{k}(s, t) = \frac{2(N-1)}{N(N-1)}t + Ct^{N/2(N+1)}. \quad (6.28)$$

Since for every interesting model we have that $N \gg 1$, which implies that we can approximate $\bar{k}(s)$ with

$$\bar{k}(s, t) \approx \frac{2}{N}t + Ct^{1/2}. \quad (6.29)$$

In the previous cases, we have determined the constant C from the initial condition $\bar{k}(s = t, t) = m$ —that every node at its introduction has m edges. Clearly, no new nodes are introduced in this amputated model. In this model, we can think of the time t_{sel} when node s is selected for the first time, changing its degree from 0 to 1. Equation (6.28) is only valid

for $t > t_{sel}$ and all nodes will not start following this dynamic until after $t \gtrsim N$. We find C from the condition that $\int_0^t du \bar{k}(u) = 2t$, which implies that $C = 0$.

The result of the above investigation is, that after a time of $t \approx N$, the degree of the individual nodes increases linearly with time. Thus, after a transient period where we expect dynamics ‘similar to the model that includes growth’—in the sense that picking a ‘virgin node’ corresponds to adding a node to the system—we expect to see a degree distribution that is Gaussian around a mean value. This conclusion is amply supported by numerical runs. In

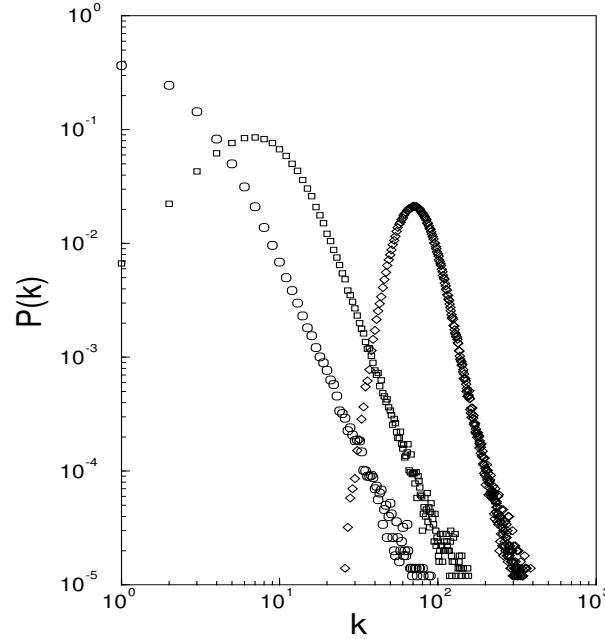


Figure 6.3: The degree distribution of the ‘no growth’ model for $N = 10,000$ and $t = N$ (circles), $t = 5N$ (squares), and $t = 40N$ (diamonds). After [35].

figure 6.3 (a), we explicitly see how the distribution looks like a power-law for $t \lesssim N$ and continuously changes into a Gaussian distribution for $t \rightarrow \infty$.

Thus it seems clear, that both the growth element and the preferential attachment is needed to recreate the scale free behavior that is seen in the citation distribution.

6.2.3 Finite Size Effects

The last aspect of the model that we shall study here, originates from the fact that a computer model (and real networks) always has a finite size. We are already aware that for low k -values, we have not yet entered the scaling regime and we must use the rate- or master equation approach to get reliable results; this region is very susceptible to changes in the initial conditions.

As to the other end of the distribution, the degree of the most popular node or, equivalently, the ‘cut-off’ of the power-law can be determined from the extreme statistics criterion $\sum_{k > k_{max}} N_k = 1$, that is, one node in the network lies in the range (k_{max}, ∞) . In the contin-

uous approximation this criterion becomes

$$t \int_{k_{max}}^{\infty} dk P(k) = 1, \quad (6.30)$$

and we find that for $P(k) = k^{-\gamma}$ this yields the cut-off

$$k_{max} \sim t^{\frac{1}{\gamma-1}}, \quad (6.31)$$

cf. also [75].

6.3 A Simple Model for SPIRES

After this compilation of the most important of the GN, we are going to change this model into a first order approximation of the citation network in SPIRES. Obviously, this model only operates on the paper-level; it does not take the level of authors into account.

The first thing we have to realize is that we need to make our model *directed*. Scientific papers have outgoing edges (references) and incoming edges (citations). The information we have about the SPIRES database is the number of inbound edges for each paper, thus this is the distribution we would like to model. Because scientific papers are published on paper, the out-bound degree distribution is frozen once and for all, when the paper is published (reference lists of published papers do not change). This aspect facilitates modelling, since we do not have to take into account internal rewiring of links, as it is the case for the internet and many other shape-shifting networks, *cf.* Chapter 3 for a further discussion.

The simple model is defined as follows:

- *Growth.* We start out with m_0 nodes. At each time step a new paper is introduced. Each new node comes with one citation to give it a non-zero probability of being cited in the following time steps. Each new node has a reference list of m papers that are already present in the model.
- *Preferential attachment.* The probability that the s th paper is cited (*i.e.* that the paper is present in the reference list of a new paper) depends on the in-degree of node s , $k(s, t)$. In other words: $\Pi(k(s)) = k(s)^\eta / \sum_u k(u)^\eta$,⁶ where room has been allowed for the preferential attachment to be non-linear, since there is no *a priori* reason for the preferential attachment to depend linearly on $k(s)$.

As we shall see in the following, this model recreates the data with surprising accuracy. In spite of the overwhelming activity in the field of complex systems, this model has actually not been solved in the literature. This is due to an amusing misunderstanding—in the first papers on the subject [4, 35], Barabási showed that *if the number of in- and out-bound edges is equal*, the slope of the distribution of edges generated by the GN model, is independent of the number of edges, m . This result is correct, *cf.* Equation (6.6), but as a result of this almost⁷ everyone—including Barabási’s group—have assumed that the slope of the degree distribution is independent of m and used the convention $m = 1$ everywhere. However, this result is true if, and only if, the number of in-bound edges equals the number of out-bound edges, *cf.* Equation (6.34). Therefore, the model shown here is original (although not all that creative).

⁶In the continuous approximation this sum is substituted by an integral.

⁷Loading papers with a number of ‘ghost’ citations is considered in [43], but in this paper, the non-linear preferential attachment (the η -coefficient) is not included.

6.3.1 Numerical Results

In Figure 6.4 the data for the theory subfield of the SPIRES database is compared to the in-degree distribution resulting from a run of the model described in the previous section. In the model, the size of the reference list of each new paper has been set to $m=14$ (this corresponds to the mean value of citations in SPIRES) and the preferential attachment has been raised to the power $\gamma = 4/5$. The agreement is excellent, and in the following, we shall

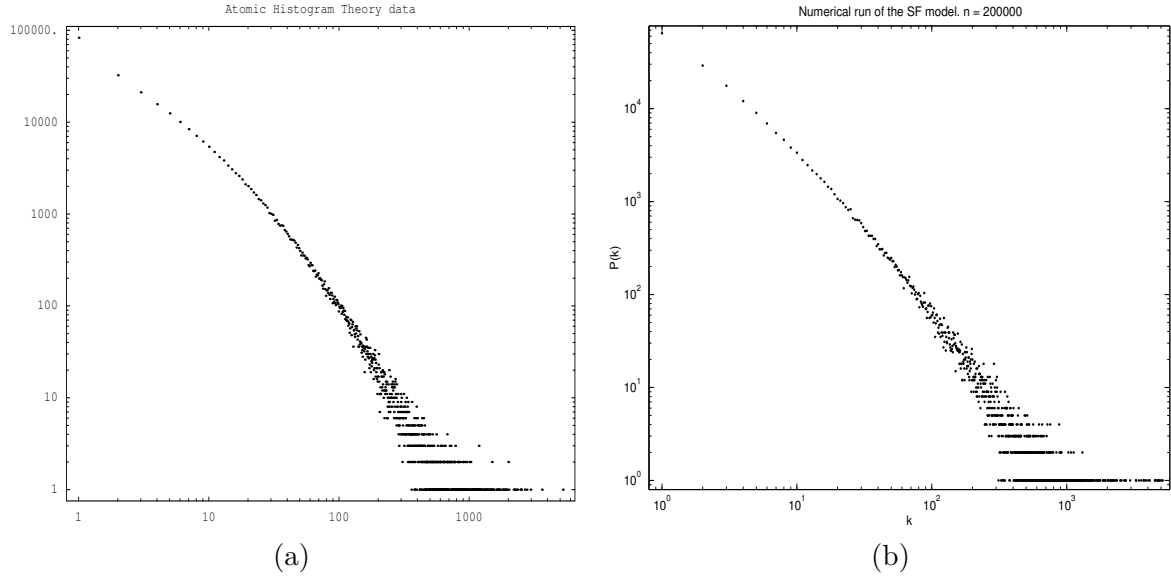


Figure 6.4: (a) Displays an ‘atomic’ histogram of the the citation distribution for the theory data (159,946 papers). (b) Is the GN model with a sub-linear preferential attachment, $\gamma = 4/5$. This version has 200,000 papers and the average number of citations per paper is 14.

see how this model can be solved analytically to find out what mechanisms are responsible for this amazing agreement between model and data.

6.3.2 Analytical Results

Warming Up

First, let’s use the continuum approach to form a rough picture of what happens when we change the way the links are added, keeping $\gamma = 1$ for now. The integral over all nodes (number of node endpoints) is $\int_0^t du \bar{k}(u) = (m+1)t$. This results in the differential equation

$$\frac{\partial \bar{k}(s, t)}{\partial t} = \frac{m \bar{k}(s, t)}{(m+1)t} \quad (6.32)$$

that can be solved with an initial condition analogous to Barabási and Albert’s that $k(t, t) = 1$, that is each node has one incoming link at its introduction. We find that

$$\bar{k}(s, t) = \left(\frac{t}{s} \right)^{\frac{m}{m+1}}, \quad (6.33)$$

where we see that the growth rate of the individual nodes do depend on the number of references in each papers reference list, and that asymptotically for $m \rightarrow \infty$ we have a linear

growth for the degree of the individual nodes. Using equation (6.21), we find an explicit expression for the inbound degree distribution. We find that

$$P_{directed}(k) \sim k^{-\gamma}, \quad \gamma = \frac{2m+1}{m} \quad (6.34)$$

in the limit $t \rightarrow \infty$. Unlike what is the case for the GN model, the slope of the power-law in our citation model *does* depend on the number of references in each paper's reference list (m). For $m = 1$ the citation model coincides with the GN model, and has slope $\gamma = 3$, but for increasing m the citation model has a slope that converges to $\gamma \rightarrow 2$ for $m \rightarrow \infty$; the slope of the power-law is thus bounded by $\gamma \in [2; 3]$. The SPIRES database has an average of 14 citations per paper which results in a slope of $\gamma_{SPIRES} = 29/14 \approx 2.07$, which is very close to the power that is empirically found in the high citation regime.

To investigate what happens in the small k regime, we switch to the rate equation approach, for which it is also possible to take into account the non-linearity in the preferential attachment. The rate equation for the citation model becomes

$$\frac{dN_k}{dt} = m \frac{(k-1)^\eta N_{k-1} - k^\eta N_k}{M_\eta} + \delta_{1k}. \quad M_\eta = \sum_j j^\eta N_j. \quad (6.35)$$

This equation is completely analogous to equation (6.7), and each term on the lhs. have the same function as in the simple case—here the possibility of adding m edges at each time step and the non-linear attachment kernel have been included.

Another important difference is the directed link addition, which plays a big role in the normalization (M_η). Since it will prove useful in the following, we start out by finding the first two moments (M_0 and M_1) of the distribution. The moments coincide with the M_η 's for integer η .

Using equation (6.8), we can do the sums explicitly to find that for \dot{M}_0 , the terms in the rightmost sum of equation (6.8) cancel out, except for the last (loss) term in the sum that is zero per definition (there are no nodes with $N + 1$ links), and the kronecker delta, resulting in $\dot{M}_0 = 1$. This can be integrated to yield $M_0 = M_0(0) + t$, or simply the total number of nodes in the system—this was to be expected; the definition of the 0th moment is just that.

The first moment is found in a similar fashion: We can either explicitly do the last sum in equation (6.8) to find that most terms cancel and leave us with $\dot{M}_1 = (1 + m)$, which can be integrated to find $M_1 = M_1(0) + (1 + m)t$, or we can simply realize that the first moment is the number of edge endpoints in the distribution, as we have already discussed when solving equation (6.32). The number of edge endpoints must clearly grow as $(1 + m)t$, since m out-going links are added at each time step and each paper comes with one citation. Doing the sums explicitly, however, gives us one important piece of information. The higher moments do depend on η , as does the total degree distribution. For the two first moments, we have the simple result that they are independent of η and they grow linearly with time.

Motivated by Figure 6.4, we embark on solving the sub-linear case, *i.e.* equation (6.35) with $0 \leq \eta \leq 1$. Again, it is instructive to solve the two limiting cases, $\eta = 1$ and $\eta = 0$. In the case $\eta = 1$ the solution corresponds to the solution of equation (6.7) with the two differences that $M_1 = (1 + m)t$ (directed link addition) and that each paper can have m references. Thus, we can solve equations (6.35) for the first N_k 's. In most interesting limit, $t \rightarrow \infty$, we find that $N_1 = (1 + m)t/(1 + 2m)$ and $N_2 = m(m + 1)t/(1 + m(5 + 6m))$, *etc.* The structure of this last solution also shows that the N_k 's are linear in time, that is, $N_k = P(k)t$, a result

we can insert in (6.35) with $\eta = 0$. So for $k > 1$ a little rewriting yields

$$P(k) = P(k-1) \frac{k-1}{k + \frac{m+1}{m}}, \quad (6.36)$$

a result that should be compared to equation (6.9); these two coincide for $m = 1$, where the two models are identical. Equation (6.36) can be solved with initial condition $P(1) = (m+1)/(2m+1)$ so that we get a closed form expression for $P(k)$,

$$P(k) = \frac{m+1}{2m+1} \frac{\Gamma(k) \Gamma(\frac{2m+1}{m} + 1)}{\Gamma(\frac{2m+1}{m} + k)}. \quad (6.37)$$

This solution is analogous to equation (6.10), only we lose a lot of simplicity because of the non-integer nature of the fractions—they fail to cancel out. In the case where m is equal to one, this reduces to equation (6.10) and in the case $m \rightarrow \infty$, the equation (6.37) simplifies to

$$P(k) = \frac{1}{k(k+1)}. \quad (6.38)$$

Thus we find the same limiting behavior as we did in equation (6.33); this was to be expected and supports the correctness of both approaches. Using Stirling's formula to approximate the Γ -functions, we find that

$$P(k) \sim k^{-\gamma}, \quad \gamma = \frac{2m+1}{m}, \quad (6.39)$$

which is exactly the result we found using the continuum approach (equation (6.33)).

In the other limiting case $\eta = 0$, we find that, using our previous results $M_0 = t$ and $N_k(t) = P(k)t$ in equation (6.35), we find a well known recursion

$$P(k) = P(k-1) \frac{1}{1+m}, \quad (6.40)$$

which if solved with initial condition $P(1) = 1/(m+1)$ gives us an exponentially decaying distribution, $P(k) = (m+1)^{-k}$. This is (not surprisingly) the same result we found for the model with no preferential attachment, since setting $\eta = 0$ corresponds exactly to eliminating the preferential attachment. Hence we expect that the solution we find for general η will allow us to tune the model from a power-law behavior for $\eta = 1$ to an exponential decay for $\eta = 0$. In the case of the pure power-law, increasing m allows us to tune the power anywhere between $\gamma = 2$ and $\gamma = 3$ for $m = 1$ and $m \rightarrow \infty$, respectively. In the case of the exponential, increasing m makes the exponential decay faster.

Solving for the Citation Model

Now, we are able to put all of this knowledge together and solve for the model with directed link addition of m references, and $0 < \eta < 1$. Much of the work has already been done. First of all, let's put the previous work on the moments to use, by noticing that $M_0 \leq M_\eta \leq M_1$. Now, since $M_0 = t$ and $M_1 = (m+1)t$, we have that

$$M_\eta = \mu(\eta)t, \quad 1 \leq \mu \leq m+1 \quad (6.41)$$

where μ is still undetermined. We plug this into equation (6.35). It is straightforward to verify that the linear time-dependence $N_k(t) = P(k)t$ is still valid; substituting this result into equation (6.35), very little rewriting yields the recursion relation

$$P(k) = n_{k-1} \frac{m(k-1)^\eta}{(\mu + mk^\eta)}. \quad (6.42)$$

The initial condition $n_1 = \mu/(\mu + m)$ is trivial to find after these considerations, simply by inserting our assumptions into equation (6.35). Solving to find $P(k)$ is a little more involved, but the same idea as in equations (6.10) and (6.37) is employed:

$$P(k) = \frac{m(k-1)^\eta}{(\mu + mk^\eta)} P(k-1) \quad (6.43a)$$

$$= \left(\frac{m(k-1)^\eta}{\mu + mk^\eta} \right) \left(\frac{m(k-2)^\eta}{\mu + m(k-1)^\eta} \right) \left(\frac{m(k-3)^\eta}{\mu + m(k-2)^\eta} \right) \cdots P(1) \quad (6.43b)$$

$$= m^{k-1} \prod_{i=1}^{k-1} i^\eta \left(\frac{1}{\mu + mk^\eta} \right) \left(\frac{1}{\mu + m(k-1)^\eta} \right) \cdots \left(\frac{\mu}{\mu + m} \right) \quad (6.43c)$$

$$= \mu m^{k-1} \prod_{i=1}^{k-1} i^\eta \prod_{j=1}^k \frac{1}{j^\eta} \left(\frac{1}{\mu j^{-\eta} + m} \right) \quad (6.43d)$$

$$= \frac{\mu}{m} k^{-\eta} \prod_{j=1}^k \left(\frac{\mu}{mj^\eta} + 1 \right)^{-1}. \quad (6.43e)$$

In equation (6.43c) the initial condition is inserted and in equation (6.43d), the first product cancels out with part of the second, simplifying the expression. This is a closed form expression for the solution of the citation model, valid for all k .

Still, the product in equation (6.43e) does not have a great deal of intuitive appeal, it would be instructive to have a functional expression that could help us understand what is going on in figure 6.4, to understand the change from the power-law distribution to the exponential. To this end, I rewrite the product in equation (6.43e) as an exponential of the logarithm of the product, the product can now be turned into a sum; and in the continuum limit ($t \rightarrow \infty$), this sum can be treated as an integral that can be solved by expanding the logarithm. In other words

$$P(k) = \frac{\mu}{mk^\eta} \exp \left\{ \ln \prod_{j=1}^k \left(\frac{\mu}{mj^\eta} + 1 \right)^{-1} \right\} \quad (6.44a)$$

$$\approx \frac{\mu}{mk^\eta} \exp \left\{ \int_1^k -\ln \left(\frac{\mu}{mk'^\eta} + 1 \right) dk' \right\} \quad (6.44b)$$

$$\stackrel{?}{\approx} \frac{\mu}{mk^\eta} \exp \left\{ \int_1^k - \left(\frac{\mu}{mk'^\eta} - \frac{\mu^2}{2(mk')^{2\eta}} + \cdots \right) dk' \right\}. \quad (6.44c)$$

For this expansion in (6.44c) to be valid, we have to make sure that $\mu/mk^\eta \leq 1$. To this end, we have to investigate $\mu(\eta)$'s dependence on η . In equation (6.41), $\mu(\eta)$ is defined as $\mu = M_\eta/t$; using this and equation (6.43e), we can find the following implicit relation for μ ,

$$m = \sum_{k \geq 1} \frac{k^\eta}{\mu} P(k) = \sum_{k=1}^{\infty} \prod_{j=1}^k \left(\frac{\mu}{mj^\eta} + 1 \right)^{-1}. \quad (6.45)$$

This equation is difficult to analyze, and except for the limiting cases we have already discussed ($\eta = 0, 1$ where $\mu = 1, m+1$, respectively) it is (to my knowledge) impossible to extract explicit information about the dependence of μ on η . We can, however, do a numerical simulation (see figure 6.5) and visually confirm that μ varies smoothly between 1 and $m+1$ as η changes from 0 to 1. Using figure 6.5 we can convince ourselves that for $0 < \eta \leq 1$ and $k > 2$, the

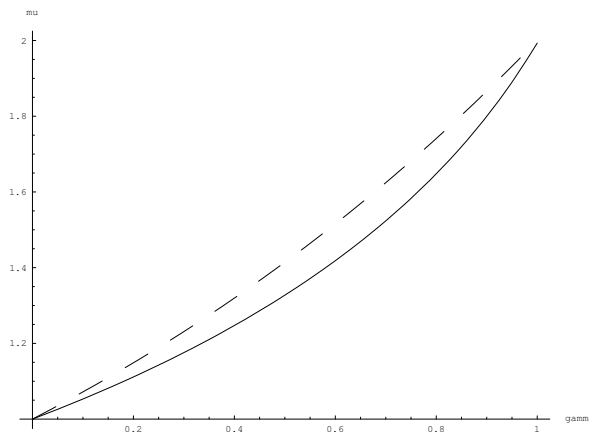


Figure 6.5: The dependence of μ on η for $m=1$. The dashed line is 2^η . If we choose $k \geq 2$ the fraction μ/mk^η is always less than or equal to one.

expansion in equation (6.44c) is allowable, *if* the integral runs from 2 to k , and we end up with the following expression for $P(k)$:

$$P(k) \simeq \frac{\mu}{\mu + m} k^{-\eta} \exp \left\{ -\frac{\mu}{m} \frac{k^{1-\eta} - 2^{1-\eta}}{1 - \eta} \right\}, \quad (6.46)$$

where only the first term of the sum inside the integral in equation (6.44c) was used. The $k = 1$ term in the sum (integral) over k results in the constant $\mu/(\mu + m)$ that is included in the normalization. For small η , it is relevant to include more terms from the expansion of the log in the integral in equation (6.44c).

Now, we can explicitly see that figure 6.4 (b) can be described as a stretched exponential. Furthermore it is clear that for η close to one ($0.8 \lesssim \eta \lesssim 1$), $P(k)$ varies very weakly with η , and it is hard to distinguish between a power-law and a stretched exponential (eq. (6.46)) for $k < e^{1/(1-\eta)}$.

This conundrum has been frequently encountered in the literature. In the case of distributions of citations, [53] found the distribution of citations of scientists to be a stretched exponential, whereas it was argued in [54] that the citation distribution of papers was described by an asymptotic power-law. The same data was attempted fitted to a curve $\sim (k_i + \text{const})^{-\alpha}$ in a later paper [55]. In a more general arena, Newman describes the distribution of the number of collaborators per publication in different databases (amongst these, SPIRES) as a stretched exponential [48], but having acquired more statistical material, the very same distribution is tentatively described as two power-laws [49] (after inspiration from [76]).

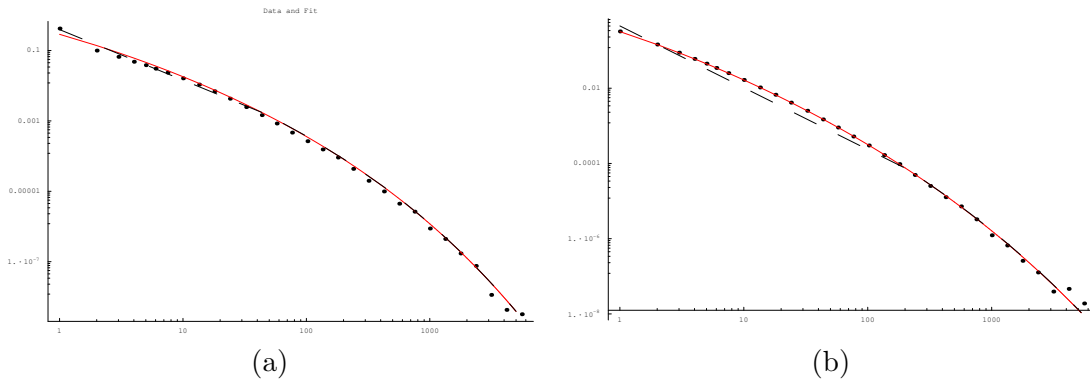


Figure 6.6: Comparing model and data. (a) The analytical solution of the citation model (red line) and data from the theory subfield (data points). The dashed line is the approximation (equation (6.46)). (b) Comparing the analytical solution of the citation model to an actual run. The data points are binned results of the data in Figure 6.4 (b). The red line is the solution and the dashed line is the approximation.

6.4 Discussion

With these analytical insights we can return to the data and compare model and solution to the real data. As was foreshadowed in Figure 6.4, the agreement is near perfect. Figure 6.6 (a) is a comparison of the simple model and the data from the theory subset; (b) is testing the analytical solution of the simple model versus an actual run of the model—it is clear that the analytical solution is correct and valid for all k , this is an impressive agreement between data and model, but as we have discussed above, and as we investigate further in the following, there is good reason to believe that this amazing correspondence is not necessarily an indication of a model that captures the structure of the citation network.

The numerics above show beyond a doubt that the simple model for the citation network is very close to the data. It is thus natural to expect that the simple model actually *does* capture important elements of how data is cited; that it does capture something universal. There are things that speak to its advantage—it is true that papers are added one by one, and that there certainly is a mechanism that makes papers that are already cited, more attractive for new papers to cite, but there is also a plethora of problems connected to this model. These will be discussed in the following.

6.4.1 Longitudinal Structure

We have spent all of the previous chapter, Chapter 5 unveiling the longitudinal correlations in the SPIRES data: some authors have citation records that are very unlikely to find in a random draw on the distribution. Clearly, the simple citation model does not contain these correlations, *simply because we have not put them there*. There are no authors in this model, only papers. Perhaps the most important step to take in modelling the citation distribution is to add a level of authors to the model so that our data acquires internal structure. This significant aspect will be discussed in Section 7.4.

6.4.2 The Age Distribution

Another problem with the simple model, is that older papers are more likely to have many citations, *cf.* the distribution of ages discussed in Section 6.1.3. The direct connection between degree and age is contradicted by the behavior of authors in SPIRES. New papers often rapidly (in the span of 3-4 years or so) acquire a massive amount of citations; they are, in other words, able to compete with the older papers in the distribution—this is not the case for the GN model. Here, all of the highly cited papers are also very old, which is a problem that needs to be addressed.

6.4.3 Measuring Preferential Attachment

Another problem relates to the nature of the preferential attachment. In the simple model, we have adjusted the preferential attachment to fit the SPIRES database. It is, in fact, possible to *measure* the nature of the preferential attachment [76, 77, 78].

The way one goes about measuring the preferential attachment is the following: Consider a network, for which we have information about the order in which each node and edge joins the system⁸. Measuring $\Pi(k(s))$ is simply keeping track of how the degree of an existing node with $k(s)$ edges grows as new nodes join the system⁹. There is a problem, however; we have that $\Pi(k(s)) = k(s)^\eta / \sum_s k(s)^\eta = c(t)k(s)^\eta$, and it is the case that the normalization $c(t)$ depends on the time at which a given node joins the system. Clearly this results in biases when measuring which old node the new nodes link to, in terms of the degree of the old nodes. The simplest way to avoid this problem is to study the attachment of new nodes in a short time frame.

Accordingly, let us call all nodes that exist in the system for T_0 nodes, and select another group of nodes that are added between T_1 and $T_1 + \Delta T$ (T_1 nodes), where $\Delta T \ll T_1$ and $T_1 > T_0$. Now, the remaining job is simply to record the degree k_i of the T_0 node to which the new node links. The $\Pi(k(s), T_0, T_1)$ function that we are looking for is simply the normalized histogram that states the number of edges acquired by the T_0 vertices with degree $k(s)$. The trick is that if the growing network develops into a stationary state (or, equivalently, if we look at small time intervals) then $\Pi(k(s), T_0, T_1)$ is independent of T_0 and T_1 , and we are then left with the preferential attachment function $\Pi(k(s))$.

Because of the need to study short time intervals, it is often practical to study the cumulated distribution $\chi(k(s))$ that is proportional to $k(s)^{\eta+1}$. In [77] the preferential attachment of several systems is measured following the procedure described above. Amongst these is data from Physical Review Letters. The nodes (1736 papers) for this investigation are papers published in 1988 and the links (83,252) are the citations that these papers have received, T_0 is chosen to be 1989. Determining η for T_1 , for 1989-1999, the value $\langle \eta \rangle = 0.95 \pm 0.1$ is found. Clearly this value disagrees with the $\eta \approx 3/4$ value that arises from fitting the analytical solution to the data for the theory subset.

It is not clear that the results from the small number of highly cited papers from PRL (average number of citations for the 1736 papers in question is circa 48 citations), but the result of measuring $\Pi(k_i)$ for a network of citations seems to indicate that another model is

⁸Unfortunately, this information is not available for the SPIRES database. It is theoretically possible to extract this information. To measure the preferential attachment for the SPIRES database is certainly an exciting project for the future.

⁹Note on notation. In this section, we are not making the continuous approximation and, therefore, there is no need for using the ‘average’ degree, $\bar{k}(s, t)$.

needed. Of course, the ideal solution would be to explicitly measure $\Pi_{SPIRES}(k(s))$, but that is not possible from the publicly accessible data.

6.4.4 The Cut-off(s)

Yet another problem, is of a different character: It occurs because of the cut-offs described earlier. First, the actual data from SPIRES has a cutoff which is discussed in Chapter 2. There is little doubt that this cut-off is a real mechanism in the SPIRES database—the idea was that ‘nobody quotes Einstein or Goldstone’. Secondly, as we have discussed earlier in this chapter, for η close to one, it is hard to distinguish the distribution created by our version of the GN model from a power-law for k , in the range that we are looking at. Let us quantify this a little. In the arguments leading to equation (6.31), we see that any model must have a cut-off, and in equation (6.31), we have calculated this cut-off for $\eta = 1$. Using the criterion in equation (6.30) on the distribution in equation (6.46) (resulting from $0 > \eta > 1$), we find that

$$k_{max} \sim \begin{cases} (\ln t)^{\frac{1}{1-\eta}}, & \text{for } 0 > \eta > 1; \\ t^{\frac{1}{1-\eta}}, & \text{asymptotically linear.} \end{cases} \quad (6.47)$$

We can use this estimate for the cut-off to test the fit generated by our model. We have $t = 159,946$ and $\eta = 0.75$, using equation (6.47), we find that $k_{max} \sim 21,000$. This, however, is a crude approximation. Using arguments similar to those of Section 2.3.1, drawing from the probability distribution defined by Equation (6.46), we would expect to find a little less than 1 paper with more than 5,242 citations, if this distribution applied to arbitrarily large k ; with a data set of 159,946 papers, we would expect the maximally cited paper to have about 4,700 citations. This fidelity to the data is alluring, but *with the data available to us at the moment*, the conclusion we arrive at is that it is impossible to draw any decisive conclusions regarding the nature of the preferential attachment. We have to look for other arguments.

6.5 Summary

In this chapter we have turned away from the data-centered investigation from the previous chapters, and begun to look at possible models for generating some of the features seen in the SPIRES database. To make the transition as smooth as possible, I have chosen to give a thorough and pedagogical introduction to the GN model, in which *three* different analytical approaches were reviewed to explain the data produced by the model: The Mean Field approach, the Rate Equation, and the Master Equation. Each of these lines of attack have strengths and weaknesses, which were pointed out. Also, limiting cases of the model were outlined, in which either of the two elements of the model, growth and preferential attachment, were removed; finally, we discussed other details about the makeup of the outputted data. The next step consisted in proposing—and solving—our own little modification of the GN model, especially designed for SPIRES. This model is original work, exclusively found in this thesis, the model fits the SPIRES data remarkably well.

Finally, the subject of whether or not any deep conclusions regarding the dynamics of SPIRES could be drawn from the model’s amazing fit to the actual data was touched upon. In discussing this subject, a few problems with the model were listed. The first important problem stems from the fact that the simple model presented here lacks the longitudinal correlations we have spent Chapter 5 analyzing; that this structure is missing informs us that

this model is still far removed from SPIRES. The second important problem is derived from the fact that the age of a given node is closely related to the degree of that node (*cf.* the discussion towards the end of Section 6.1.3)—this picture is vastly different from SPIRES, where it is *not* the case that the oldest papers possess all of the incoming links. The third problem that was raised, was that recent measurements of the preferential attachment in another set of citation data, seems to indicate that the model we set up to fit SPIRES, may be wrong, since the value of the parameter η used in our model, has a different value than the value measured in the data from Physical Review Letters. The next problem we discussed, has a slightly different character, *viz.* that the data available to us from SPIRES does not allow us to draw any definite conclusions with respect to the size of the parameter η , in our model.

The next chapter is going to continue in the same direction as this chapter, only that we are going to be a little more constructive. Instead of just pointing to problems with the model, we will begin to see what happens to the model, when we modify it to be more realistic. The first part of this chapter will mainly be a review of existing work, but rather exciting nonetheless, and in the latter part look at the possibilities of including the author-level in the description.

In the previous chapter, the GN model was scrutinized from many different angles. In the present chapter, we are going to benefit from this intimate knowledge of the GN model, and begin to think about how we can change the simple version into a more realistic model of SPIRES. The chapter begins with a continuation of the discussion of the cut-off, this time seen from a new angle.

We then review a few minor objections to the model and the possible modifications of the model that can be implemented in order to take these objections into account. The subjects are initial attractiveness of nodes, edge redirection, and ageing. None of these models are original; they are, however, interesting from a theoretical point of view, and as such well worth our while. The most interesting of these discussions, is the one on ageing. This is partially due to the fact that the inclusion of gradual ageing in the model pushes it to the verge of being analytically intractable. Further, as we have discussed in some length earlier, it is the case for the GN model that the nodes with the most citations are also the oldest (recall, that in the model $k(s, t) \sim (s/t)^{-\beta}$, $\beta = m/(m + 1)$, in the simple case, $\eta = 1$), but surely age has an effect on nodes: In the case of citation data, we know from experience that some papers age gracefully, and indeed keep picking up citations, but surely age takes its toll on other papers, leaving them stagnant or ‘dead’ with a static number of citations.

The second part of the chapter regards a more fundamental change of the GN model, in which we consider a population of authors publishing papers, and where the probability of being cited is determined by the number of citations by the paper’s author rather than by the number of citations of the paper itself.

7.1 Initial Attractiveness

The primary virtue of the simple SPIRES model from Chapter 6, is that it results in power-laws (at least for $\eta = 1$) and even when it does not ($\eta < 1$) it results in stretched exponentials; the criticisms in the concluding sections of last chapter, should have convinced us not to be too awe-stricken by the model’s truly fantastic reproduction of the degree distribution of papers in SPIRES. In the context of modelling real citation networks, the exponential cut-off of the

power-law seems to be a valuable trait, since many real networks with close-to-power-law shapes, still appear a little curved just before their power-law stops, either because of a finite size effect¹ or because of other cut-off mechanisms, *cf.* Sections 6.2.3 and 6.4.4. The following discussion will add yet another argument to this discussion.

We can think of the parameter η as a way of tuning the preferential attachment. Say we have the probabilities $p_k \sim 1$ and $q_k \sim k$, the $\Pi(k(s)) \sim k(s)^\eta$ model makes a *geometrical* interpolation from p_k to q_k by creating new probabilities

$$p_k^{1-\eta} q_k^\eta. \quad (7.1)$$

This interpolation leads to the stretched exponentials described above. Instead of this geometrical interpolation, we could just as easily make an *arithmetical* interpolation, with new probabilities that are proportional to

$$(1 - \nu)p_k + \nu q_k. \quad (7.2)$$

In practice, this transition can be realized simply via loading each paper with w citations instead of just one, as we have previously done². These ‘ghost’ citations can be subtracted later, but in the present context, they serve as a way of increasing the initial attractiveness of the nodes. After the calculations, we can simply let these ghosts disappear back into nothing. In the limit $\nu \rightarrow \infty$, we know that the distribution is described by exponential decay, *cf.* Equation (6.26), since all preferential attachment has been eliminated from the system. Now, let us investigate what happens in between.

We have that $\Pi(k(s)) \sim k(s) + w$. For a rough estimate of the asymptotic behavior, we can use Barabási’s method. Solving the DE

$$\frac{\partial \bar{k}(s, t)}{\partial t} = \frac{m \bar{k}(s, t)}{(m + w)t}, \quad (7.3)$$

with the now obvious initial condition $\bar{k}(t, t) = w$, we find that

$$\bar{k}(s, t) = w \left(\frac{t}{s} \right)^\beta, \quad \beta = \frac{m}{m + w}. \quad (7.4)$$

We can use Equation (6.21) to directly estimate $P(k)$, resulting in

$$P(k) = \left[t \frac{\partial \bar{k}(s, t)}{\partial s} \right]^{-1} \bigg|_{s=s(k, t)} = \frac{w^{1/\beta}}{\beta} k^{-\gamma}; \quad \gamma = \frac{1}{\beta} + 1 = \frac{w + 2m}{m}, \quad (7.5)$$

where $s = s(k, t)$ is the solution of Equation (7.4). Thus, the asymptotic power-law can be tuned to any value $\gamma \in [2, \infty[$.

¹For example, this must be the case for the co-author network. This network has a physical limit to the number of collaborators that a given author can have, since only so many papers can be written within a lifetime

²This calculation was first performed in [43], using the master equation approach (this paper is also an excellent illustration of how ‘involved’ the calculations become, when using the master equation without going to the continuous limit). The calculations found here were performed independently hereof, using the rate equation, and were inspired by discussions with A. D. Jackson.

To become acquainted with the exact form of the solution in the low k regime, we have to resort to the rate equation. We proceed as we have done quite a few times in the previous chapter, to find the following recursion relation

$$P(k) = P(k-1) \frac{m(k+w-1)}{m+w+w(k+w)}, \quad (7.6)$$

which along with the initial condition that $P(w) = (m+w)/(m+w+2mw)$, yields the exact expression for the degree distribution

$$P(k) = \frac{(m+w)\Gamma(3+w+\frac{w}{m})\Gamma(w+k)}{(1+m+w+2mw)\Gamma(w+1)\Gamma(2+w+\frac{w}{m}+k)}, \quad (7.7)$$

reinsuring us that we indeed have an asymptotic power-law with the slope predicted in Equation (7.4). It is interesting to note that the arithmetical interpolation results in power-laws for any finite value of w , and yields an exponential only in the case $w \rightarrow \infty$. In the case of the geometrical interpolation, we have the relation that we only find power-laws for $\eta = 1$ and stretched exponentials in all other cases. This suggests that these limits are (at least somewhat) non-trivial, and it would be interesting to study this in more detail.

Plotting Equation (7.7) against the data from the theory subfield, makes another highly interesting point regarding the subject of the cut-off that was touched upon earlier (*cf.* Sections 6.2.3 and 6.4.4). Thus, this calculation further strengthens the argument that because of the cut-off in the SPIRES data, it is impossible to distinguish between a model that has an exponential cut-off and a model, such as this one, that results in an asymptotic power-law. As demonstrated in Chapter 2, our data is of a much higher quality than the ISI and PRD data sets, discussed in [54]. But it seems that even with access to the highly homogeneous SPIRES database, the ‘scientific saint’ cut-off mechanism still leaves room for speculation as to the topology of the citation distribution. Arguments regarding the ‘microscopic’ citation mechanisms will have to be made before any model of the citation network based on the data presently available, can be taken seriously. For further input on this discussion, the reader should also recall Section 2.3.1.

7.2 Edge Redirection

The next step in adapting the GN model to the real world stems from taking human laziness and vanity into account. In an amusing paper, entitled *Read before you cite!* [79], Simkin and Roychowdhury have developed a method to estimate how many papers, in a given paper’s reference list, the author(s) have actually *read*. Citing Freud (!), they connect scientific misprints to freudian slips and find that an alarming number of misprints in scientific papers are identical: Take a 4-digit page number with one digit misprinted. There are 10^4 possible misprints, which makes the probability of a repeat misprint 10^{-4} . Since the probability of coincidental misprints is so small, it is natural for Simkin and Roychowdhury to conclude that misprints are due to copying someone else’s misprint, without reading the paper in question. They argue that it is, of course, possible to actually read a paper and proceed to copy the bibliographic entry from some unreliable source, but the authors claim that this is highly unlikely because of the following mechanism: If someone has gotten a hold of the original paper from a faulty reference list, then they are likely to have encountered problems with

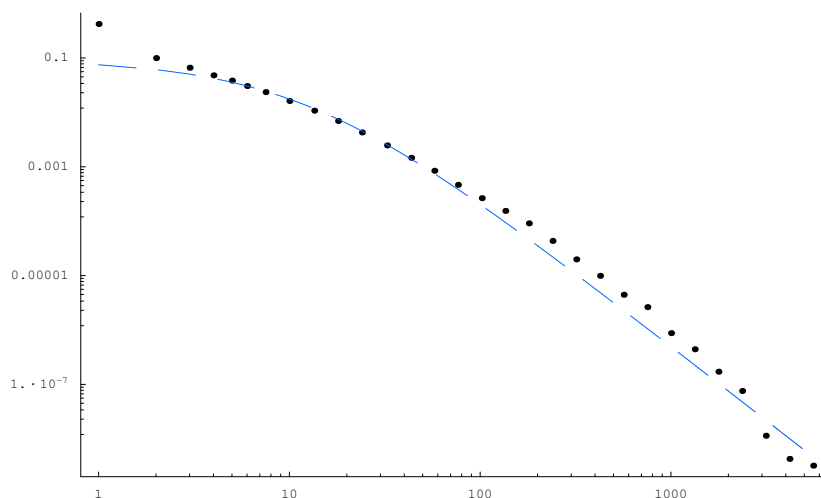


Figure 7.1: Comparing Equation (7.7) and data from the theory subfield. The data is represented using dots, whereas the dashed line is given by equation (7.7). The values of m and w are set to 15 and 9, respectively, this corresponds to an asymptotic power-law with slope $\gamma_B = 2.6$

finding the right paper, and therefore noticed the misprint and corrected it in their own reference list.

This is not the place to go into the details of their investigation. It is sufficient to merely state that Simkin and Roychowdhury conclude that approximately only 20% of papers, in a given reference list, have actually been read by the author. Whether or not this number is correct can be debated, but it raises an important issue, namely the influence of *edge redirection* on the citation distribution. Even though Simkin and Roychowdhury may have overestimated the number of unread papers in science, it is beyond questioning that—at a certain level—they have revealed an important detail about citation networks. When writing a paper, one cites a number of papers that have been studied with various degrees of intensity. Furthermore, one is also likely to include a couple of important papers from the reference list of these papers; in summary, many citations appear via sloppily copying from other people's reference lists. We can think of this as rewiring links in the network.

The effects of re-wiring links was investigated in [75]. A simple model of rewiring is the following: We only discuss the in-degree distribution, and at each time step, a new node \mathbf{n} is added to the system and an earlier node \mathbf{x} is selected *uniformly* as a possible target for attachment. With probability $1 - r$, a link from \mathbf{n} to \mathbf{x} appears, and with probability r the link is redirected to the ancestor node of \mathbf{x} , that is the node that \mathbf{x} links to \mathbf{y} . This process is illustrated in Figure 7.2.

Let us set up the rate equations for the redirection network. According to the rules defined above, we find that the degree distribution $N_k(t)$, evolves by the equations

$$\frac{dN_k}{dt} = \frac{1-r}{M_0} [N_{k-1} - N_k] + \frac{r}{M_0} [(k-2)N_{k-1} - (k-1)N_k] + \delta_{1k}. \quad (7.8)$$

For redirection probability $r = 0$ we are left with the GN model without any preferential attachment³, (the uniform selection of \mathbf{x}). The two rightmost terms account for the redirection

³Had we chosen to select the target node preferentially rather than uniformly, this model simply collapses

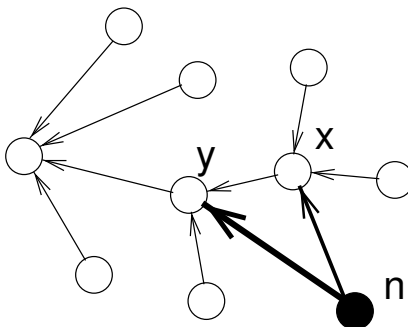


Figure 7.2: The processes in the redirection network. The new node \mathbf{n} selects a target node \mathbf{x} . With probability $1 - r$, \mathbf{n} links to \mathbf{x} (dashed arrow) and with probability r , it links to the ancestor of \mathbf{x} , \mathbf{y} (thick arrow). From [75].

process. The first of these is the gain term: Since the first node was selected uniformly, if redirection indeed does occur, then the probability that a node with $k - 1$ pre-existing links receives the new redirected edge is proportional to $k - 2$, the number of pre-existing incoming links. The argument for the loss term (the next term to the right) is analogous. Equation (7.8) applies to all $k \geq 1$.

Now, we can rewrite Equation (7.8) so that it reduces to the original rate equation with a preferential attachment $\Pi(k) \sim A_k$, where $A_k = r[(k - 1) + (1 - r)/r]$. Scaling out the factor r , we find that A_k can be reduced to $A_k = k + w$, where $w = 1/r - 2$. In other words, the redirection is equivalent to adding initial attractiveness, which is exactly the case we have solved above (Section 7.1). For $r = 1/11$ we find $w = 9$, which fits the data quite well, *cf.* Figure 7.1. Note, however, that $r = 1/11$ is a rather small probability for redirection, since Simkin and Roychowdhury estimated it to be about 80%. On the other hand, the model presented here is crude and only intended to explain some of the basic mechanisms of edge redirections.

7.3 Ageing

7.3.1 Measuring Ageing in SPIRES

In order to measure explicit ageing in SPIRES, one needs to know *when* individual papers received their citations. Unfortunately, this information is not available in the data from SPIRES. However, in 1999 Benny Lautrup was working on a project on the ageing of papers, which I am using here (this is still work in progress [80]). In Figure 7.3, we see the result of Prof. Lautrup's investigations. The age of papers is measured relative to 1999, so a paper published in 1999 has age 1, a paper published in 1995 has age 5, *etc.* A 'live paper' is defined as a paper that was cited in 1999,⁴ and the figure shows the fraction of live papers compared to the number of papers published in the same year. There are, however, further

to the regular GN network, with the only difference that $\Pi(k(s)) \sim k - 1$.

⁴This seems to be a reasonable assumption. If a paper has not been cited for the last 12 months, it is not likely that it will be cited again.

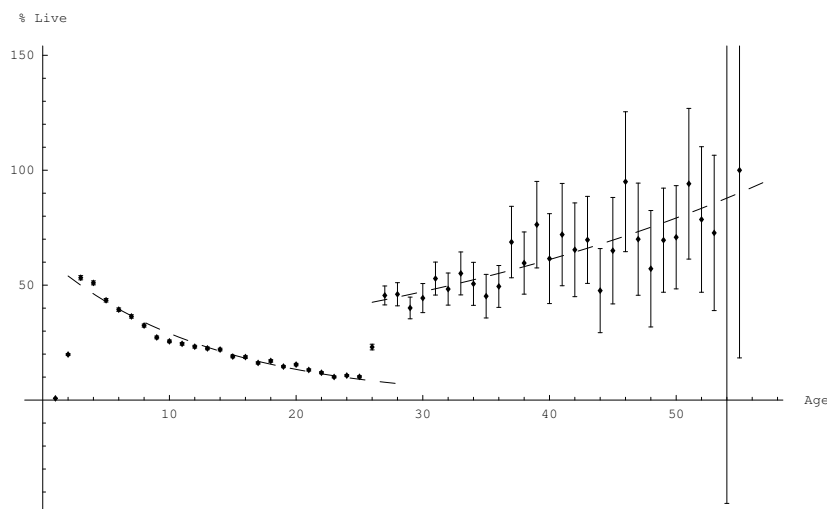


Figure 7.3: The fraction of live papers. The age of papers is measured relative to 1999, so a paper published in 1999 has age 1, a paper published in 1995 has age 5, *etc.* A ‘live paper’ is defined as a paper that was cited in 1999, and the figure shows the fraction of live papers compared to the number of papers published in the same year. The error bars are the square roots of the paper count.

complications: There is a significant bump around 1974, when SPIRES began collecting papers systematically. It is relatively easy to understand the nature of this bump; before 1974, not all papers were collected. It was only the ones that the Stanford librarians found interesting, since they were inclined to collect what they considered ‘important’ papers, it is only natural that a higher fraction of the papers from before 1974 are still ‘live’, *cf.* Section 1.4. After 1974, as we know, the collection of papers is more complete, and almost every paper in high energy physics was added to the SPIRES database. My observations in the following is based on the 1974+ population.

The important thing here is that this histogram can be interpreted as an indication that the popularity of papers seems to fall off; that we don’t cite as many old papers. If we assume *a priori* that the quality of cited papers is approximately the same year after year⁵, then, after 1974, an ageing of papers is exactly what we see in the figure. Even though there is a steady drop in the live fraction year by year, the older population is an indication that some papers indeed have *very* long life spans; these are the classics.

Since these are not my own data, I will not investigate this *highly interesting* subject any further, at least empirically. This, however, constitutes solid proof that the real network of scientific communications displays ageing. Let us investigate the consequences for the GN model.

7.3.2 Analytical Results

Treating explicit ageing in networks analytically was done by Dorogovtsev and Mendes [81]. In the following, let us review the most important calculations in a pedagogical manner, since the results here are very interesting and important. It is clear from our intuitions *and from*

⁵This is another reasonable assumption, since there is no reason to believe (most older men would probably debate this, but I strongly suspect that this is a mistake) that there should be a drop in IQ over time.

the data (*cf.* Section 7.3.1), that including ageing is an important step towards adapting the citation network to the real world.

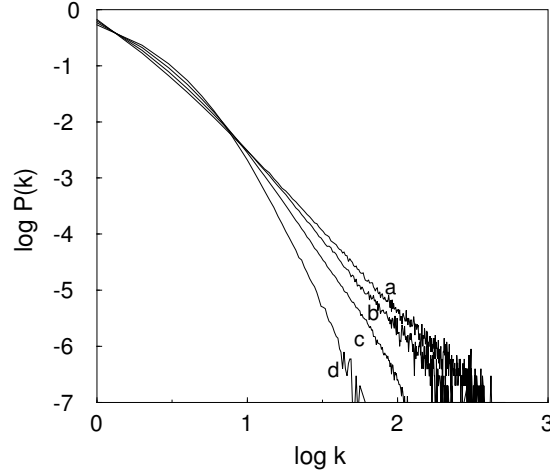


Figure 7.4: The degree distribution for different values of the ageing exponent. Curve a, $\alpha = 0$; curve b, $\alpha = 0.25$; curve c, $\alpha = 0.50$; curve d $\alpha = 0.75$. It is clear that the number of highly connected nodes are reduced as it is expected. Also, the data still follows asymptotic power-laws. From [81].

So, what happens if the probability for a new node to connect to a node, already in the system, depends—not only on the degree of the old node—but was also proportional to the power of its age, $\tau^{-\alpha}$? Let us just solve for the simplest imaginable case where each paper comes with one citation, since these results are easily generalized to m papers. In Figure 7.4, numerical runs of this model are presented. This calculation is exceedingly comprehensive in the exact version, so let us only regard the scaling and proceed in the continuous approximation. The introduction of ageing leads to the differential equation:

$$\frac{\partial \bar{k}(s, t)}{\partial t} = \frac{\bar{k}(s, t)(t - s)^{-\alpha}}{\int_0^t du \bar{k}(u, t)(t - u)^{-\alpha}} \quad (7.9)$$

that we solve with the initial condition $\bar{k}(t, t) = 1$. To solve this equation, we look for a solution of the scaling form, that is

$$\bar{k}(s, t) \equiv \kappa(s/t), \quad s/t \equiv \xi. \quad (7.10)$$

Inserting this into Equation (7.9) and rewriting it a little, we can find

$$-\xi(1 - \xi)^\alpha \frac{d \ln \kappa(\xi)}{d \xi} = \left\{ \int_0^1 d\zeta \kappa(\zeta)(1 - \zeta)^{-\alpha} \right\}^{-1} \equiv \beta, \quad (7.11)$$

where the initial condition becomes $\kappa(1) = 1$. Here, β is a constant that is still unknown, but will turn out to be the exponent of the average degree, $\bar{k}(s, t)$. We know from earlier, that $\int_0^t dk \bar{k}(s, t) = 2t$ and, using Equation (7.10), this transforms into $\int_0^1 d\zeta \kappa(\zeta) = 2$. The solution

to Equation (7.11) is found simply by integrating over ξ to find

$$\kappa(\xi) = C \exp \left(-\beta \int \frac{d\xi}{\xi(1-\xi)^\alpha} \right), \quad (7.12)$$

where C is a constant. The solution of the indefinite integral in Equation (7.12) can be expressed via special functions to yield

$$\int \frac{d\xi}{\xi(1-\xi)^\alpha} = \ln \xi + \alpha {}_3F_2 \left[\begin{matrix} 1, 1, 1+\alpha \\ 2, 2 \end{matrix}; \xi \right]. \quad (7.13)$$

Here, ${}_3F_2()$ is the generalized hypergeometric function, *cf.* Appendix B.2 and [62]. We can use the boundary condition $\kappa(1) = 1$ to determine the constant C and find the final expression for $\kappa(\xi)$, we have that

$$\kappa(\xi) = \xi^{-\beta} \overbrace{\exp[-\beta(\gamma_{EM} + \Psi(1-\alpha))]}^C \exp[-\beta\alpha {}_3F_2(1, 1, 1+\alpha; 2, 2; \xi)]. \quad (7.14)$$

In the normalization, γ_{EM} is the Euler-Mascheroni constant, $\gamma_{EM} \approx 0.57721566490153286\dots$, and $\Psi()$ is the digamma function [62]. That β is indeed the exponent of $\bar{k}(s, t)$, can be seen from the fact that $\kappa(\xi) \sim \xi^{-\beta}$ for $\xi \rightarrow 0$ (for $t \gg s$); in the scaling regime. Finding an explicit expression for β is not easy, but by substituting the solution for $\kappa(\xi)$, from Equation (7.14), into the differential equation (equation (7.11)), we can find the following transcendental expression for β (we could have used $\int_0^1 d\zeta \kappa(\zeta) = 2$ to find an equivalent expression)

$$\beta^{-1} = e^{-\beta(\gamma_{EM} + \Psi(1-\alpha))} \int_0^1 d\zeta \frac{\exp[-\beta\alpha {}_3F_2(1, 1, 1+\alpha; 2, 2; \zeta)]}{\zeta^\beta(1-\zeta)^\alpha}, \quad (7.15)$$

from which we can analyze the properties of the network with ageing. The solution of Equation (7.15) is plotted in Figure 7.5 (a). Before we begin considering equation (7.15), the reader should note the relationship $\beta(\gamma - 1) = 1$ from equation (6.22). The general nature of this relation makes it valid also in the case of ageing, so we can directly extract information on the scaling properties of the degree distribution from β . In Figure 7.5 (b), the dependence of γ on α is plotted.

Equation (7.15) has a solution in the range $-\infty < \alpha < 1$. It is possible to find explicit expressions for $\beta(\alpha)$ and $\gamma(\alpha)$ in a couple of simple cases. For $\alpha \rightarrow 1$, we find that

$$\beta \cong c(1-\alpha), \quad \gamma \cong \frac{1}{c(1-\alpha)}; \quad (7.16)$$

here, the relation ${}_3F_2(1, 1, 2; 2, 2; \zeta) = -\ln(1-\zeta)/\zeta$ was used, and the constant $c = 0.8064659942\dots$ is the solution of the implicit relation $1+1/c = \exp(c)$. In the limit $\alpha \rightarrow -\infty$, we find that

$$\beta \rightarrow 1 \quad \text{and} \quad \gamma \rightarrow 2. \quad (7.17)$$

Thus we have that the ranges of the exponents are $0 < \beta < 1$ and $2 < \gamma < \infty$.

In summary, the main results of the investigation of the GN model with the inclusion of gradual ageing of the sites (the factor τ^α) are: Both simulations and the continuum solution show that the network exhibits power-law behavior in the case $\alpha < 1$. The results for the exponents of the degree distribution $P(k, t) \sim k^{-\gamma}$ and the average degree $\bar{k}(s, t) \sim s^{-\beta}$ were determined analytically (Equation (7.15)); we also discussed some of the limits of α , where reasonably simple expressions for $\beta(\alpha)$ and $\gamma(\alpha)$ could be found.

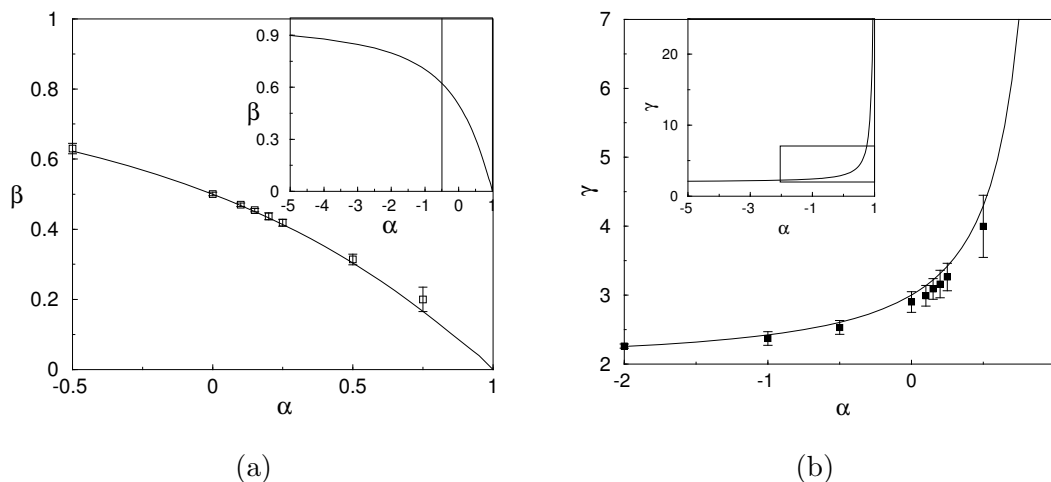


Figure 7.5: (a) The solution of Equation (7.15) (line). The points originate from the numerical simulation. The inset shows the same solution for a larger range of α values, $-5 < \alpha < 1$. (b) The solution of equation (7.15) (line) in terms of γ , see the main text for details. The points are the results of simulations. Again, the inset shows the same solution for a larger range of α values, $-5 < \alpha < 1$. From [81].

7.4 An Author Model

In this final section we are going to discuss an obvious modification of the GN model, in which we let the preferential attachment depend on the citations of the *author* of a given paper, rather than on the citations of the paper itself. This is a reasonable assumption, when we consider aspects of the mentality of the citation network: First of all because there is a tendency to pay more attention to papers written by ‘big names’ within the subfield one works and, second of all, because ‘big names’ have a way of showing up on the list of references of most papers. This is not only because people have actually used their work, but also because these ‘big names’ on a list of references, by some psychological mechanism, seem to give credibility to the paper in question (something like “... author \mathcal{U} doesn’t seem to be all that bright, but if he is doing the same stuff as author \mathcal{C} , he’s gotta be doing *something* right...”).

7.4.1 Defining the Author Model

At a first sight, letting the author citations supply the preferential attachment sounds like a promising augmentation of the GN model—and easy to implement. The notation introduced in Section 3.1.2 is ideally suited to describe this model, so we will use it here. We start out by defining \mathcal{N} authors, $\mathbf{K}_i(t)$, and define the model using the same two elements that we know from the GN model:

- *Growth.* We initialize the model by letting the first m_0 authors publish a paper with one citation each (their initial attractiveness). Author \mathbf{K}_{m_0+1} then proceeds to publish his first paper, $k_{m_0+1,(1)}$ —each new paper has one citation to begin with (again to supply the initial attraction). This paper refers to $m \leq m_0$ papers already present in the

database, *etc.* At each time step, a new author publishes a paper that links to m papers in the database. Thus, during the first $\mathcal{N} - m_0$ time steps, the model is identical to the GN model, but after $\mathcal{N} + 1 - m_0$ time steps, author \mathbf{K}_1 publishes his second paper $k_{1(2)}$, *etc.*

- *Preferential attachment.* The probability $\Pi(k_{a(b)}(t))$ that, at time step t , a new node attaches to a node at site $a(b)$ (the b th paper by the a th author) is proportional to the total number of publications by author $\mathbf{K}_a(t)$, that is, the sum of citations of his publications, $\sum_b k_{a(b)}(t) = n_a(t)$. The nature of the preferential attachment thus becomes:

$$\Pi[k_{a(b)}(t)] = m \frac{n_a(t)}{\sum_c n_c}. \quad (7.18)$$

Now, let us investigate where this model takes us.

7.4.2 Results

The results from a numerical run with $\mathcal{N} = 1000$ authors with 15 publications each, and each publication distributing 10 references. The resulting paper citation distribution is plotted in Figure 7.6. This figure immediately leads our thoughts to Figure 6.3. Also, this ‘atomic

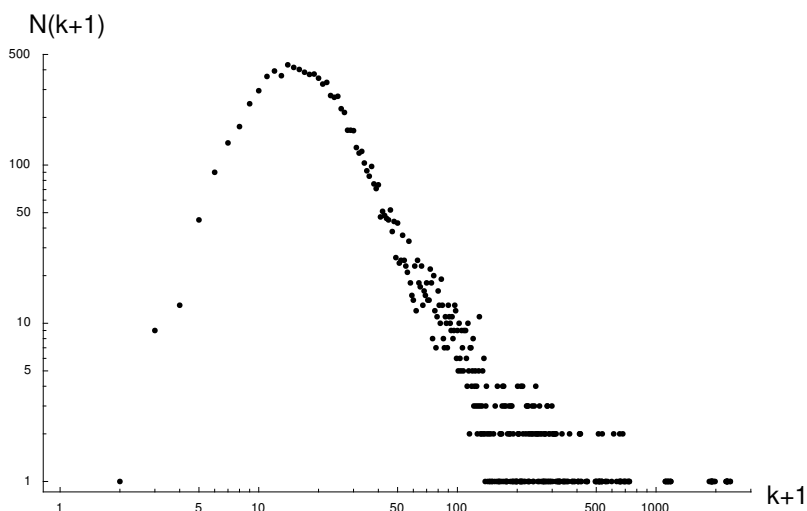


Figure 7.6: An ‘atomic histogram’ plot of the degree distribution for the author-level GN model.

histogram’ immediately puts us on track of the problem with this model. The level of authors *is identical to the ‘no growth’ version of the GN network* that we discussed in Section 6.2.2; therefore, the results found there (Equation (6.29)) can, of course, be directly transferred to this case. With the preferential attachment governed by a Gaussian distribution, we find the same distribution of paper citations. This is what we see in Figure 7.6.

The idea of adding a level of authors to the model is sound, but new attempts to model SPIRES are going to have to produce an author population with publication lengths approximately described by the distribution number of papers per author seen in Figure 3.5—instead of all authors having the same number of publications as it is the case for this model. One

idea could be to simply remove and add one author from a random site once in a while; this would keep the power-law distribution from saturating and turning into a Gaussian. We will, however, not investigate this model further, but the concept of including a level of authors, of course, provides an interesting starting point for further modelling.

7.5 Summary

We began this chapter, where the previous one left off—with a discussion of the nature of the preferential attachment in the GN model. Instead of tuning the probability of citation with the η -parameter, we chose to tune the preferential attachment in a different manner, *viz.* by loading each paper with a number of ‘ghost’ citations—to add initial attractiveness. There were two lessons to learn here: Firstly, that the model modified in this fashion results in power-law degree distributions for any finite number of initial ‘ghost’ citations, and not the stretched exponentials we know from Chapter 6. Secondly, and more importantly, we realized that on the basis of the SPIRES data, we cannot make any decisive conclusions on the asymptotic shape of the paper citation distribution; the SPIRES data is well described by both power-law *and* stretched exponential models.

After this, we proceeded to discuss edge redirection, inspired by [79]. Using the rate equation, we found that a network with redirection of links is equivalent to a network, where each paper is loaded with an initial number of papers—thus connecting to the ‘initial attractiveness’ solved immediately above.

The next subject was the ageing of papers. The discussion here, was motivated by a plot of the ‘live’ papers in SPIRES, which clearly indicated that papers in SPIRES age gradually. We were able to draw this conclusion, because the fraction of ‘live’ papers is a monotonically decaying function of the age (after 1974). This motivated a pedagogical⁶ review of the effects of including gradual ageing in the network model, first presented in [81]. Using tools from the continuous approximation (derived in Section 6.1.3)—and refreshing our knowledge of special functions—we found an explicit expression for the two exponents β and γ , characterizing the GN network.

Finally, we considered the effect of letting the preferential attachment in the GN model be governed, not by the number of citations but, by the number of citations of a level of *authors* of the papers in the database. We learned, however, that this level of authors is an exact representation of the ‘no growth’ version of the GN network, discussed in Section 6.2.2. The idea of adding a level of authors in order to create longitudinal correlations in the model is an extremely interesting (and easy⁷) starting point for further modelling.

⁶At least far more pedagogical than the original paper, *cf.* [81].

⁷Admittedly, I began working with this model at an unreasonably late stage in the writing of this thesis; therefore, the results are not at all as impressive as they easily could have been.

Concluding Scientific Postscript

So where are we now? After more than a hundred pages of L^AT_EX-output, close to 150 distinct *Mathematica* notebooks, around a thousand lines of MATLAB and PERL code, endless sub-directories of graphics output, various versions of raw data from SPIRES, and bulk data output from various models and simulations. With all of this adding up to some 400+ MB of hard disk space, and an unhealthy amount of man-hours, what lessons have we learned?

This question is answered in the beginning of this thesis; the reader who needs a brief answer to this question can simply flip back to the Abstract, where the main conclusions are summarized rather succinctly. In this concluding postscript, I am going to consider a slightly different question, namely: *Where will we go from here?*

8.1 Future Directions

In this thesis, I have done my best to disguise the fact that the investigation of SPIRES is very much a work in progress. Much more than a closed work, this thesis is a snapshot of an ongoing investigation, and writing it has been a necessary—and at most times somewhat tedious—break away from what is truly important: The fascinating study of principles of order emerging from behind a veil of complexity; away from using the tools of physics to explore the world around us. In the following, I will point in some of the directions the continued unveiling of the SPIRES network are likely to take us.

8.1.1 More Data

As was made clear in Section 3.1.2, the data from SPIRES is not complete. We lack access to *when* and *from whom* a given paper's citations came. One of the first orders of business in the continued investigation, is to get a hold of all of the reference lists of all papers in SPIRES (this is virtually tantamount to downloading the entire database). The new data will enable the calculation of every conceivable quantity of interest in the network, *e.g.* path lengths, out-bound degree structure, *etc*; we will be in possession of an exact representation of the network of references in high energy physics.

8.1.2 Temporal Aspects

The acquisition of more data will enable us to initiate work with the temporal evolution of the database. This aspect of SPIRES has been suppressed in the work so far. We will be able to ‘build’ SPIRES from ‘paper one’ and watch it grow: What is the structure—both temporally and ‘spatially’ that emerges for the paper network—for the author network? How do the different distributions of authors, papers, author paper averages, *etc.*, evolve in time? These are all interesting questions that it will be inspiring to answer.

Another interesting venue in this respect, is the study of the time evolution of the principal components and the Power of Excellence for authors in general—refining the quest to pinpoint excellence early in the careers of authors; recall that this subject was touched upon in Chapter 4.

8.1.3 Paper Citation Histories

Within the question of the temporal evolution of SPIRES, we find yet another quantity of interest. The typical ways in which *papers* are cited. Intuitively, we expect that there are several types of papers; the ‘canonized’ papers that stop being cited because of their acceptance in another realm beyond this one; the classics that continue to harvest citations, but are not quite ‘canonized’, the papers that are heavily cited initially, but are replaced by more general results after a while, *etc.* So far our only results are the (somewhat meager) ones found in Section 7.3.

A way to make quantitative investigations of how papers age, is by performing PCA on the paper histories. This course of action will allow us to determine exactly the archetypes of papers’ evolution in time.

8.1.4 Refining the Model(s)

There is two approaches to modelling. The type we have used here is based on setting up a *simple* minimalistic model, trying to capture the *essential mechanisms* of a system. The idea is to use the model to investigate the *structure* of the system in question. If the output of the model resembles the actual system, we usually assume that we understand the system. This is the modelling approach used in the physics of complex systems, *cf.* Chapter 1. There is no doubt that our ability to choose which generic mechanisms to include in this type of modelling, of the citation network, can benefit from seeing the network grow ‘from paper one’, as described above. Further, in Chapters 6 and 7 there are many suggestions to other improvements that can be added to a new model—this is one of the main directions to take in the project of analyzing SPIRES.

The other approach to modelling is the one that is used in models for the climate, or for modelling something that has any kind of practical use (*e.g.* the dissipation of heat in some part of a new car engine). With the information on the temporal evolution of SPIRES, we will be able to enter this new modelling terrain; to create models that imitate every aspect of SPIRES, rather than trying to generate the right network structure. We will be able to create very precise models that will be able to generate *more* data, predict how the structure of the citation network changes in time, *etc.*

8.1.5 Down the Rabbit Hole

A completely different direction to take this project, is to cross the field boundaries and present the results found in the preceding chapters to sociologists and anthropologists—to open the door to a more qualitative analysis based on *their* theorists and theories. As I have pointed out earlier, these sciences are often based on qualitative investigations and not on statistically significant sets of data.

The results provided here, will surely fuel discussions within these fields; many of the concepts—for example the author citation histories, modelling, or the concept of the Scientific Staff—are new in these fields and will hopefully change them completely. As to the existing concepts, for example the paper citation histories, ‘socio-scientists’ in the humanities have already shown a considerable interest in attaining real data on the subject. Some of my ideas along these lines are described in Section 8.1.3. In initiating such a cooperation, we leave behind the solid ground that the method of physics provides us with; however, the enterprize of leaving the solid ground behind is not necessarily uninteresting—just ask Alice¹.



¹Although eccentric, the Red Queen appears to be an interesting acquaintance.

A.1 The Acquisition of Data

The data used in this *Cand. Scient. Thesis* was collected on August 14th 2002 by the author from the SPIRES hep mirror at the university of Durham¹. The information on collecting data is placed in this appendix, because we believe it to be of relatively little interest to the general reader. We have supplied detailed information on the subject of gathering information for completeness and in order to facilitate recreating results in this paper.

A.1.1 Navigating in SPIRES

SPIRES is first and foremost a programming language that was written in order to create and search databases, *cf.* Section 1.4. The commands used in the following are all explained in detail in the SPIRES manual [82].

The Durham server was accessed via a telnet connection established by the freeware program `putty.exe`. One enters the SPIRES database by entering `spires` at the prompt, and the hep database is selected by typing `select hep`. The SPIRES database is divided into six sub-groups that are accessed by applying the search modifiers (to set the search modifiers one has to type `set sea mod [modifier]` at the prompt):

- Theory: `and ps T not ps E`
- Phenomenology: `and ps T, E`
- Experiment: `and ps E not ps T`
- Instrumentation: `and ps I`
- Review: `and ps R`
- Published: `and ps published`

¹Further information is available at <http://www.dur.ac.uk/>.

| | 1945 – 79 | 1980 – 89 | 1990 – 94 | 1995 - 99 | 2000 – 02 | Sum |
|-----------------|-----------|-----------|-----------|-----------|-----------|---------|
| Total | 68,848 | 133,280 | 94,173 | 117,876 | 87,354 | 501,531 |
| Theory | 29,130 | 65,332 | 43,904 | 45,268 | 25,778 | 210,412 |
| Phenomenology | 9,209 | 16,098 | 13,574 | 27,374 | 17,850 | 84,105 |
| Experiment | 10,794 | 15,346 | 8,823 | 10,643 | 9,200 | 54,806 |
| Instrumentation | 6,873 | 10,903 | 8,981 | 9,926 | 4,695 | 41,405 |
| Review | 1,393 | 3,741 | 2,212 | 2,268 | 1,255 | 13,081 |
| Test | 57,399 | 111,447 | 77,494 | 95,479 | 59,778 | 403,809 |
| Discrepancy | 11,449 | 21,833 | 16,679 | 22,397 | 27,576 | 97,722 |
| Discr. % | 16.6 | 16.4 | 17.7 | 19.9 | 31.6 | 19.5 |

Table A.1: The data categorized according to the searches in the Durham database. The first column contains the different categories that the database can be divided into. The next five columns are the different time intervals that the database was divided into and the final column is the sum of the numbers in the previous five columns, thus representing the total number of papers in each category. The row labelled ‘Test’ is the sum of the number of papers in the sub categories within each time period. Ideally this number should correspond to the number of papers in the ‘Total’ category, but as it is apparent from the table, this is not the case. The discrepancy is listed below, both as the number of papers that are missing, and in percent of the total number of papers in each time period.

The SPIRES programming language is designed to find single papers and thus it requires a bit of ingenuity to read out the entire database. The “date added” (**date-added**) command makes it possible to select the entire database, simply by choosing the entire lifespan of SPIRES, as the dates in between which the paper one is looking for is added. In practice, we divided the database into 5 time periods of approximately equal byte-size, since this drastically reduces the search time. Furthermore, it is necessary to set the system up to do big searches using the command **set big search**.

A.1.2 Retrieving the Data

The next task one has consider, is to set up SPIRES to output citation information. One way to do this is via the output format **citation**, which is set by typing **set for citation**. Then all there is left to do is to type the **type** command to output the data. The data is output on the telnet prompt, which I saved using **putty.exe**’s logging function.

The data on the 30 different searches is summarized in Table A.1, and an example of the data acquisition can be found in Appendix A.1. Finally, the raw data is cleaned using simple PERL programs. A generic example of one of these can be found in Appendix A.3.

A.1.3 Problems Regarding the Quality of the Data

It is clear from Table A.1 that the SPIRES system is less than perfect. Applying the search modifiers leaves out a number of papers that are not included in any of the sub categories. There is a significant discrepancy between the sum of the papers in the subgroups and the total number of papers reported by SPIRES – including the entire database, 19.5 percent of the entries in the database do not belong to any of the five sub categories.

Random samples of these papers indicate that these extra records refer to papers that have later been either removed or relocated. On account of this, we *define* the SPIRES hep database as the sum of papers in the sub categories, thus reducing the database to 403.805 papers.

| | Cited | Uncited | N/A | Sum |
|-----------------|---------|---------|---------|---------|
| Theory | 113,818 | 46,128 | 50,424 | 210,360 |
| Phenomenology | 53,811 | 14,738 | 15,516 | 84,065 |
| Experiment | 20,890 | 7,637 | 26,236 | 54,763 |
| Instrumentation | 7,523 | 12,114 | 21,748 | 41,385 |
| Review | 3,813 | 1,096 | 5,813 | 10,871 |
| Total | 200004 | 81713 | 119,837 | 401,444 |

Table A.2: The final data. This table contains the data after the PERL parsing. Note that a few of the papers (2365) ‘disappear’ in the process of parsing the data, cf. the final column of Table A.1. The ‘N/A’ column refers to papers that are non-journal papers for which no citation information is available i.e. conference proceedings and the like. The ‘Total’ row is obtained directly from the subfield data.

Another problem is that a small fraction of the data ($\approx 0.6\%$) is output in some way erroneously, such that the PERL script is unable to extract the citation information. The differences are displayed in the final columns of Tables A.1 and A.2.

It is also apparent from Table A.2 that a significant fraction of the papers are non-journal papers (conference proceedings, preprints, etc.), i.e. no citation information is available. Thus, we are left with a total of 281,717 papers (i.e. roughly 56% of the SPIRES database) for which both citation information and subfield designations are available. This data is what is referred to when we speak of the *SPIRES hep database*.

A.2 An Example of a Typical Run

The following is an example of how the data from the durham SPIRES hep mirror was collected. The example chosen is a search of Review papers added before 1980.

SunOS 5.6

```
login: lehrmann Password: Last login: Wed Aug 14 20:32:45 from
pci38.valkendorf
```

```
Sun Microsystems Inc. SunOS 5.6 Generic August 1997
```

```
> spires
```

```
Loading module: spirest
```

```
Loading completed: 868352 bytes.
```

```
-Welcome to SPIRES 00.07
```

```
-> set big sea 10000k
```

```
-> sel hep

-Command logging in effect for this subfile

-> set sea mod and ps R

-> find da before 1980

-Result: 1393 DOCUMENTS

->set for citation

-> type
```

Warning: the citation search should be used and interpreted with great care. At present, the source for the citation list in the HEP database is only the preprints received by the SLAC Library, and not the (unpreprinted) journal articles. Citations of a paper during the months it was circulated as a preprint may also be lost, because only references to journal articles and e-print-archives papers are indexed. Still, the citation index in HEP (SPIRES-SLAC) is formed from an impressive number of sources. For example, in 1994, the citation lists were collected from 10,000 preprints.

- 1) Jean-Loup Gervais, A. Neveu, INTRODUCTION.
Published in Phys.Rept.23:240-244,1976.

Cited 28 times. To get the listing use: FIN C PRPLC,C23,240

...

A.3 A Perl Program

This appendix contains an example of one of the simple PERL programs used to clean up the data outputted from spires, cf. Appendices A.1 and A.2.

```
open INPUT,"<review.spires";

open OUTPUT,">review.dat";

while(<INPUT>)

{
  next unless m/Cited/;
  split;
```

```
print OUTPUT $_[1], "\n";
}
```

```
close OUTPUT;
```

This particular piece of code, when executed, reads out the number of citations of each cited paper in the file containing the raw information on the Review subgroup.

A.4 Longitudinal Cleaning

The longitudinal data is obtained from the data acquired in appendix A.1, using a slightly more complicated PERL script to parse the data²:

```
#!/usr/local/bin/perl -w

$#ARGV==1 || die "usage: $0 infile outfile\n";

#initiations of database

%authors = ();

%articles = ();

$article_no = 0;

%carticle = ();
$carticle = ""; $ctitle = ""; $fnames = ""; $surnames = "";
$no_more_authors = 0; @array = ();

open (INFILE, "<$ARGV[0]>") || die("Can't open $ARGV[0]: $!");
@lines = <INFILE>;

foreach $line (@lines) {
    # first article
    if ($article_no == 0) {
        if ($line =~ /\d+(\)|A-Z)/) {
            # initiations
            $carticle = $line;
            $ctitle = "";
            $article_no++;
            $no_more_authors = 0;
        }
    }
    elsif ($line =~ /\d+(\)|A-Z)/) {
        # print "Match $line\n";
        $carticle =~ s/[ ]*$//;
        @array = split(/\\n|/, $carticle);

        $carticle{'authors'} = ();
        $carticle{'title'} = "";

        # # if you had a unique number. # $array[0] =~
        /(\d+)(\)|A-Z)/; # $cnumber = $1;

        for ($i=0; $i<=$#array; $i++) {
            if ($array[$i] =~ /[A-Z]{2,}/) {
                $no_more_authors = 1;
            }
            if ($array[$i] =~ /([A-Z]+[\\w\\s\\.\\-]+)\\s([\\w+|\\-]+[a-z]+)/ && !$no_more_authors) {
                $fnames = $1;
            }
        }
    }
}
```

²I am thankful to Jens Munk for his hours of help on writing this rather complicated bit of code.

```

$surnames = $2;
$fnames = ~ s/[a-z\.\s\-\]//g;

push @{$carticle{'authors'}}, "$surnames $fnames";
}
elseif ($array[$i] =~ /[([A-Z]+[\w\s\.\-]+)\s([([w-]+l)\set\s\al\.$/ && !$no_more_authors) {
$fnames = $1;
$surnames = $2;
$fnames = ~ s/[a-z\.\s\-\]//g;
push @{$carticle{'authors'}}, "$surnames $fnames";
}
if ($array[$i] =~ /Cited\s(\d+)|(N\A)|(This\swork)/) {
# |( no citatio) matched af $2
if ($1) {
$carticle{'quotations'} = $1;
}
elseif ($2) {
$carticle{'quotations'} = 0;
}
else {
$carticle{'quotations'} = '';
}
}
}
if ($carticle =~ /[([A-Z]{1}[A-Z\s,\:\+\-\=0-9]+\.)]/) {
# ret 2 til 1
$title = "$1";
$title = ~ s/\s{2,}/ /;
$carticle{'title'} = $title;
}
else {
$carticle{'title'} = 'Problems identifying title';
}
}

# actions - create database
foreach $author (@{$carticle{'authors'}}) {
if ($authors{"$author"}{'quotations'}) {
push @{$authors{"$author"}{'quotations'}}, $carticle{'quotations'};
push @{$authors{"$author"}{'articles'}}, $carticle{'title'};
}
else {
$authors{"$author"}{'quotations'} = [$carticle{'quotations'}];
$authors{"$author"}{'articles'} = [$carticle{'title'}];
}
}
if (!$articles{"$carticle{'title'}"}) {
$articles{"$carticle{'title'}"}{'authors'} = "@{$carticle{'authors'}}";
}

# initiations
$no_more_authors = 0;
$carticle = "$line";
$title = "";
$article_no++;
}
else {
$carticle .= $line;
}
if ($article_no == 20) {
# exit;
}
} close(INFILE);

# Here you can modify the output written to the output file
# e.g. if you are a windows person, you can make a csv file.

```

```
open (OUTFILE, ">$ARGV[1]") || die("Can't open $ARGV[1]: $!");
foreach $author (sort (keys %authors)) {
    print OUTFILE qq($author @{$authors{"$author"}{'quotations'}}\n);
} print OUTFILE "End authors\n";

# There are still titles that are matched wrong (only names). You
# can correct it if you feel like #print OUTFILE
"-----\n";
foreach $article (sort (keys %articles)) {
    # print OUTFILE qq($article
    $articles{"$article"}{'authors'}\n); #} close (OUTFILE);
```

This script collects papers, written by the same author into strings, that contain the year the paper was written and the number of citations acquired by the particular paper. It is possible to specify whether one wants to sort authors after last name only, or after initials *and* last name.

B.1 Residual Analysis

The Q-Statistic, defined in Section 5.3.5 is defined as the sum of squares of the residuals:

$$Q = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}). \quad (\text{B.1})$$

This represents the sum of squares of the distance of $\mathbf{y} - \hat{\mathbf{y}}$ from the k -dimensional space that the PCA model defines. To obtain an upper limit for Q , let:

$$\theta_\alpha = \sum_{i=k+1}^p \lambda_i^\alpha, \quad \alpha = 1, 2, 3, \quad (\text{B.2})$$

and

$$h_0 = 1 - \frac{2\theta_1\theta_2}{3\theta_2^2}. \quad (\text{B.3})$$

Then the quantity

$$c = \theta_1 \frac{\left[\left(\frac{Q}{\theta_1} \right) - \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} - 1 \right]}{h_0 \sqrt{2\theta_2}} \quad (\text{B.4})$$

is approximately normally distributed with zero mean and unit variance. On the contrary, the critical value for Q is

$$Q_\alpha = \theta_1 \left(\frac{c_\alpha h_0 \sqrt{2\theta_2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{1/h_0}, \quad (\text{B.5})$$

where c_α is the normal deviate cutting off an area of α under the the upper tail of the distribution, if h_0 is positive, and under the lower tail, if it is negative. This distribution holds whether or not all of the significant components and even if non-significant ones are employed. Values of Q that are higher than Q_α (for a given α that we choose) are an indication that a data-vector cannot be adequately represented by a k component model. This appendix draws heavily on [67].

B.2 Special Functions

The generalized hypergeometric function is given by a Hypergeometric Series, *i.e.*, a series for which the ratio of successive terms can be written

$$\frac{a_{k+1}}{a_k} = \frac{P(k)}{Q(k)} = \frac{(k+a_1)(k+a_2)\cdots(k+a_p)}{(k+b_1)(k+b_2)\cdots(k+b_q)(k+1)}x. \quad (\text{B.6})$$

(The factor of $(k+1)$ in the denominator is present for historical reasons.) The resulting generalized hypergeometric function is written:

$$\sum_{k=0}^{\infty} a_k x^k = {}_pF_q \left[\begin{matrix} a_1, a_2, \dots, a_p \\ b_1, b_2, \dots, b_q \end{matrix} ; x \right] \quad (\text{B.7})$$

$$= \sum_{k=0}^{\infty} \frac{(a_1)_k (a_2)_k \cdots (a_p)_k}{(b_1)_k (b_2)_k \cdots (b_q)_k} \frac{x^k}{k!}, \quad (\text{B.8})$$

where $(a)_k$ is the Pochhammer Symbol,

$$(a)_k \equiv \frac{\Gamma(a+k)}{\Gamma(a)} = a(a+1)\cdots(a+k-1). \quad (\text{B.9})$$

If the argument, $x = 1$, then the function is abbreviated to

$${}_pF_q \left[\begin{matrix} a_1, a_2, \dots, a_p \\ b_1, b_2, \dots, b_q \end{matrix} \right] \equiv {}_pF_q \left[\begin{matrix} a_1, a_2, \dots, a_p \\ b_1, b_2, \dots, b_q \end{matrix} ; x \right]. \quad (\text{B.10})$$

${}_2F_1(a, b; c; z)$ is ‘the’ Hypergeometric Function, and ${}_1F_1(a; b; z) \equiv M(z)$ is the Confluent Hypergeometric Function. A plethora of information on the hypergeometric function and other special functions can be found on the www, *cf.* [62]. My main source on identities between special functions and non-trivial integrals has been MathWorld and the excellent program *Mathematica*.

- [1] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.
- [2] A. Vlado. Pajek: Program for large network analysis. <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>, 2003.
- [3] S. Lehmann, A. D. Jackson, and B. E. Lautrup. Citation distributions in high energy physics. *arXiv:physics/0211010*, 2002.
- [4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–12, 1999.
- [5] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of $1/f$ noise. *Physical Review Letters*, 59(4):381–384, 1987.
- [6] J. M. Yeomans. *Statistical Mechanics of Phase Transitions*. Clarendon Press, Oxford, 1997.
- [7] A. Bird. *Philosophy of Science*. Fundamentals of Philosophy. UCL Press, London, 1998.
- [8] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–7, 1959.
- [9] D. J. Watts and S. H. Strogatz. Collective dynamics of “small-world” networks. *Nature*, 393:440–2, 1998.
- [10] M. E. J. Newman. Ego-centered networks and the ripple effect—or—Why all your friends are weird. *Social Networks*, 25:83–95, 2003.
- [11] M. E. J. Newman, C. Moore, and D. J. Watts. Mean-field solution of the small-world network model. *Physical Review Letters*, 84(14):3201–4, 2000.
- [12] M. E. J. Newman. Models of the small world. *Journal of Statistical Physics*, 101(3/4):819–41, 2000.
- [13] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74:47–97, 2002.

- [14] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances of Physics*, 51:1079–1187, 2002.
- [15] T. J. Fararo and M. Sunshine. *A Study of a Biased Friendship*. Syracuse University Press, Syracuse, NY, 1964.
- [16] J. F. Padgett and C. K. Ansell. Robust action and the rise of the Medici, 1400-1434. *American Journal of Sociology*, 98:1259–1319, 1993.
- [17] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.
- [18] A. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world-wide web. *Nature*, 401:130–1, 1999.
- [19] R. Kumar *et al.* In *Proceedings of the 9th ACM Symposium on Principles of Database Systems*, 1999.
- [20] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web, www9. *Computer Networks*, 33(1-6):309–20, 2000.
- [21] M. E. J. Newman, S. Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66:035101–1–4, 2002.
- [22] M. Faloutsos, P. Faloutsos, C. Faloutsos. On power-law relationships of the internet topology. *Computer Communications Review*, 29:251–262, 1999.
- [23] H. Jeong, S.-H Yook and A.-L. Barabási. Modeling the internet’s large-scale topology. preprint: cond-mat/0107417, 2001.
- [24] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–2, 2001.
- [25] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118:1–17, 2001.
- [26] I. Farkas *et al.* Networks in life: Scaling properties and eigenvalue spectra. *Physica A*, 314:25–34, 2002.
- [27] A. L. Barabási *et al.* Evolution of the social network of scientific collaborations. *Physica*, 311:590–614, 2002.
- [28] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651, 2000.
- [29] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.
- [30] W. Aiello, F. Chung, and L. Lu. Random evolution in massive graphs. In James Abello *et al.*, editor, *Handbook on Massive Data Sets*. Kluwer Academic Press, Handbook on Massive Data Sets.

- [31] S. H. Strogatz. Exploring complex networks. *Nature*, 410:263–76, 2001.
- [32] M. Barthélemy and L.A.N. Amaral. Small-world networks: Evidence for crossover picture. *Physical Review Letters*, 82(15):3180–3, 1999.
- [33] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60(6):7332–42, 1999.
- [34] M. A. de Menezes, C. F. Moukarzel, and T. J. P. Penna. First-order transitions in small-world networks. *Europhysics Letters*, 50(5):574–79, 2000.
- [35] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173, 1999.
- [36] C. Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678–82, 2000.
- [37] R. Cohen *et al.* Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626–8, 2000.
- [38] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–71, 2000.
- [39] A. Barrat and M. Weigt. On the properties of small-world network models. *European Physics Journal B*, 13:547–60, 2000.
- [40] C. F. Moukarzel. Spreading and shortest paths in systems with sparse long-range connections. *Physical Review E*, 60:R6263–66, 1999.
- [41] S. N. Dorogovtsev and J.F.F. Mendes. Exactly solvable small-world network. *Europhysics Letters*, 50(1):1–7, 2000.
- [42] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629–32, 2000.
- [43] S. N. Dorogovtsev, J.F.F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(21):4633–6, 2000.
- [44] R.V. Kulkarni, E. Almaas, and D. Stroud. Exact results and scaling properties of small-world networks. *Physical Review E*, 61:4268–71, 2000.
- [45] J. M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.
- [46] R. Kumar *et al.* Stochastic models for the web graph. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pages 57–65, New York, 2000.
- [47] Science Citation Index. <http://www.isinet.com/isi/>. This website contains a great deal of historical information on the history of counting scientific papers.
- [48] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences. USA*, 98(2):404–9, 2001.

- [49] M. E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64:016131, 2001.
- [50] M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64:016132, 2001.
- [51] A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, pages 317–323, 1926.
- [52] W. Shockley. On the statistics of individual variations of productivity in research laboratories. *Proceedings of the Institute of Radio Engineers*, 45:279 – 290, 1957.
- [53] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *European Physical Journal B* 2, pages 525 – 539, 1998.
- [54] S. Redner. How popular is your paper? An empirical study of the citation distribution. *European Physics Journal B*, 4:131–4, 1998.
- [55] C. Tsallis and M. P. de Albuquerque. Are citations of scientific papers a case of nonextensivity. *European Physics Journal B*, 13:777–80, 1999.
- [56] S. Bilke and C. Peterson. Topological properties of citation and metabolic networks. *Physical Review E*, 64:036106, 2001.
- [57] H. O'Connell. Physicists thriving with paperless publishing. *High Energy Physics Libraries Webzine*, March 2002. <http://library.cern.ch/HEPLW/6/papers/3/>.
- [58] S. Cole and J. R. Cole. Scientific output and recognition; a study in the operation of the reward system in science. *American Sociology Review*, 32:377–90, 1967.
- [59] E. Garfield. Citation indexing for studying science. *Nature*, 227:669–671, 1970.
- [60] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, 1949.
- [61] A. Einstein. Zur elektrodynamik bewegter körper. *Annalen der Physik*, 17:891, 1905.
- [62] Eric Weisstein. MathWorld. Online, 2003. <http://mathworld.wolfram.com>.
- [63] H. Quin. Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations. *arXiv:math-ph/0007010*, 2001.
- [64] J. Vlachý. Citation histories of scientific publications. The data sources. *Scientometrics*, 7:505–528, 1984.
- [65] N. H. Timm. *Applied Multivariate Analysis*. Springer Texts in Statistics. Springer Verlag, Heidelberg, 2001.
- [66] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley series in probability and mathematical statistics. Wiley, New York, second edition, 1984.
- [67] J. E. Jackson. *A User's Guide to Principal Components*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1991.

- [68] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:414–441,498–520, 1933.
- [69] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.
- [70] H. A. Simon. *Models of Man*. Wiley, New York, 1957.
- [71] M. H. Ernst. Kinetics of clustering in irreversible aggregation. In L. Pietronero and E. Tosatti, editors, *Fractals in physics*. Elsevier, Amsterdam, 1986.
- [72] A. J. Bray. Theory of phase-ordering kinetics. *Advances of Physics*, 43:357–459, 1994.
- [73] A. Pimpinelli and J. Villain. *Physics of Crystal Growth*. Cambridge University Press, Cambridge, 1998.
- [74] S. N. Dorogovtsev and J. F. F. Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, 63:056125, 2001.
- [75] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63:066123, 2001.
- [76] A.-L. Barabasi, *et al.* Evolution of the social network of scientific collaborations. *Physica A*, 311, 2002.
- [77] H. Jeong, Z. Néda, and A.-L. Barabási. Measuring preferential attachment of evolving networks. *arXiv:cond-mat/0104131*, 2001.
- [78] M.E.J. Newman. Clustering and preferential attachment in growing networks. *arXiv:cond-mat/010409*, 2001.
- [79] M. V. Simkin and V. P. Roychowdhury. Read before you cite! *arXiv:cond-mat/0212043*, 2002.
- [80] B. E. Lautrup. Dynamics of excellence. Private Communications, 1999.
- [81] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Physical Review E*, 62:1842, 2000.
- [82] *SPIRES Manual*. <http://www.slac.stanford.edu/~guertin/SPIRES.HTML>.