Popularity Distributions with Social Influence

Comparing YouTube with Hollywood

V.A. Traag

Department of Sociology
Faculty of Social and Behavioural Sciences
University of Amsterdam
Email: V.A.Traag@uva.nl

Thesis submitted for the degree
of Master in Sociology

University of Amsterdam

May 9, 2008

Supervisors dr. J.P. Bruggeman
dr. D.C. Gijswijt

"Popularity is the one insult I have never suffered."

— *Oscar Wilde*

Abstract

This thesis considers a model for popularity based on a 'rich-get-richer' effect. Basically, popular items tend to become increasingly popular. However, quality of items also plays a role in popularity. The model thus incorporates both a 'rich-get-richer' as well as a 'good-get-richer' effect. The balance between these two effects is interpreted as the amount of social influence. Formal analysis of the model suggests that the distribution of popularity becomes more unequal and more uncertain with rising social influence. Higher quality has a dual role: it results in a higher average popularity, but increases the uncertainty as well. The model is tested against data from the YouTube market and the Hollywood movie industry. Comparing the results for these two markets suggests that social influence is higher for on-line markets than for traditional markets. When markets go on-line, producers should be prepared to take the increase in risk into account. Some books might break all records, while others remain on the shelf, and it becomes harder to predict which books that will be.

Keywords: Social Influence, Popularity Distribution, Power law Distribution, Cultural Market, YouTube, Hollywood, On-line Market, Traditional Market

Contents

Preface

Sociology and mathematics is a combination rarely seen. Recently, however, an increasing number of scientists from other fields, such as biology and physics, are using their mathematical modelling experience to conduct sociological investigations, in particular in the field of social networks (Barabási 2003). There are also some sociologist who use mathematical modelling techniques (Edling 2002). Yet, mathematics is not a common phenomena in the social sciences.

Mathematics is often seen as 'too' abstract: how can any social phenomenon be captured in equations? It does not seem possible to capture things such as nationalism or ethnicity in mathematical symbols, especially the emotions aroused by such concepts. To understand what nationalism does to a person, you don't need mathematics, you need a detailed in-depth account of it. Or maybe, you have to experience it for yourself.

The power of mathematics lies not with emotions or descriptions in that way. Mathematics does not make you aware of what nationalism does to a person, nor does it tell you what is important and what not. What processes are important for the emergence of nationalism will not be revealed by mathematics. Whether to include economical developments, technological advancements or changing ideologies to explain nationalism cannot be decided by mathematics. However, mathematics can provide other insights.

In sociology hypotheses are not always deduced clearly from theory, since most deductions in sociology are done verbally. Verbal reasoning leaves the deduced hypotheses ambiguous and open for debate. The power of mathematics lies with the deduction of hypotheses from theory. Seeing how various principles, causes and effects interplay to produce outcomes is difficult, especially when multiple causes and feedback loops are present. Through mathematical analysis this caveat of verbal reasoning can be addressed. It provides a grounded, well established framework wherein theories can be modelled carefully and thoroughly. In sum, mathematics is no substitute for theoretical reasoning as verbal reasoning is no substitute for mathematical analysis. Several authors, such as McElreath and Boyd (2007) and Turchin (2003) have made similar arguments.

Let me illustrate this point by considering the geopolitical theory of Randall Collins. He predicted in 1978 the collapse of the Soviet Union, based on a geopolitical model of state-breakdown (Collins 1999). His model can be summarised as following. A good geopolitical position ('marchland' advantage) leads to a higher success rate in war. A higher success rate in war leads to an increase in

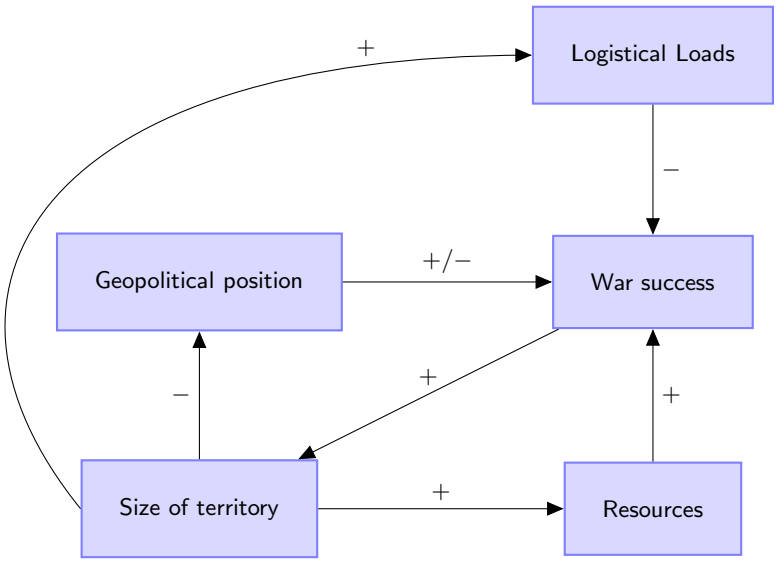

Figure 1: Illustration of the geopolitical model of state breakdown of Collins. After Collins (1999:figure 1).

territorial size. The increase in size has several consequences. It increases the logistical loads (the resources spent to control the empire), changes its geopolitical position (usually worsens it, as the state now comes in contact with more powerful neighbours) but also increases the resources with which wars can be won. These resources increase the success rate in war again, but the logistical loads decrease the success rate. This model is illustrated in figure 1.

Although this model is stated quite clearly, it is rather difficult to draw clear conclusions from it. When will a state break down? Does the geopolitical model of state breakdown predict cyclic Great Powers? Or will it lead towards a stationary configuration of states? Maybe it will predict the emergence of a world state? The answers to these questions are not easy to infer verbally from the model.

More specifically, how do the various variables interact? What effect is stronger than the other? Can we pick any one state at random, and determine whether it is about to expand, collapse or remain stable? More problematic: what exactly constitutes a falsification of the model? Can we always argue that a specific case is indeed 'explained' by the model? And how would we compare it to rival models of state-breakdown?

For example, Goldstone (1993) focuses on the strain that population growth puts on the state, while Skocpol (1979) suggests that defeat in war in pivotal in state breakdowns. In the French Revolution for example, we see various suggested principles at work. There was the Seven Year's War (1756–1763) in which France experienced serious territorial losses and incurred fiscal problems. But population growth also led to inflating wheat prices which resulted

in widespread famine and food revolts. What model gives a more realistic picture? Using post-hoc arguments, it is difficult to asses in any specific way, which model better approximates reality. If we only provide arguments post-hoc, we will only be able to try and fit history into the model.

Of course it should be the other way round. We should try to fit the model to history; not to fit history to the model. We should not try to 'explain' the English Civil War of 1642–1651 or 'explain' the collapse of the Soviet Union in 1990 with hindsight. Besides the fact that we may ignore many instances at which states did not breakdown in this way—more formally known as 'sampling on the dependent variable'—hindsight blurs clear vision.

The question is: could we have predicted the English Civil War starting in 1642, using only data from before 1642? We should put on a historical veil of ignorance so to speak. Blind to whatever may happen after a certain date, could we still make correct predictions?

Given these problems more specific to historical sociology, one rather general problem remains: in what way can we infer actual predictions from our model? How are assumptions of the processes that take place translated into consequences? It is here that mathematics comes into play, so that we may see the consequences more vividly, and thus draw conclusions more rigorously.

This is largely in congruence with Collins' own approach:

> "Can successful historical predictions be made? Obviously they can. But it is important to distinguish between a sociological prediction and a guess or wishful thinking. A valid prediction requires two things. First, the prediction must be based on a theory that explains the conditions under which various things happen or do not happen—that is, a model that culminates in if-then statements. This is a more stringent standard of theory than what sociologists generally mean by the term. It is not a category scheme, or a metatheory, or even a process model, which lacks observable if-then consequences. Second there must be empirical information about the starting points, the conditions at the beginning of the if-then statement." (Collins 1999:57)

And Collins also recognises the need for a more explicit formulation of theories:

> "Theory usually enunciates general tendencies: for example, rulers require legitimacy, conflict produces solidarity, a military-industrial complex promotes war. Each proposition stands alone as a ceteris paribus generalization. Deductions about the behavior of the systems described by such statements are often far from obvious for a variety of reasons. Most important are multiple causes and feedback processes among them. Even in very simple theoretical models, there can be unexpected outcomes. Positive loops accelerate basic processes and bring some of them to ceilings at which they rest; negative feedback provides counteracting forces, which sometimes lead

> to a stable equilibrium, sometimes to oscillation, and sometimes to chaos.
>
> When a theory is formulated verbally, such as Weber's or Simmel's classic statements about conflict, these alternatives are left open. We do not know what is implied in a theory as long as it is left on the level of separate general principles and is abstracted out of time. One way to overcome this ignorance is to perform experimental research on such theories by means of computer simulation. This activity is a discovery-making process in the sense that one does not really understand what the theory is saying about the world until one has experimented with it as a dynamic model" (Collins 1999:239)

Here, Collins argues for a simulation of various interacting processes, which he thereafter indeed undertakes. Gilbert and Troitzsch (2005) are other proponents of this 'social simulation'. This is a solid step forward: it becomes more apparent what the dynamics of the suggested model are. It has one drawback however: understanding is limited to what initial conditions have been experimented with. Although computing power has grown tremendously over the past few decades, it is still limited. Hence, not all initial conditions can be tried. Moreover, simulations are largely written in a non-standardised form, and depend on specific implementations. We cannot 'read' simulations as we can mathematics. It is therefore rather difficult to replicate simulations and results from other scientists.

Through mathematical analysis, more definite results can be obtained. We do not have to 'try' every single initial condition, since mathematical deductions provide us with more clear-cut answers. The interaction and consequences of various initial conditions can be made quite clear with mathematical analysis. Furthermore, analytical results can be verified easily, since they can be mathematically proven. Other scientists can therefore always replicate the results and check the formal deductions.

Of course, some models are of such inherent complexity (including explicit spatial dimensions for example), that simulation is the only way we can understand its dynamics. Simulation provides a mechanism by which to understand these more complex dynamics. But it should mainly be limited to exploratory research or for intractable problems. Mathematical formalisation has clear advantages in comparison with simulation.

One such formalisation of Collins' model is given by Turchin (2003). He also promotes a more formal approach to historical sociology:

> "In general, non-linear dynamical systems have a much wider spectrum of behaviors than could be imagined by informal reasoning. (...) Thus, a formal mathematical apparatus is indispensable when we wish to rigorously connect the set of assumptions about the system to predictions about its dynamic behavior." (Turchin 2003:4)

We will outline the approach Turchin used to formalise the theory of Collins. His approach is that of *dynamical systems*, which is a quite well-defined concept.

Within the body of sociological literature, this concept is usually interpreted as some sort of interaction between various variables, which produces somehow certain dynamics. But this does not cover the concept.

The concept of dynamics (usually) entail some form of temporal behaviour. That is, over time resources are accumulated, populations grow, and territory increases. The growth (or decline) of these variables are dependent on the actual value of the variable. For example, suppose we have state A with a population of $1,000$ people and a state B with a population of 1 million people. With a birth rate of 5% per year, state A would have 50 new-borns, and state B would have $50,000$ new-borns. So, the growth of a population (in absolute terms) is dependent on the size of a population.

More formally, let the size of the population be given by P. The rate of change of P over time will then be denote by $\dot{P}$. A dot over a variable thus signifies the rate of change of the variable, not the actual value of the variable. We can then state that a population grows linearly with its size, or

$$\dot{P} = rP.$$

It says here, that the rate of change of the population size is proportional to the size of the population. The rate at which the population grows is denoted by r. We might call r the growth rate. Such an equation is called a *differential equation*.

Now lets say that we have a number of resources Q. As the population grows, resources become more depleted, so Q will decline. On the other hand, the population will grow faster if resources are plenty, but slows down if they become depleted. If we assume that resources are added at a constant base, we might arrive at the following idea

$$\begin{aligned} \dot{P} &= aQP - bP, \\ \dot{Q} &= c - dQP. \end{aligned}$$

Here, the population P grows relative to the number of resources and the population at some rate a and shrinks at a rate b proportional to the population size. The resources Q grow at a base rate of c, and are consumed by the population at a proportional rate d. There are now two differential equations which are *coupled*. With coupled, I mean that P and Q both influence each other.

We might wonder whether there is point where the system is in an equilibrium. If it is in equilibrium, we do not expect the population numbers or the resources to change. In other words, we expect $\dot{P} = 0$ and $\dot{Q} = 0$. Assuming the first, we obtain

$$(aQ - b)P = 0,$$

suggesting that either the population size is zero, $P = 0$, or that $(aQ - b) = 0$. Writing out the latter option gives $Q = b/a$. If we solve $\dot{Q} = 0$ we obtain

$$c = dQP.$$

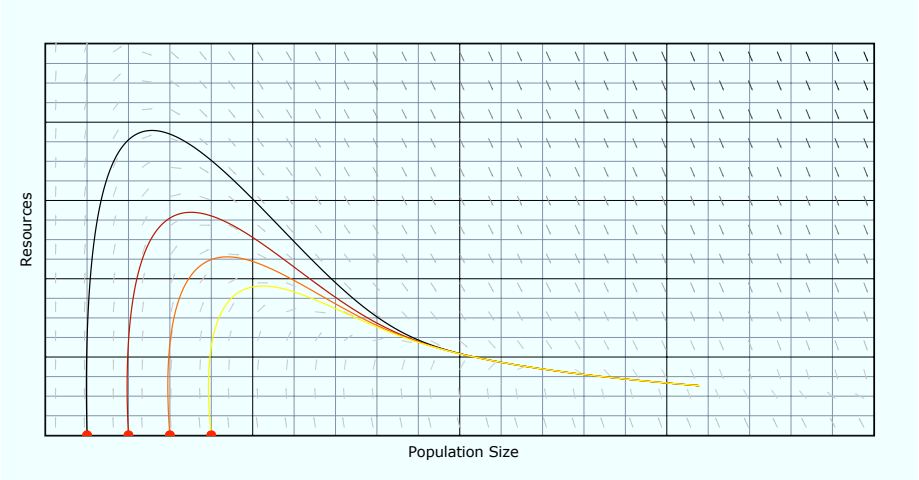

Figure 2: Illustration of the population and resource dynamics. The red dots represent the initial conditions. At first, population growth is rather slow, but resources accumulate over time, and the population starts to grow. As the population grows, the resources decrease, and the population growth slows down. In the end, resources are consumed by the population at a rate equal to the rate at which the resources replenish. An indication of how resources and population size change for different values is given by the small arrows. Darker arrows mean a larger change.

If we assume that $P \neq 0$ we can substitute $Q = b/a$, at which point the population is in equilibrium. We then obtain

$$\begin{aligned} c &= d\frac{b}{a}P, \\ P &= \frac{ca}{db}. \end{aligned}$$

So if $P = ca/db$ and $Q = b/a$ the system is in equilibrium. Since the resources grow independently of the size of the population, it will always replenish. Hence, if we start of with some non-zero population, the equilibrium is always reached. This process is illustrated in figure 2

We can now define the concept of a dynamical system more specifically. A dynamical system[1] consists of one or more (coupled) differential equations. Should dynamical systems be of interest to the reader, Strogatz (2001) is a good place to start.

Turchin uses this approach of dynamical systems to formalise the model of Collins. He thus creates a differential equation, for which the basic relations are given by the model of Collins as portrayed in figure 1. Let us review the formalisation by Turchin briefly[2].

[1] This definition is only a rather informal one; it is somewhat looser than definitions usually provided in mathematical textbooks.

[2] Should it be of interest to the reader, the model of Collins is discussed on pages 16–28 in Turchin (2003).

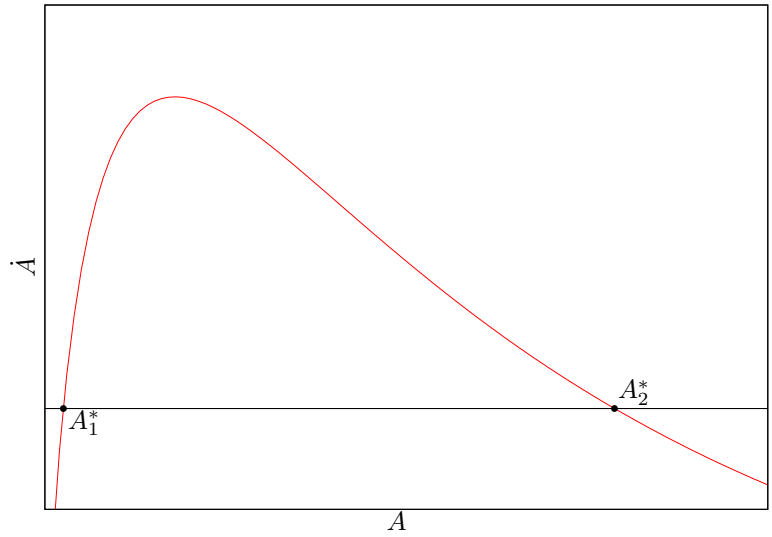

Figure 3: Rate of change $\dot{A}$ of the territory A for the model with logistical loads incorporated. It can be seen that there are two equilibria, A_1^* and A_2^*, of which only the latter is stable. Hence, for any initial condition between A_1^* and A_2^* the territory will grow until, restrained by logistical loads, it will slow down approaching the equilibrium A_2^*.

Turchin defines A to be the area, or territory of a state. The territory is only affected by the succes in war W. Turchin assumes the relationship to be linear, and arrives at $\dot{A} = c_1 W$, where c_1 is some constant translating succes in war into territory. Resources R are simply proportional to territory, or $R = c_2 A$. The succes in war is determined by the state's own resources relative to the resources of other states. Thus $W = c_3 R - c_4$, where c_4 is the constant indicating the constant military force of opponents. Putting all this together gives us

$$\dot{A} = cA - b,$$

where $c = c_1 c_2 c_3$ and $b = c_1 c_4$. This is a simple linear model, which has an equilibrium at $A^* = b/c$. Below this equilibrium A^*, the growth of the state is negative, and hence it will decline. Above the equilibrium, growth is positive, and hence the state will grow indefinitely. But the logistical loads have not yet been incorporated.

The logistical loads have a negative effect on success in war W, and thereby on the growth of the state. Turchin assumes that the projection of power declines exponentially with the distance r from the center, or e^{-r/c_5}. Since the distance is related to the area as $r \sim A^2$, this can be written as $e^{-\sqrt{A}/h}$. Since power is proportional to the territory, we can incorporate this as:

$$\dot{A} = cAe^{-\sqrt{A}/h} - b.$$

This model is illustrated in figure 3. It can be seen that two equilibria A_1^* and A_2^* exist. For any initial condition below A_1^*, the rate of change is negative, and the state will decline. For any initial condition above A_1^*, the state will grow until it slows down and reaches the equilibrium A_2^*. Should somehow the state grow above that equilibrium, the logistical loads make sure the state declines again, until it reaches the equilibrium A_2^*. So equilibirum A_1^* is unstable and equilibrium A_2^* is stable.

The formalisation gives rise to some conclusions. Firstly, Collins' model does not seem to give an argument as to why states collapse. The negative feedback from logistical loads does not produce a full collapse, but only sets an upper limit to the size of a state. Secondly, the model leads to a stable equilibrium. The model predicts a rather smooth trajectory towards an equilibrium. History suggests that empires rise and fall however. Turchin shows that some adaptions of the specific implementation of Collins' postulates do not change the essence of the conclusion: the model is incapable of generating cyclic behaviour, which is not in congruence with the rise and fall of empires.

Turchin shows that other models—without relying on exogenous factors—are capable of generating this behaviour. It allows for both state failures as well as for boom-burst cycles of territory size. It is based somewhat on the work of Goldstone (1993). The basic idea is this: populations prosper in peaceful times, thus increasing the population size. As the state becomes more populous, so does the elite, which gives rise to fiercer competition, and strains the state fiscally. The competition and fiscal problems lead to conflict, which reduces the population, which induces the beginning of another cycle. The actual model considered is more complicated, but it conveys the general idea.

This example clearly shows that mathematics may prove useful for the social sciences. The formalisation of the model illustrates that the conclusions that are drawn can be quite contrary to what may have been deduced verbally. The essential point of the model—that states collapse when the logistical burden becomes too heavy—does not follow from the postulates described by Collins. This does not mean that the processes discussed by Collins are invalid or incorrect, it just implies that these processes by themselves do not lead to the collapse of states.

To summarise, using mathematics it can be shown more clearly what conclusions can, and what conclusions cannot be drawn. Through formal analysis we may more vividly see the consequences of theoretical statements. Mathematics provides a clear, well established framework for modelling social phenomena. Because of the prominence of mathematics in other fields, many results and techniques are available. We should take advantage of that, and use mathematics to advance the social sciences.

Chapter 1

Introduction

1.1 General Idea

Why do some books, such as *Harry Potter* or *The Da Vinci Code* become popular, and why do others not? If someone publishes a magnificent book—evocative characters, imaginative prose and a quivering plot—what are the odds of it becoming popular? And how does that contrast with a book of lesser quality? Of course we might refer to the rich characters created by J.K. Rowling to explain her success. But can we do the same for Dan Brown?

More generally, what does popularity look like? We all hear on the radio, read in the newspaper or see on television how many copies of a book have been sold, or that a movie has broken all records. But we never hear of books that didn't become so popular. How many books and movies actually achieve some level of popularity? What are the chances of success?

These are questions which beg serious answers. Publishers in all sorts of markets have to deal with it daily. Why has almost everyone read *Harry Potter* or *The Da Vinci Code*? Is it because we all like them so much? Or is it just that everyone else has read it, and then: why shouldn't you? That is probably the key to popularity: social influence.

Humans exchange more information than any other animal. They gossip, they talk, they mimic and they imitate. This thesis discusses how mimicking preferences for books, songs or movies lead to popularity. For example, suppose someone, say Jane, has a preference for the song *School* by Supertramp. She tells al her friends about it, some of whom may or may not decide to go and listen to the song themselves. They, in their turn, tell their friends about the song, and the process so replicates itself.

Now let's evaluate this process a bit more in-depth. We start out with Jane, and assume she tells a number of people, say n. Let's assume, for sake of argument, that there is some fixed probability p that a friend of Jane, upon hearing from her, will go and listen to the song. So, after Jane has spoken to those n people, there will be about $1 + np$ people (including Jane herself) who

will have listened to the song.

Now, since this process replicates itself, another iteration of this process is done. We will assume that people not having listened to the song, will not mention it to their peers. Now each of the np persons who have in fact listened to the song will mention it to n other people. To keep it simple, friends of Jane have no common friends (besides Jane of course). So after this second round of music-gossip $1 + np + (np)^2$ people have listened to the song.

This simple model would predict that after t rounds of music-gossip, there would be

$$\sum_{i=0}^{t} (np)^i = \frac{1 - (np)^{t+1}}{1 - np}$$

people who have listened to the song. So if there is less than one new listener on average, or $np < 1$, there will only be a finite number of listeners. This is no surprise, since on average, less than one person will be persuaded by his friend. So, the process dies out if on average less than one person will be persuaded. If at least one person is persuaded on average ($np \geq 1$), the total number of listeners will tend to infinity.

We can look at the process another way. Suppose that k people have already heard the song. If each of them will tell n (distinct) friends, and each person has a probability p of actually going to listen to it, we will probably gain about knp listeners. Summing up, the probability of gaining additional listeners is linearly dependent on the number of listeners it already has. This effect can be summarised as the 'rich-get-richer' effect.

Of course, this process is not completely realistic. In the real world, a lot of people have common friends, and we might hear from a dozen people that this or that movie is "so great", or "so much fun". This changes the process. Maybe hearing it from two friends has even more impact than twice the impact of hearing it from only one. Moreover, hearing it from your best friend is quite something different than hearing it from some classmate (unless of course, that classmate happens to be your best friend).

These are indeed all kinds of complicating factors which deserve to be treated in their own regard. My basic assumption here is very simple: popularity increases the chances of becoming popular. In more complex systems, this assumption will not be very realistic. As a first approximation however, it might serve its purpose.

This central assumption allows us one big advantage. We don't need to know anything about the social network. All we care about is how much a song has been listened to, a book has been sold, or a movie has been watched. This is a big advantage because finding out about social networks in a robust manner is a painstaking, time consuming and expensive process.

This assumption is the basis for the model I will develop here. We might call this the 'compounded' approach, to offset it from an approach taking the social network into account. More specifically I try to model the phenomenon that (in the limit) an infinite number of people (large enough anyhow) can choose from an ever growing market. At each time step, m items from the market will be

picked by the population, I will speak of *votes* for those items. After this time step a new item is introduced into the market without any votes. In addition, each item has an associated 'intrinsic quality', which will partly be responsible for attracting votes.

Markets such as these are easily found on the internet, of which YouTube[1] is perhaps one of the most renowned examples. New videos are being uploaded to YouTube each day, and the number of viewers are practically unlimited. I have collected data on the number of views (votes in my terminology) for approximately 200,000 movies from YouTube, including the currently all time most viewed *Evolution of Dance*, which for reasons unknown to me, some people find hilarious. In addition I have recollected the data on the number of views about a week later in order to see how many additional views the movies attracted in the meantime.

The results of YouTube are compared with the results of mainstream Hollywood movies. Considering the amount of movies released each year, Hollywood can be seen as a growing market as well. One of the differences for Hollywood movies however, is that relatively much of the revenue is generated in the first few weeks. This might set cinema apart from the YouTube market, where movies from years ago still can be seen. Since millions of people go to the movies, the population might be considered virtually unlimited.

I will discuss some previous works on social influence in the remainder of this chapter. A formalisation of the model is given in chapter 2. I also discuss the relevance of my model for generating networks, which has drawn quite some attention over the past few years. In chapter 3 I will construct the methods for both estimating parameters and testing my model. The empirical analysis—based on data from the YouTube market and the Hollywood market—is done in chapter 4. Finally, a summary of my main conclusions is given in chapter 5. Additional material can be found in several appendices.

1.2 Cumulative Advantage

A 'rich-get-richer' effect was considered by Merton (1968). Merton describes in his work how scientists are credited with their research. He finds that well-known scientists are disproportionately more credited than their lesser known peers. He terms this the Matthew effect, after the Gospel According to St. Matthew:

> "For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath." (Merton 1968:3)

A more explicit formulation of such an effect was undertaken by Simon (1955) and De Solla Price (1965; 1976). Price talks more explicitly of a 'cumulative advantage', or as he puts is:

[1] http://www.youtube.com

> "The simplest expression of such a principle is to suppose that successes fall equally on the head of all *previous* successes, so the frequency of transitions from state n tot $n+1$ will be proportional to n."(De Solla Price 1976:294)

My central assumption is similar to the principles put forth by Merton, Simon and Price. Popular items tend to become even more popular. The difference however is that I do not assume that only popularity plays a role, but that intrinsic quality also plays its part, which will become more apparent when I discuss the model in chapter 2. The model of Price is a direct predecessor of the model considered by Barabasi and Albert (1999), but they term the cumulative advantage effect 'preferential attachment'. I discuss that model briefly in section 2.4.1.

The terms 'Matthew effect', 'rich-get-richer', 'cumulative advantage', 'preferential attachment' all signify the same thing: popular items tend to become more popular. My approach differs from previous models however, because of the social influence parameter and inherent quality. It is not only popular items that tend to become more popular: good items also tend to become more popular.

We may wonder however, why there is such an effect. Why do we imitate peers? Is there some innate drive to imitate others? And if so, why do we have such a drive? An interesting evolutionary argument is made by Boyd and Richerson (1985), but I will not pursue this question further.

1.3 Social Influence on Networks

I try to model social influence from a compounded point of view. This contrasts with models that try to capture influence and opinion dynamics on a network. They assume, as I do, that people influence each other through relationships. Instead of taking only some sort of average social influence, as I do with the compounded approach, they try to model this behaviour more explicitly.

In general they start off with some network, where vertices have some initial opinion or preference. Then, at each time step, the opinions and preferences of vertices are updated. Most of the time this points the research in another direction than that of mine. Preferences and opinions are usually formulated in juxtaposition. You cannot be both Republican and Democratic, or pro-life and pro-choice, or vote for Hillary Clinton as well as Barack Obama. So, the main focus is usually on investigating segregation of opinions, or looking at conditions by which a full consensus is reached.

This differs from the type of phenomena I wish to model. We usually don't have only one book in our bookcase, which we throw away if we buy a new book, or view only one film in our lifetime. Moreover, it is irreversible. If we have read a book, viewed a film or listened to some music, we cannot unread, unview or unlisten it. On the other hand, we can change our opinion, or our preference. We might change our taste in music, or our opinion on a movie, but not the fact that we have listened to it or viewed it.

1.3.1 Friedkin

Noah E. Friedkin has worked on the subject of social influence which resulted in numerous publications (Friedkin 1991; 1993; 1998; 1999; 2001). Instead of looking at dichotomous opinions or preferences, he analyses a model in which opinion is a continuous variable. He assumes some network of (weighted) relationships. Through these relationships, people are being influenced. The pressure being exerted by peers is a weighted combination of their opinions (Friedkin 2001:171, Friedkin 1991:1481, Friedkin 1999:860, Friedkin 1998:24).

More formally[2], let W be an $n \times n$ row stochastic matrix of the weighted relationships. So w_{ij} is the weight given to the relationships between vertex (actor) i and j and $\sum_j w_{ij} = 1$. Let A be a $n \times n$ diagonal matrix describing how much of the influence is determined by the social network. Some amount $1 - a_{ii}$ of the 'opinion formation' is done independently of others, while the rest a_{ii} is formed because of others. Friedkin analyses m opinions which are described by the $n \times m$ matrix Y. Then, given initial opinions Y_0 somehow formed exogenously the opinions are updated in each step as

$$Y_{t+1} = AWY_t + (I - A)Y_0,$$

where I is the identity matrix. This model attains an equilibrium at

$$\begin{aligned} Y^* &= AWY^* + (I - A)Y_0, \\ &= (I - AW)^{-1}(I - A)Y_0, \end{aligned}$$

provided that $(I - AW)$ is in fact invertible.

The equilibrium solution is the theoretical argument used by Friedkin to support various centrality measures (Friedkin 1991). We can write $(I - AW)^{-1}$ as $\sum_{k=0}^{\infty}(AW)^k$, and with the observation that $(AW)^k$ gives the weighted influence running between each pair of vertices in k steps, this gives a measure of how much influence is being asserted by each actor on the complete network, which might be a good interpretation of centrality.

The use of a continuous variable for modelling opinion dynamics, is something which cannot be translated to my model easily. I am using a discrete dichotomous analogy. Either you have viewed the movie, or you haven't. But, the general idea of how social influence is being exerted is similar to my idea. However, a dichotomous preference cannot propagate through the network in the model of Friedkin, since the preferences are continuous, which makes it less relevant to us.

1.3.2 Cascading

Another model of social influence processes is considered by Watts (2002). The basic idea is similar to my idea:

[2]This is the model being considered in Friedkin (1998) which gives the most general formulation of the 'Social Influence Network Theory', as Friedkin terms his own theory.

> "When deciding which movie or restaurant to visit, we often have little information with which to evaluate the alternatives, so frequently we rely on the recommendation of friends, or simply pick the movie or restaurant to which most people are going. (...) In all these problems, therefore, regardless of the details, individual decision makers have an incentive to pay attention to the decisions of others." (Watts 2002:5767).

Watts analyses a model in which people[3] make binary choices, that is, we either go to a restaurant or not, we either join a movement or we do not etc... This is then restricted to a single issue. This is similar to the spreading of a preference for a book, song or movie.

More specifically, Watts' model is constructed as follows. We build a network of n vertices having degree k with probability p_k and mean degree z. Each vertex can either be 'on' (having a preference for the issue) or 'off'. Initially all vertices are set to 'off', and we begin the process by setting one vertex (or more if we like) to 'on'. Each vertex has a threshold $0 < \phi < 1$, and if the fraction of neighbours that are switched on exceeds the threshold ϕ the vertex will also be switched on. In this way, the preference propagates through the network.

Watts researches the phenomena of 'global cascades', or the condition that the complete network[4] is switched on. He finds for uniform random graphs[5] and homogenous thresholds (meaning that each vertex has the same threshold) that global cascades take place only in a specific region. For a certain threshold ϕ^* global cascades take place below a mean degree z^*. The boundary z^* for which global cascades take place decreases if ϕ^* increases. This means that if a network is densely connected, we need to have a lower threshold in order to let the preference propagate throughout the network entirely.

Analysis suggest that if thresholds become more heterogeneous, the window of obtaining a global cascade increases. Oddly, if the degree distribution becomes more heterogeneous the window lessens.

Watts' model suggests that the actual structure of the network plays an important role in how the process unfolds itself. Depending on connectivity and the threshold distribution, the results for how many vertices have actually been switched on can vary widely. So, indeed this needs to be taken into account when modelling social influence.

However, the main interest of Watts lies with describing how 'hypes' might arise from a social influence model. He describes under what conditions a hype is more likely. My interest however, does not lie with the hypes only. I also would like to take the less successful items into account. The analysis of Watts does show that 'tipping point' behaviour might arise from social influence.

[3] Watts extends the model to other entities, such as power houses which may or may not fail. These are however mostly irrelevant for my discussion here.

[4] More specifically, the connected component, since some vertices might not be connected to the initial vertices which have been switched on.

[5] In a uniform random graph each edge has an equal probability z/n of appearing, leading to a Poisson probability distribution $p_k = e^{-z}z^k/k!$.

Chapter 2

Formalisation

2.1 Model

Below I describe the model I developed. Two distributions are my main concern: the 'uncertainty' distribution and the 'popularity' distribution. The uncertainty distribution will indicate what the probability is of receiving a specific number of votes for items with a given quality. The popularity distribution gives the probability that any random item has a specific number of votes.

Each item has an associated quality $\phi_i > 0$ drawn from a quality distribution $\rho(\phi)$ with an average quality of μ and variance σ. We start out with s items, having m votes each and an average quality. The number of votes obtained by item i is denoted by k_i. At each time a new item will be introduced without any votes, while there will be m votes cast between two successively introduced items. The probability that an item will obtain another vote is dependent on its quality and on the number of votes it already received. The balance between attracting votes because of its quality and because of the number of votes is thought of as the 'social influence', and is denoted by $0 \leq \lambda \leq 1$. An overview of the parameters used in the model is found in table 2.1. We will use the so-called continuum approach, which assumes that k is continuous, and not discrete, after Albert and Barabasi (2002). The derivation of the results follows largely their approach.

The probability that item i obtains another vote is

$$\Pi_i = (1-\lambda)\frac{\phi_i}{\sum_l \phi_l} + \lambda\frac{k_i}{\sum_l k_l}. \tag{2.1}$$

The rate at which item i will attract votes can be given by

$$\frac{\partial k_i}{\partial t} = m\left[(1-\lambda)\frac{\phi_i}{\sum_l \phi_l} + \lambda\frac{k_i}{\sum_l k_l}\right], \tag{2.2}$$

since m votes will be cast at each time.

ϕ	Quality
ρ	Quality Distribution
μ	Average quality of ρ
σ	Variance of the quality
k	Number of votes
λ	Social Influence
m	Number of votes per newly introduced item

Table 2.1: Overview of the general set of parameters for the model.

Now $\sum_l k_l$ is simply the total number of votes cast, plus the initial votes. So this sums to $(t+s)m$, which we approximate by tm for $t \gg s$. On average the quality of items is μ. So, after a long enough period, we can expect the average quality of all items to be about μ, or $\sum_l \phi_l = (t+s)\mu$, which we approximate by $t\mu$ for $t \gg s$. Thee rate at which item i will attract votes can then be written as

$$\frac{\partial k_i}{\partial t} = m\left[(1-\lambda)\frac{\phi_i}{t\mu} + \lambda\frac{k_i}{tm}\right],$$

or

$$\frac{\partial k_i}{\partial t} = \left[(1-\lambda)\frac{m\phi_i}{t\mu} + \lambda\frac{k_i}{t}\right]. \tag{2.3}$$

Since new items are introduced at t_i without any votes $k_i(t_i) = 0$. The solution of the differential equation is then

$$k_i(t) = \left[\left(\frac{t}{t_i}\right)^{\lambda} - 1\right](1-\lambda)\frac{m\phi_i}{\lambda\mu}, \tag{2.4}$$

where $0 < \lambda < 1$, $t_i > 0$ and $\mu > 0$. Details of this derivation are given in appendix C. This equation shows that the number of votes an item gets grows with t, as expected. It also allows for items later introduced to attract votes at a higher rate, as long as they have a higher quality.

This entails both a 'rich-get-richer' and a 'good-get-richer' effect. So, for items having a similar quality, older items will have attracted more votes than younger items. But, good items that are introduced at a later time can still increase the number of votes they receive above the number of votes older items have. An illustration of this can be seen in figure 2.1.

We would like to find out what the distribution of votes after a while looks like. More specifically, let us examine items having a specific quality ϕ after time t. Let us draw a random item from those items having quality ϕ after time t, and denote the number of votes of that item by $X_{t,\phi}$ and the time of introduction by $\tau_{t,\phi}$. We wish to examine the probability $\mathbb{P}(X_{t,\phi} < k)$ for some k.

If items have the same quality ϕ they all attract votes at the same rate. The only reason that different items have different number of votes is that they

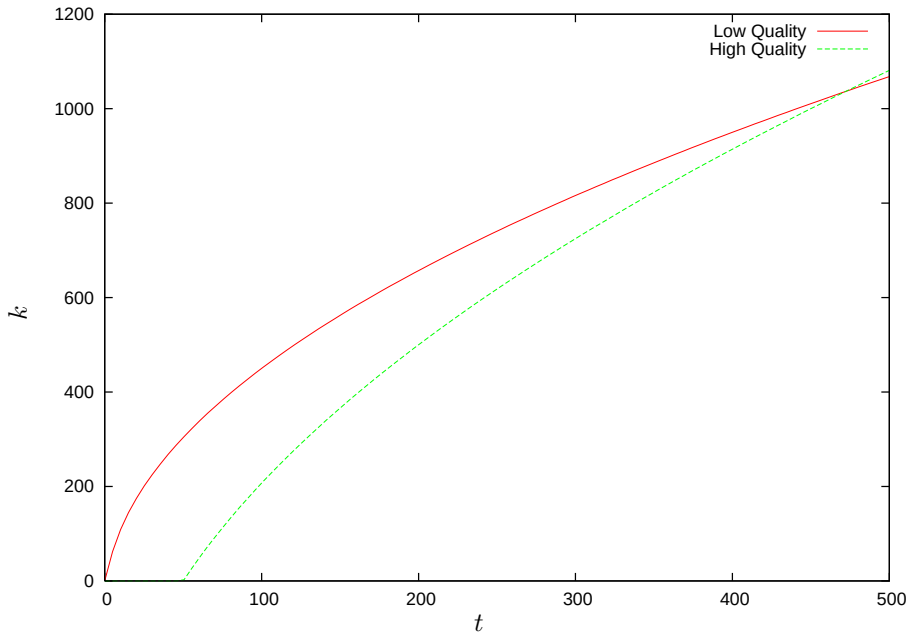

Figure 2.1: Growth trend of two different items. The first one is introduced at $t_i = 1$ and the second at $t_i = 50$. The quality of the latter however is 10 times as high as that of the former. Social influence is set to $\lambda = 0.5$ and $m = 100$ votes are cast before new items are introduced.

were introduced at different times. So, using equation 2.4, we can write the probability as

$$\mathbb{P}(X_{t,\phi} < k) = \mathbb{P}\left(\tau_{t,\phi} > \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}} t\right).$$

We started out with s items, so there is a probability of $s/(t+s)$ that an item was one of those first items. Since we introduce a new item at each 'time step', $\tau_{t,\phi}$ is uniformly distributed with $1/(t+s)$ (proportional to the probability at which items with quality ϕ appear) hence $\mathbb{P}(\tau_{t,\phi} < c) = 1 - c/(t+s)$. Using this we can write

$$\mathbb{P}(X_{t,\phi} < k) = 1 - \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}} \frac{t}{t+s},$$

which we differentiate with respect to k, take the limit for $t \to \infty$ and obtain

$$\mathbb{P}(X_\phi = k) = \mu((1-\lambda)m\phi)^{\frac{1}{\lambda}} \left(k\lambda\mu + (1-\lambda)m\phi\right)^{-\left(1+\frac{1}{\lambda}\right)}. \tag{2.5}$$

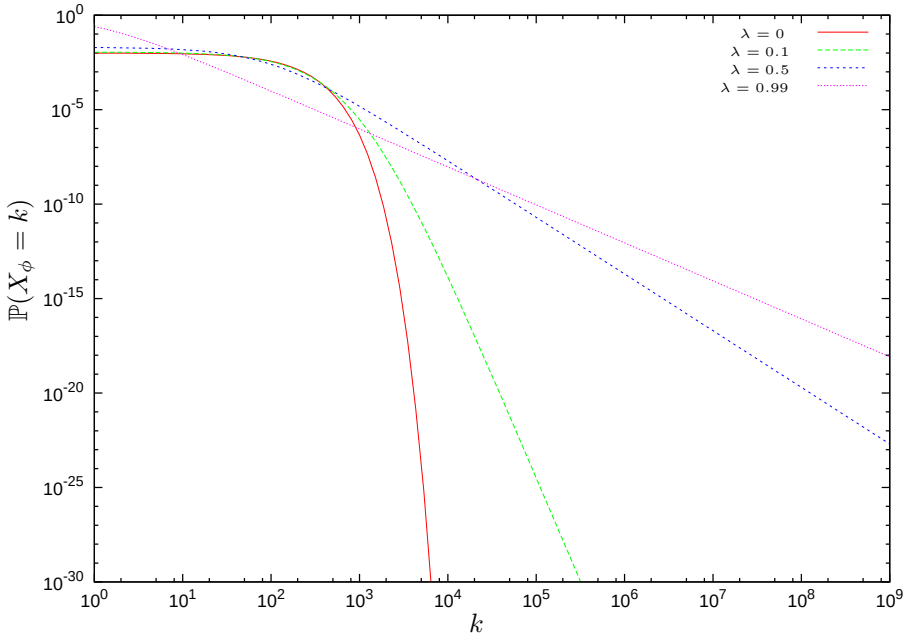

Figure 2.2: Uncertainty distribution for various values of $\lambda = (0, 0.1, 0.5, 0.99)$. Quality is taken to be the average quality, while we assume that 100 votes are cast ($m = 100$) before a new item is introduced.

where X_ϕ is the number of votes of a random item after a long enough time having a specific quality ϕ. Details are again given in appendix C. This distribution is then the uncertainty distribution. It shows how the number of votes may vary for items having a certain quality.

Two special cases need to be derived in another way, since the above analysis was for $0 < \lambda < 1$ only. For $\lambda = 0$ the uncertainty distribution follows an exponential distribution with exponent $\frac{\mu}{m\phi}$. If $\lambda = 1$ the uncertainty distribution follows a power law distribution[1] with an exponent of 2. The distribution for various levels of social influence can be seen in figure 2.2.

Some properties of this distribution can be calculated relatively straightforward. The expected number of votes for $0 < \lambda < 1$ is given by

$$\overline{k} = \frac{m\phi}{\mu}, \tag{2.6}$$

and the variance for $0 < \lambda < 1/2$ is given by

$$\frac{m^2\phi^2}{(1-2\lambda)\mu^2} = \frac{\overline{k}^2}{1-2\lambda}. \tag{2.7}$$

[1] Of course, if items will be introduced without any initial votes, the process never gets started, so they need to be introduced with some votes in order to get the process started.

For $\lambda \geq 1/2$ the variance is infinite. For details on the derivation see appendix C.

The expected number of votes is independent of social influence. So social influence does not affect the number of total votes cast. It only affects the way the number of votes are distributed among items with a certain quality. More specifically, the variance increases with λ, reaching infinity for $\lambda \geq 1/2$. Higher social influence thus makes it harder to predict whether an item will become popular or not. The variance is actually the mean number of votes squared divided by $1 - 2\lambda$. In other words, a higher mean implies higher variance.

Furthermore, the expected number of votes increases with quality. Items with a higher quality will gain more votes than items with a lower quality. More surprisingly, the variance also increases with quality. This means that it is harder for items with a higher quality to predict how many votes it will attract, even though the expected number of votes is higher.

Finally, the number of votes cast m between two successively introduced items increase both the average popularity and the variance. A market in which items are introduced rather quickly (i.e. few votes cast per introduced item, or a low m) results in a lower variance, while a more static market (i.e. more votes cast per introduced item or a high m) results in a higher variance. This is the result of the longer time that social influence can play a role in a more static market. If m is high, it provides more time for the existing items to attract votes biased by social influence.

The increasing uncertainty with a higher social influence and a higher quality is confirmed empirically by Salganik et al. (2006). They performed an on-line experiment, where visitors of their webpage were able to listen and download various songs. Each visitor would be assigned to one of three different settings. In the first (independent) setting, no figures on the number of downloads were provided, and songs were displayed in random order (i.e. no, or little, social influence). In the second (social influence) setting, songs were still displayed in random order, but this time the download count was provided. In the third setting, not only were the download counts provided, but the songs were sorted in decreasing order of popularity.

They found that uncertainty (unpredictability[2] in their terms) increased from setting one through three. In other words, a higher social influence increases the variance. They also suggest that uncertainty varies with quality:

> "In general the 'best' songs never do so badly, and the 'worst' songs never do extremely well, but almost any other result is possible. Unpredictability also varies with quality—measured in terms of market share, the 'best' songs are the most unpredictable, whereas when measured in terms of rank, intermediate songs are the most unpredictable." (Salganik et al. 2006:855)

So my model may serve as a theoretical basis for the findings of Salganik et al.

[2] Unpredictability is taken to be the difference in market shares across multiple experiments in the same setting.

2.2 Popularity Distribution

As the probability distribution for items with a given quality—the uncertainty distribution—is available, we can turn to the probability distribution for all items—the popularity distribution. The uncertainty distribution was for items having a specific quality ϕ, hence is proportional to $\rho(\phi)$. So, let X denote the number of votes of a random item after a long enough time. The probability of obtaining a number of votes k, $\mathbb{P}(X = k)$ which we shorten as $\mathbb{P}(k)$, can be written as

$$\mathbb{P}(k) = \int_{\phi_{\min}}^{\phi_{\max}} \rho(\phi)\mathbb{P}(X_\phi = k)\mathrm{d}\phi, \tag{2.8}$$

which means that we take the average of the uncertainty distribution over each level of quality ϕ.

The mean uncertainty for items with a given quality ϕ is $m\phi/\mu$. The mean popularity is thus given by the integral over the quality distribution ρ, or

$$\int \frac{m\phi}{\mu}\rho(\phi)\mathrm{d}\phi.$$

Since μ is the mean quality, which is $\int \phi\rho(\phi)\mathrm{d}\phi$, we obtain

$$\begin{aligned} \frac{m}{\mu}\int \phi\rho(\phi)\mathrm{d}\phi &= \frac{m}{\mu}\mu, \\ &= m. \end{aligned}$$

So the mean popularity is simply m, irrespective of the quality distribution. This is quite logical, since after time t, a total of $(t+s)m$ votes have been cast for $t+s$ items. On average this is $(t+s)m/(t+s)$ votes per item, which of course reduces to m.

The variance of popularity can be deduced using

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Since the last term is simply the mean squared this simplifies to

$$\mathbb{E}(X^2) - m^2.$$

Solving $\mathbb{E}(X^2)$ and plugging it back in gives

$$\mathrm{Var}(X) = \frac{2m^2(1-\lambda)}{\mu^2(1-2\lambda)}\int_{\phi_{\min}}^{\phi_{\max}} \phi^2\rho(\phi)\mathrm{d}\phi - m^2.$$

If σ is the variance of the quality distribution, then $\sigma + \mu^2 = \int \phi^2\rho(\phi)\mathrm{d}\phi$. Simplifying then gives

$$\mathrm{Var}(X) = \frac{m^2(2\sigma(1-\lambda)+\mu^2)}{\mu^2(1-2\lambda)}, \tag{2.9}$$

We will use this equation to calculate the variance for the various distributions. It can be seen that the variance of the popularity distribution will increase with the variance of the quality distribution.

The asymptotic behaviour of $\mathbb{P}(X_\phi = k)$ is $k^{-(1+1/\lambda)}$, or more formally

$$\mathbb{P}(X_\phi = k) \in \mathcal{O}\left(k^{-(1+\frac{1}{\lambda})}\right),$$

where $\mathcal{O}$ is the so-called big-Oh notation. We write this somewhat loosely as $\mathbb{P}(X_\phi = k) \sim k^{-(1+1/\lambda)}$. All linear combinations of $\mathbb{P}(X_\phi = k)$ for various values of ϕ have the same asymptotic behaviour of $k^{-(1+1/\lambda)}$. The popularity distribution then has the same asymptotic behaviour, regardless of the quality distribution.

We will review three quality distributions, the delta Dirac distribution (in which all the items have a similar quality), the uniform distribution and the exponential distribution.

2.2.1 Dirac Quality Distribution

Let us review the situation that $\rho(\phi) = \delta(\phi - 1)$. The $\delta(\phi - \phi^*)$ distribution is called a Dirac distribution. It can be loosely thought of as a distribution such that the value ϕ^* (which equals 1 in this case) is drawn with certainty, and all other values have a probability of zero of being drawn[3]. Obviously, the average quality $\mu = 1$ and the variance $\sigma = 0$. In other words, all items have the same quality. Then, each item will attract votes at a similar rate, only depending on how much weight is given to the preferential attachment. Using equation 2.8, we write

$$\mathbb{P}(k) \quad = \quad \int_0^\infty \delta(\phi - 1)\mathbb{P}(X_\phi = k)\mathrm{d}\phi,$$

which equals

$$\mathbb{P}(k) = (m(1-\lambda))^{\frac{1}{\lambda}} \left(k\lambda + m(1-\lambda)\right)^{-(1+\frac{1}{\lambda})}, \tag{2.10}$$

which is equation 2.5 with $\mu = \phi = 1$. So, the expected value and the variance also remain similar, only with $\mu = \phi = 1$.

2.2.2 Uniform Quality Distribution

Now let us look at a somewhat more complex distribution. Let us assume that quality is uniformly distributed on the $[0, 1]$ interval, so that 'good books' are just as likely as 'bad books'. This distribution has a mean of $\mu = 1/2$ and a variance of $\sigma = 1/12$. If $\lambda = 1$, we of course maintain a power law distribution

[3] More formally, the Dirac distribution can be defined such that $\int_{-\infty}^{\infty} f(\phi)\delta(\phi - \phi^*)\mathrm{d}\phi = f(\phi^*)$ for a smooth enough f.

with an exponent of 2, since quality then has no influence. The distribution for $\lambda = 0$ remains approximately exponential. We write

$$\mathbb{P}(k) = \int_0^1 \mathbb{P}(X_\phi = k)\mathrm{d}\phi,$$

or

$$\mathbb{P}(k) = \int_0^1 \frac{1}{2}(m(1-\lambda)\phi)^{1/\lambda} \left(k\lambda\frac{1}{2} + m(1-\lambda)\phi\right)^{-(1+1/\lambda)} \mathrm{d}\phi.$$

Simplifying the integral gives us

$$\mathbb{P}(k) = \frac{1}{2m(1-\lambda)} \left| B\left(\frac{2m(\lambda-1)}{k\lambda}, 1+\frac{1}{\lambda}, -\frac{1}{\lambda}\right)\right|, \tag{2.11}$$

where B is the incomplete Beta function

$$B(x,a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t, \tag{2.12}$$

and $|\cdot|$ denotes taking the absolute value.

The expected value is of course again m, while the variance increases with λ and with m and approaches infinity as $\lambda \to 1/2$. The variance for the uniform distribution is given by

$$\frac{m^2(5-2\lambda)}{3(1-2\lambda)}.$$

This distribution is a bit more stretched than if we assume the Dirac distribution, as in equation 2.10. A more heterogeneous quality distribution thus results in a more skewed distribution.

2.2.3 Exponential Quality Distribution

Perhaps a somewhat more realistic distribution is an exponential distribution $\rho(\phi) = \gamma e^{-\gamma\phi}$ with an average quality of $\mu = 1/\gamma$ and a variance of $\sigma = 1/\gamma^2$. Here, 'good books' are rare, while the market is overwhelmed by 'shitlit'. Again, the distribution becomes even more stretched out, and the distribution for $\lambda = 0$ no longer is an exponential distribution. The general equation is

$$\mathbb{P}(k) = \int_0^\infty \frac{1}{\gamma}(m(1-\lambda)\phi)^{1/\lambda} \left(\frac{k\lambda}{\gamma} + m(1-\lambda)\phi\right)^{-(1+1/\lambda)} \gamma e^{-\gamma\phi}\mathrm{d}\phi$$

for which we obtain

$$\mathbb{P}(k) = \frac{\Gamma\left(1+\frac{1}{\lambda}\right) U\left(1+\frac{1}{\lambda}, 1, \frac{k\lambda}{m(1-\lambda)}\right)}{m(1-\lambda)} \tag{2.13}$$

where Γ is the gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$$

and U is the so-called 'confluent hypergeometric function of the second kind' and has an integral representation as

$$U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{b-a-1} \mathrm{d}t$$

and has asymptotic behaviour

$$U(a,b,z) = z^{-a} \left(1 + \mathcal{O}\left(\frac{1}{z}\right)\right)$$

for $z \to \infty$, where $\mathcal{O}$ is the so-called big-Oh notation (Abramowitz and Stegun 1970). The asymptotic behaviour is thus again $\mathbb{P}(k) \sim k^{-(1+\frac{1}{\lambda})}$.

Remarkably, the parameter γ of the exponential quality distribution does not affect the popularity distribution at all. So, if a quality distribution seems to follow an exponential distribution, we do not even have to estimate the parameter γ, since this has no effect on the popularity distribution. This is the result of the fact that for an exponential quality distribution the variance is the mean squared, or $\sigma = \mu^2$. The variance for the exponential distribution is given by

$$\frac{m^2(3-2\lambda)}{(1-2\lambda)},$$

which, again, is higher than for the uniform distribution. However, since γ does not play a role, a more skewed quality distribution per se need not result in a more skewed popularity distribution.

2.3 Analysis

In general, we can state that the more heterogeneous the quality distribution becomes, the more stretched the resulting popularity distribution becomes, although this result it somewhat ambiguous if we study an exponential quality distribution. Still, all retain an asymptotic behaviour of $\mathbb{P}(k) \sim k^{-(1+1/\lambda)}$ for $k \to \infty$, thus showing power law behaviour in the tail, regardless of the quality distribution.

2.3.1 Volatility

Since m is the mean for all popularity distributions, we may simply interpret m to be the average number of views a movie gets, or the average number of sales a book achieves. This interpretation allows for a simple estimation of m.

However, another, more interesting interpretation is also possible. Since m is also the number of votes cast between two successively introduced items, we may interpret the reciprocal $1/m$ to indicate the relative volatility of a market.

So, let us define volatility as $\nu = 1/m$. A volatility of $\nu = 1$ indicates a highly volatile market. It indicates that relative to the number of views, new items are introduced at a fast pace ($m = 1$). A volatility near $\nu = 0$ then

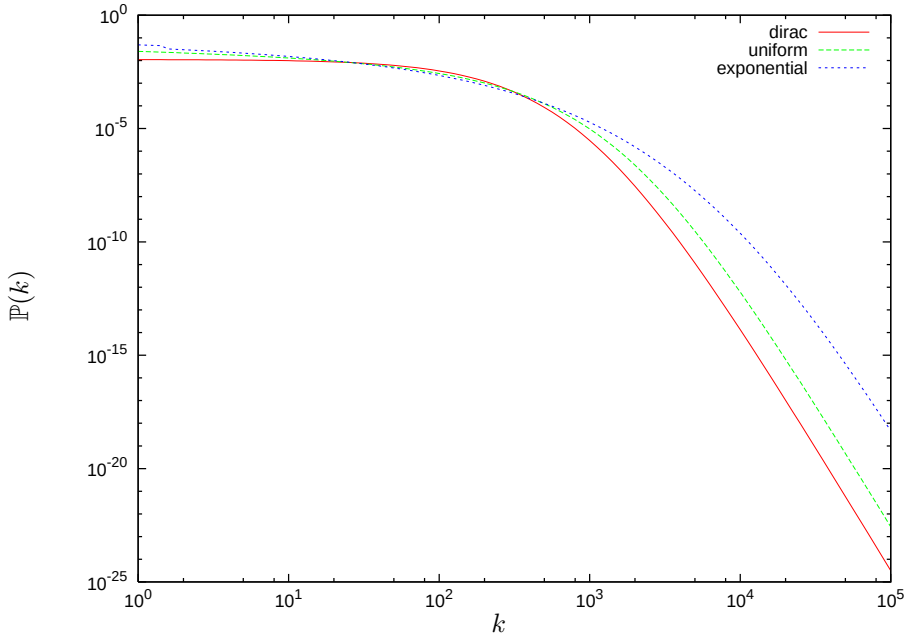

Figure 2.3: Popularity distribution for the various quality distributions. Here $m = 100$ and $\lambda = 1/10$. It can be seen that the Dirac distribution is the least stretched out, the uniform distribution is a bit more stretched out, while the exponential distribution is even more stretched. The asymptotic behaviour can be seen to be similar for all three distributions.

indicates a nearly static market ($m \to \infty$), since new items are almost never introduced. In a highly volatile market, new items are frequently introduced, while in a less volatile market, new items tend to be introduced less often.

From the mean number of votes m, it can be deduced that a market in which volatility is higher receives less votes on average per item. This makes sense. After all, when we are confronted with new products daily, we tend to loose track of previous products relatively quickly. Similarly, in a rather static market, we are confronted with similar items for days on end, giving a higher average number of votes per item.

From the variance of the uncertainty distribution (equation 2.7) it can be seen that a more volatile market results in less uncertainty and a lower average number of votes.

2.3.2 Inequality

The popularity distribution is quite extremely skewed. It suggest that the bulk of the items—usually more than halve—never achieve anything above the average. Vice versa, most of the total number of votes are accounted for by a small

number of popular items. This distribution approximately follows the 80/20 rule of thumb, which implies that the most popular 20% of the items account for 80% of the votes. The exact figures depend on social influence of course.

Let us study the simplest version of the model, assuming the Dirac distribution. How many votes does the less popular halve obtain maximally? Or what is the K for which $\mathbb{P}(X < K) = 1/2$. So we need to solve

$$\mathbb{P}(X < K) = \int_0^K \mathbb{P}(k)\mathrm{d}k = \frac{1}{2}.$$

Since

$$\mathbb{P}(X < K) = 1 - (m(1-\lambda))^{\frac{1}{\lambda}}(K\lambda + m(1-\lambda))^{-1/\lambda},$$

we obtain

$$K = \frac{\left(2^\lambda - 1\right) m(1-\lambda)}{\lambda}.$$

So, for $\lambda = 0.5$ and $m = 100$ for example, about half of the items have less than approximately 42 votes. The fraction L that is accounted for by the less popular halve can be given by the average number of votes for items having less than K votes, compared to the general average number of votes. This is

$$L = \frac{\int_0^K k\mathbb{P}(k)\mathrm{d}k}{\int_0^\infty k\mathbb{P}(k)\mathrm{d}k} = \frac{\mathbb{E}(X|X < K)}{m},$$

where $\mathbb{E}(X|X < K)$ is the expected number of votes for items having less than K votes. In the example this evaluates to about 8.5%. So the less popular halve of the items account for only 8.5% of the votes, while the more popular halve account for about 91.5%.

More generally, for the Dirac distribution, the level K at which a proportion p of the items remains below is given by

$$K = \frac{m(1-\lambda)((1-p)^{-\lambda} - 1)}{\lambda}$$

Now the fraction L of the votes that are accounted for by the least popular p items can be given as

$$L(p) = \frac{1 - (1-p)^{1-\lambda} - p(1-\lambda)}{\lambda}, \tag{2.14}$$

which is known as a *Lorenz curve*. Rather surprisingly, the Lorenz curve is independent of m. This means that inequality is determined by λ only. So, whether the market is volatile or not does not have an influence at how unequal the distribution is. The behaviour of this Lorenz curve is displayed for various levels of social influence in figure 2.4. It can be seen that with rising social influence the inequality rises.

Clearly, the more a Lorenz curve deviates from the line of full equality, the more unequal it will be. The area below the line of full equality thus signifies the

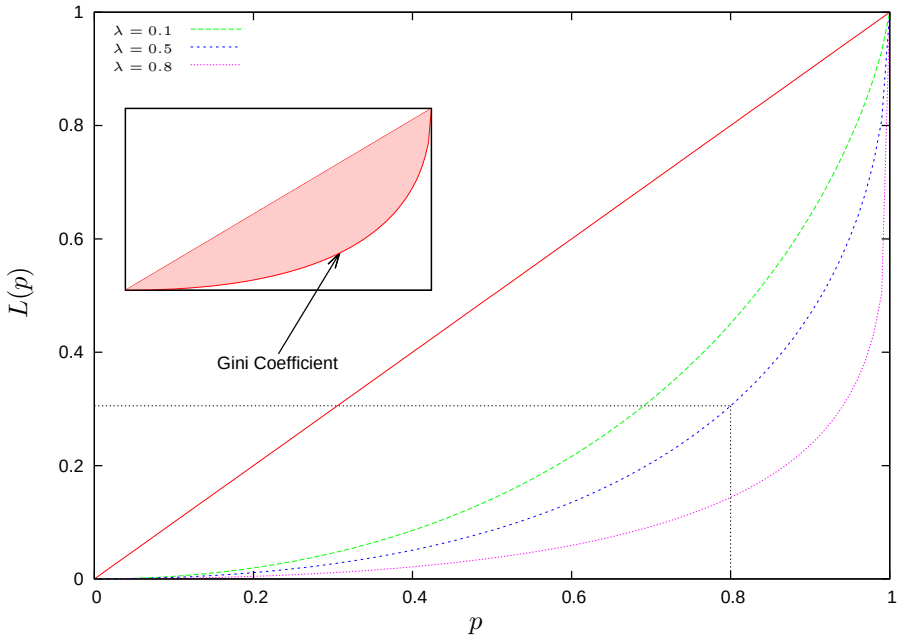

Figure 2.4: Lorenz curves for various levels of social influence illustrating the inequality for the Dirac popularity distribution. For example, for $\lambda = 0.5$ the bottom 80% of the items account for a little over 30% of the votes. Stated otherwise: the top 20% is responsible for almost 70% of the votes. The main diagonal is the line of full equality. The inset shows visually how the Gini coefficient is obtained.

amount of inequality, as is illustrated in the inset of figure 2.4. The following definition quantifies this inequality and is known as the *Gini coefficient*

$$G = 1 - 2\int_0^1 L(p)\mathrm{d}p,$$

which for the Dirac distribution equals

$$G = \frac{1}{2-\lambda}, \tag{2.15}$$

giving a Gini coefficient of $G = \frac{1}{2}$ even for $\lambda = 0$. It should be borne in mind that this analysis is for the Dirac quality distribution only. The other distributions probably show even more extreme inequality. For more information on Lorenz curves and Gini coefficients see Gastwirth (1972).

The Gini coefficient is widely used to give an indication of the inequality of the distribution of wealth and income in countries. Although this type of distribution is unrelated to the distribution under study here, to get a sense of what a low and high Gini coefficient is, we might as well give some examples. For

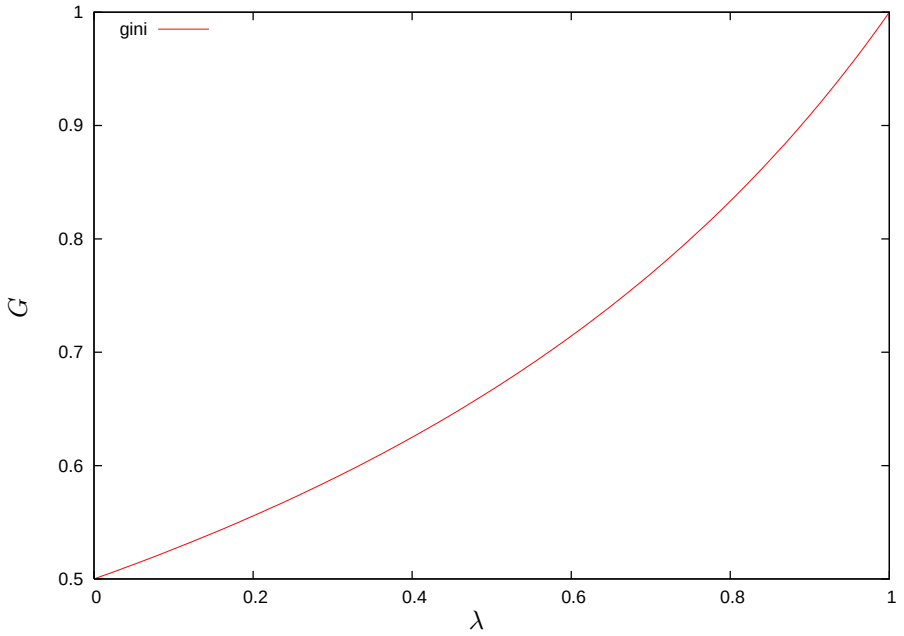

Figure 2.5: Gini coefficient G for the Dirac distribution. Even for $\lambda = 0$ inequality as measured by the Gini coefficient is $\frac{1}{2}$, which is quite high.

comparison, most countries[4] have Gini coefficients between 0.3–0.6 for the income distribution. Sweden has a traditionally low Gini coefficient of 0.25, while inequality in the United States is somewhat higher with 0.40, but compared with one of the highest Gini coefficients of 0.63 for Sierra Leone, the United States does not have an extremely skewed income distribution. My model thus shows a greater inequality than most income distributions.

In the experiment of Salganik et al. (2006) they find a Gini coefficient for the independent setting of somewhere between 0.2–0.3. The Gini coefficient rises when social influence is increased, ranging somewhere between 0.4–0.55. These figures are substantially lower than what I predict. It is possible however, that if the experiment lasts longer, and if the market grows, the Gini coefficient might rise to match the level I predict. Still, the fact that inequality increases with social influence is confirmed by the theoretical model.

If we apply the same analysis of the Lorenz curve and the Gini coefficient to the uncertainty distribution, we obtain the same results. The mean number of votes for the less popular p items is proportional to the mean number of votes $m\phi/\mu$. Hence, the Lorenz curve is independent of the average $m\phi/\mu$, and so is the Gini coefficient. If we express uncertainty as a Gini coefficient, it is only effected by social influence and equals equation 2.15. Details of the derivation

[4]See UNDP (2007).

of the Lorenz curve and the Gini coefficient are given in section C.5.

From the model point of view, it is then no coincidence that Salganik et al. (2006) find rising inequality and uncertainty with social influence. After all, inequality and uncertainty are two side of the same coin, namely social influence. Both inequality and uncertainty rise with social influence.

2.4 Generating Networks

The model suggested here has connections with models for growing networks. Many networks, such as the internet, actor collaboration, power grid network and scientific citations show power law behaviour in their degree distribution. They are said to be scale-free. A network, or graph, has several vertices (websites, actors, power stations, scientific publications) and several edges or arcs[5] (hyperlinks, collaborations, power lines, citations) running from one vertex to another.

The number of incoming edges/arcs is said to be the indegree k_i of vertex i. Each vote in the model can then be seen as an incoming link for vertex (item) i, and thus signifying its degree. An overview of several models and their behaviour can be found in Newman (2003) and Albert and Barabasi (2002).

2.4.1 Classic Model

The first[6] model suggested to account for the observed degree distribution on the web is the BA-model, named after their authors Barabasi and Albert (1999). Their model works as following. We start out with some m_0 vertices, and introduce a new vertex at each time step. The new vertex establishes $m < m_0$ new arcs, where the other end of the arc is chosen with 'preferential attachment', that is, vertices with a higher degree have a higher probability of receiving an incoming arc. This process is then iterated.

Their conclusion is that the degree distribution stabilises after a while. That is, although the network keeps on growing, the degree distribution essentially remains the same. The distribution is then said to be a stationary distribution. They found for the indegree distribution $\mathbb{P}(k) = \frac{2m^2}{k^3}$ which is similar to my results for $\lambda = 1$ if we realise that $\sum k_i = 2mt$ for the BA-model (they examine both incoming and outgoing links) and equals mt for my model (I examine only incoming 'links', which I call votes).

Albert and Barabasi then go on to demonstrate that both growth and preferential attachment are essential for obtaining a scale free stationary distribution. They consider two alternative models, model A and model B. In model A there

[5] Edges usually refer to undirected links (i.e. they have no direction), while the term arcs refer to directed links (i.e. they point from some origin vertex to some destination vertex).

[6] De Solla Price (1965) presented an earlier model, which I mentioned at page 15. He used it to account for the observed power law distribution of scientific citations. This is approximately the same model as the one from Barabasi and Albert. Barabasi and Albert are the ones suggesting however, that a similar model could also be used for degree distributions of the internet for example.

is no preferential attachment, while in model B there is no growth. So in model A the endpoints are chosen with equal probability. If we set $\lambda = 0$ and assume a delta Dirac quality distribution my model mimics this behaviour, and my solution is the same as for their model A. In model B growth is absent but preferential attachment is present. The model seems to exhibit power law behaviour at first, but since the number of arcs only increase while the number of vertices does not, the network soon becomes completely connected[7]. Hence, the power law behaviour is unstable.

Our model generalises these results from Barabasi and Albert, and shows that 'preferential attachment' need not be fully present in order to obtain a power law distribution in the tail. In fact my model interpolates nicely between model A and the actual model suggested by Barabasi and Albert.

2.4.2 Fitness model

It seemed unrealistic that the BA-model would account for the complete growth behaviour of these complex networks. For example, when the search engine Google[8] was launched, it began attracting hyperlinks to itself at a much higher rate than other websites, despite its initial low indegree. So, something else was affecting the way vertices accumulate edges.

Bianconi

This is taken into account by Bianconi and Barabasi (2000), who have introduced 'fitness' of a vertex to account for this behaviour.

The process is largely the same as in the original model. We start with some initial number of vertices m_0, and introduce a new item every time step which links to m of the existing vertices. But, the probability of linking to a vertex is now not only dependent on the degree k_i of a vertex i but also of the fitness η_i. They assume the probability of getting an edge to be

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j}.$$

So the final degree distribution will depend on the fitness distribution ρ. For $\rho(\eta) = \delta(\eta - 1)$ a distribution in which all the items have a similar fitness, the model reduces to the original BA model, which of course coincides with $\lambda = 1$ in my model.

For a uniform fitness distribution on the $[0, 1]$ interval they obtain the solution

$$\mathbb{P}(k) \sim \frac{k^{-(1+C^*)}}{\log k},$$

where $C^* = 1.255$, which is more stretched out than the original BA model. This suggests that when more variance is introduced in the fitness, the variance

[7] A completely connected network or *complete graph*, is a graph in which every vertex is connected with all other vertices.

[8] http://www.google.com

of the degree distribution also increases, as is the case in my model. The exact forms do differ however.

My model might serve as an alternative fitness model to the model suggested by Bianconi and Barabasi (2000). It allows a qualitatively similar process of accumulating links (or votes in my vocabulary), but allows the possibility of varying the social influence as well. So quality can be given more or less weight depending on the context.

Pennock

Another model taking 'fitness' into account in a certain way is considered by Pennock et al. (2002). They consider the preferential attachment to be[9]

$$\Pi_i = (1-\lambda)\frac{1}{s+t} + \lambda\frac{k_i}{2mt}.$$

When compared to my version of preferential attachment (as stated in equation 2.1) it becomes clear that the model considered is the same as my model for the Dirac distribution (cf. equation 2.10). The only difference here is that they consider the total connectivity in the graph (i.e. $2mt$ instead of only the incoming links mt). Hence, the stationary indegree distribution is the same as mine.

The authors consider histograms with logarithmic binning for the model. The binning size thus increases exponentially, and the bounds for the i-th bin are $10^{i/6} - 1$ to $10^{(i+1)/6} - 1$. By substituting $k = 10^{k'/6}$ into the cumulative distribution, differentiating it with respect to k, and substituting back $k' = 6\log_{10} k$ they obtain a version of the probability distribution appropriate for visualising the model in a histogram with logarithmic binning. More generally, one can assume the bounds of the bins to be $e^{\alpha i} - 1$ to $e^{\alpha(i+1)} - 1$ for $\alpha > 0$. The transformed probabilities for the Dirac quality distribution then becomes

$$\alpha k(m(1-\lambda))^{\frac{1}{\lambda}}(m(1-\lambda)+k\lambda)^{-(1+\frac{1}{\lambda})},$$

where Pennock et al. have set $\alpha = \log 10/6$. As the authors also show, the maximum of the transformed probabilities is $m(1-\lambda)$ (in their article it is $2m(1-\lambda)$ because of the difference in counting the total degree). This means that for logarithmic binning, the mode is larger than zero (if $m > 0$ and $\lambda < 1$). The transformed probabilities first increase, obtain their maximum at $m(1-\lambda)$, and then decrease again. This means that the probabilities decrease slower than exponentially for low k. Beyond the maximum, the decrease in probabilities is faster than exponential.

The authors are mainly interested in the low k region of the connectivity of the web, as this deviates significantly from a power law. They show this model captures the behaviour for various subcategories of webpages well. They report

[9] In the original article they have used α instead of λ and m_0 instead of s. But substituting λ for α and s for m_0 seems a good idea, since it then agrees with my definitions.

relatively high social influence estimates[10] for companies (0.950) and newspapers (0.948) and somewhat lower scores for universities (0.612) and scientists (0.602). This seems to indicate that for websites of universities and scientists quality plays a higher role than for companies and newspapers.

My model and the model of Pennock et al. differ in two aspects. Firstly, I reserve the possibility of another quality distribution, while they only investigate a Dirac quality distribution. Secondly, they do not interpret the parameter λ as being a measure of social influence. These two differences allow us to investigate what the effects of quality and social influence are. My conclusions regarding uncertainty and inequality go beyond the analysis Pennock et al. have undertaken.

[10]They use a histogram with logarithmic binning and least squares to fit the distribution. Although a preliminary analysis shows this method to be quite good, the standard errors are larger than with maximum likelihood estimation (see chapter 3.)

Chapter 3

Estimation and Testing

3.1 Estimation

Having formalised the model, it can be empirically analysed. More specifically, we would like to know what the amount of social influence is and how many votes are cast in between two successively introduced items. In short, we need to estimate λ and m on the basis of a sample.

Suppose we have a random sample $x_1, \ldots, x_n$ of size n from some market of items. The *population* consists of all items, and we have drawn a sample from this population. Of course we do not know what the distribution of popularity in the population looks like, and we do not know what the amount of social influence in the population is. Therefore, we try to estimate the social influence from our sample. We denote the estimate of social influence with $\hat{\lambda}$, where the hat over the parameter signifies an estimate of the actual value λ in the population.

Suppose we know λ. Then we could describe the probability that our first sampled item has exactly x_1 votes by our popularity distribution. Given λ, the probability of drawing x_1 from the population distribution is $\mathbb{P}(x_1|\lambda)$. Of course, the probability of drawing x_2 from the population is similar: $\mathbb{P}(x_2|\lambda)$. The probability of drawing both x_1 *and* x_2 is then $\mathbb{P}(x_1|\lambda)\mathbb{P}(x_2|\lambda)$. Hence, if we would know λ, the probability of obtaining the exact sample $x_1, \ldots, x_n$ is

$$\prod_{i=1}^{n} \mathbb{P}(x_i|\lambda).$$

As I said earlier however, we do not know what the social influence in the population is. But let us turn things around. We may ask how likely a specific value of λ is given the sample. For some values of λ we will see a higher probability of having drawn the sample than for other values of λ. We say that the value of λ for which that probability is higher is a more *likely* value. We

can write the *likelihood function* as

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \mathbb{P}(x_i|\lambda).$$

The values of λ for which the likelihood $\mathcal{L}(\lambda)$ is greater are more likely candidates for the population parameter λ. So the value for which the likelihood obtains a (global) maximum can be considered to be the most likely value. This gives our estimate $\hat{\lambda}$ of the population parameter λ.

The product of the probabilities is usually rather difficult to analyse. A useful transformation in that regard is taking the logarithm of the likelihood. Doing so yields

$$\begin{aligned} \log \mathcal{L}(\lambda) &= \log \prod_{i=1}^{n} \mathbb{P}(x_i|\lambda) \\ &= \sum_{i=1}^{n} \log \mathbb{P}(x_i|\lambda). \end{aligned}$$

Since the logarithm is a monotonically increasing function, the log-likelihood will obtain its maximum at the same value of $\hat{\lambda}$ as for the ordinary likelihood. So we can maximise the log-likelihood instead of the ordinary likelihood. In order to find the maximum, we need to set the derivative equal to zero, and solve for λ. So the λ for which

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = 0$$

holds gives us the estimate $\hat{\lambda}$. This procedure is known as *maximum likelihood estimation*, and gives us the *maximum likelihood estimator* (MLE). Estimators of this kind are known to provide good estimates. For more information on maximum likelihood estimation see Myung (2002).

Lets work out this procedure for the simplest model. According to equation 2.10 for the Dirac quality distribution, the probability is:

$$\mathbb{P}(k) = (m(1-\lambda))^{\frac{1}{\lambda}} \left(k\lambda + m(1-\lambda)\right)^{-(1+\frac{1}{\lambda})}.$$

So given the sample of $x_1, \ldots, x_n$, we can write the log-likelihood as

$$\log \mathcal{L} = \frac{n}{\lambda} \log m(1-\lambda) - \left(1 + \frac{1}{\lambda}\right) \sum_{i=1}^{n} \log \left(x_i\lambda + m(1-\lambda)\right). \tag{3.1}$$

Now we would like to maximise the likelihood. In order to do so, we need to set the derivative equal to zero. Unfortunately, this yields no closed expression to find λ. So we will have to approximate λ numerically. Various methods are available to numerically maximize the likelihood, see section B.2 in the appendix for more information.

We have simplified the log likelihood for the other models as much as possible. For the uniform distribution we obtain

$$\log \mathcal{L} = -n \log 2m(1-\lambda) + \sum \log \left| B\left(\frac{2m(\lambda-1)}{k\lambda}, 1+\frac{1}{\lambda}, -\frac{1}{\lambda}\right)\right|, \qquad (3.2)$$

and for the exponential distribution

$$\begin{aligned} \log \mathcal{L} &= n\left(\log \Gamma\left(1+\frac{1}{\lambda}\right) - \log m(1-\lambda)\right) + \\ &\quad \sum \log U\left(1+\frac{1}{\lambda}, 1, \frac{k\lambda}{m(1-\lambda)}\right). \end{aligned} \qquad (3.3)$$

Similarly for the uniform and exponential behaviour there is no closed-form expression, so we will numerically determine for which values of λ the log likelihood obtains its maximum.

We can estimate m in a simple way as it is the mean popularity, with $\hat{m} = \overline{x} = \frac{1}{n}\sum x_i$. It is easy to see that this gives an unbiased estimate since

$$\begin{aligned} \mathbb{E}(\hat{m}) &= \mathbb{E}\left(\frac{1}{n}\sum x_i\right) \\ &= \frac{1}{n}\sum \mathbb{E}(x_i) \\ &= \frac{1}{n} nm \\ &= m. \end{aligned}$$

This estimate is not the MLE given by the above procedure, however. If we would use the MLE, it would need to be determined numerically, thereby increasing the required computation power. Furthermore, since we know m does not depend on λ or on the quality distribution, we would like to keep it constant when estimating λ for the various quality distributions. That way we can compare the various estimates of λ somewhat better. Using the mean for estimation provides us with such a steady estimate.

3.2 Testing

Having an estimate $\hat{\lambda}$ available does not imply that the sample was in fact drawn from our hypothesised distribution. It merely gives the most likely value of λ should the sample be drawn from our distribution. We still need to find out whether our theoretical distribution is a 'good' model for the empirical distribution at all. In order words, we need to perform a goodness-of-fit test.

Such a test is the Kolmogorov-Smirnov test, or KS-test for short. This test provides a statistic by which we can judge whether our distribution fits the sample nicely. I follow the procedure as recommended by Clauset et al. (2007).

Two quantities are computed with the KS-test: the so-called D-statistic and a p-value indicating how probable it is to have found such an extreme D-statistics as we did (assuming that the data in fact did come from the hypothesised distribution). The D-statistic quantifies the deviation of the empirical data from our hypothesised distribution. The p-value states whether the deviation as measured by the D-statistic is large enough to reject the hypothesis that the empirical data was sampled from our hypothesised distribution. If this deviation is sufficiently large, we obtain a small p-value, and if the deviation is small, we obtain a large p-value. Hence, if we find a large p-value, the data might be—we are never certain—from the hypothesised distribution. On the other hand, if our p-value is small, we should reject this hypothesis, and conclude that the sample was not drawn from the hypothesised distribution.

More formally, the D-statistic is defined as

$$D = \max |S(x) - P(x)|,$$

where $P(x)$ is the cumulative distribution function of the hypothesized distribution and $S(x)$ is the empirical cumulative distribution function. In order to find out whether such a result is probable or not, we need to calculate the p-value. Formulas for calculating a p-value are available, but only if the parameters of the hypothesized distribution are known instead of estimated. So I will use the method outlined by Clauset et al. (2007). That is, we generate some large number of samples of the same size as the empirical sample. Then, for each generated sample we calculate the best fit and the D-statistic. We subsequently check what fraction of the generated samples had a D-statistic as extreme as ours. That fraction is our p-value.

In order to generate these samples, we need a method to obtain a random number from the hypothesised distribution. The following method was used for obtaining a random sample. Most programming languages and statistical programs have a uniform distribution random number generator on the interval $[0, 1]$ available. Now if k is a random variable distributed according to a probability density function $p(k)$, then $P(k) = r$, where $P(k)$ is the cumulative distribution function associated with $p(k)$ and r is uniformly distributed on the $[0, 1]$ interval. This is known as the integral probability transform. Solving for k then gives a random number from the distribution, or $k = P^{-1}(r)$. For the Dirac quality distribution we obtain

$$k = \frac{1}{\lambda} m r^{-\lambda} (r^{\lambda} - 1)(\lambda - 1)$$

Unfortunately, we cannot obtain similar results for generating random numbers for the other distributions. This does not prevent us completely from generating such random numbers however. We could generate a random number r and numerically find k such that $P(k) = r$. This number k is then a random number drawn from our distribution. It should be noted however, that this is computationally intensive, and thus prevents us from obtaining a random sample in most practical settings for the uniform and exponential distribution.

Summarising, we need to do two things: estimate λ and perform the KS-test. In order to estimate λ we need to numerically maximise the (log) likelihood function. For the KS-test we need to generate a large number of samples (in the order of 10,000). These samples require us to generate a large number of random numbers, depending on the empirical sample size. So in total, quite some computational power is required.

Chapter 4

Empirical Analysis

For the empirical analysis, a number of questions need to be addressed. Firstly, is the assumption of 'preferential attachment' satisfied? Secondly, are items indeed being added at a steady rate? Thirdly, what does the underlying quality distribution look like? Finally, do the popularity distributions fit the data well?

Sampling was based on the notion of 'related' movies. YouTube provides a mechanism to see what movies are related[1] to each other. I started out with the movies from the top viewed, top rated and recently featured list. This was done, so that I at least have movies in the tail of the distribution. The recently featured list was taken in, so that movies that were less popular were not left out. Unfortunately, this beclouds the possibility to say the sample is fully random. So, I have no way to guarantee randomness, but I will assume that the sample provides a relatively good impression of the population.

In order to provide an answer to the first question, I have sampled data from YouTube at two times. First I collected a sample at 23 December 2007, and one week later I revisited these movies, in order to obtain an estimate of how many views each movie attracted in the meantime. In total I collected valid information on 200,201 movies[2]. The information collected was the age (difference between the time it was uploaded and the time I retrieved the information), the rating and the view count of the movie. And, as stated, one week later I again collected the view count information which provides an estimate of how many views the movies attracted in the meantime.

The first collection of the data took about two days, and the second collection of the data took about one day. Some request gave bad data, which I subsequently threw out, but I do not expect this to bias the sample, since these errors seemed to be randomly distributed. For technical information on how I retrieved the data see appendix A.

[1] Full details of how movies are related are undisclosed. It probably depends on commonalities in tags, viewed by the same person, et cetera...

[2] Data are available from the author upon request

4.1 Assumptions

4.1.1 Preferential Attachment

The number of views added in a week are expected to be proportional to k. Let k be the view count, and Δk be the number of video's added over a period of Δt, about a week. A logarithmic regression ($r^2 = 0.6453$) suggests that $\Delta k \approx 0.0281 k^{0.78}$. From equation 2.1 we obtain that $\Delta k_i \approx (1-\lambda)\phi_i + \lambda k_i$ where ϕ_i is some normalised quality. For $\lambda = 0$, we would expect an exponent of 0, and for $\lambda = 1$ an exponent of 1. An exponent of 0.78 is then a quite plausible result. The results can be seen in figure 4.1.

Some amount of preferential attachment is present, although it still shows quite some variation, which might be due to the intrinsic qualities of the items. The time frame was comparatively long. After one day, the number of views have already increased. This has an impact on the probability of obtaining more votes. Hence, the time frame should be small enough to measure preferential attachment accurately, but it should be long enough to allow items to obtain additional votes. More accurate results might have been obtained if the time frame was a little shorter.

Having sampled YouTube movies at multiple times, Cheng et al. (2007) can more specifically determine whether indeed the growth trend follows approximately equation 2.4. They report a sublinear growth of the kind $k \sim t^\alpha$ where $\alpha < 1$ for most movies. Hence, the growth flattens off, as predicted in equation 2.4. This confirms that the preferential attachment assumption is not entirely unrealistic.

The solution from equation 2.4 seems to be confirmed as well by Johansen (2001; 2004; 2005). After Johansen was interviewed, and thus had some media exposure, he measured the number of downloads on each day. The cumulative number of downloads is found to fit the model $k_i = (1-b)^{-1}(t/t_i)^{1-b} + ct$. In general Johansen suggests that the download rate follows a power law $\partial k_i/\partial t \sim t^{-b}$. This corresponds well with the solution as stated in equation 2.4.

It is suggested by Chessa and Murre (2004) however that an exponentially decaying function better fits the data collected by themselves as well as the data collected by Johansen. Their results are based on a cognitive memory theory, wherein response times roughly vary with individual memory duration. It seems more appropriate however to model response times in a social context. Not everybody is exposed to an event at the same time, and the variance in response times is probably not due solely to individual memory duration. More likely, only some people are exposed a certain event (or the information regarding the event), and these people spread the information throughout the network. Chessa and Murre investigate whether such propagation through a social network is present, and indeed find such behaviour for their data. The better fit of the exponentially decaying function is disputed by Johansen (2005) who suggests that the cumulative downloads are better fitted by his model.

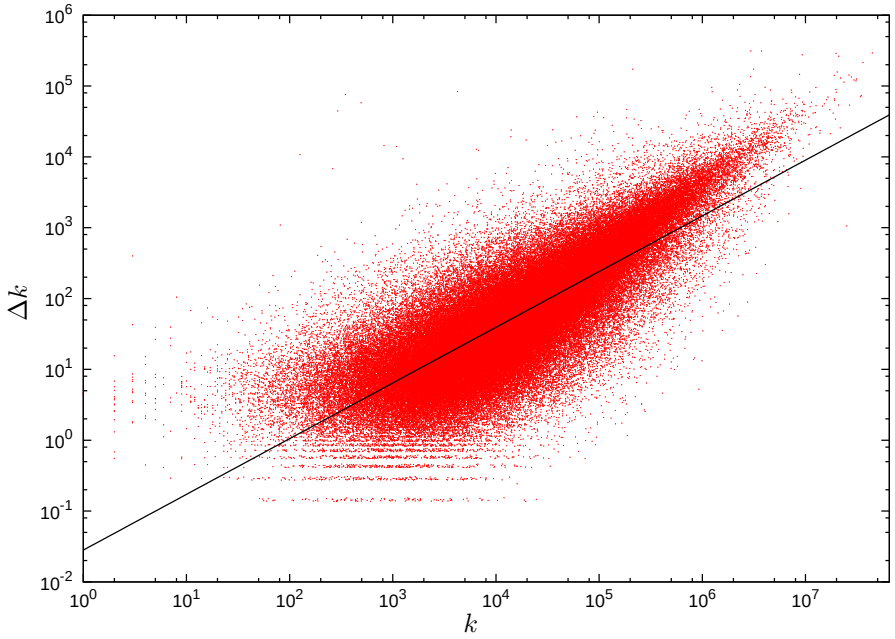

Figure 4.1: The average number of views received daily versus the view count. Logarithmic regression ($r^2 = 0.6453$) suggests $\Delta k \approx 0.0281 k^{0.78}$ which is displayed as the solid black line.

4.1.2 Uniform Introduction Rate

In order to answer the second question we can simply plot the time of introduction of the items in our sample, which can be seen in figure 4.2. It shows the time of introduction of the items in days, versus the rank of the items in the sample. If items were to be introduced at a uniform rate, then between any two items having a similar difference in age we should see approximately a similar difference in rank, or simply put, a linear trend[3].

Figure 4.2 seems to show such behaviour in the cumulative number of movies, at least from $t \approx 400$ onwards. This makes sense since the sample was drawn at 23 December 2007, and YouTube started at 15 February 2005, which is a difference of just over a thousand days. So, presumably, the first year things needed to get going, and we can see the rate at which items are being added is increasing from time 0 until 400, and after that remains relatively stable.

The rate at which movies are being added seems to be going slightly down after time $t \approx 700$, whilst the increasing popularity of YouTube would suggest an increasing rate. This could be explained by the fact that I collected data

[3]In fact the empirical cumulative distribution is displayed. If the items are being added in at a uniform rate, or $p = 1/b$ where $b \approx 1000$ is the maximum age, the cumulative distribution is $P(x) = \int_0^x 1/b \, \mathrm{d}t = x/b$ for $0 \leq x \leq b$, hence is linear.

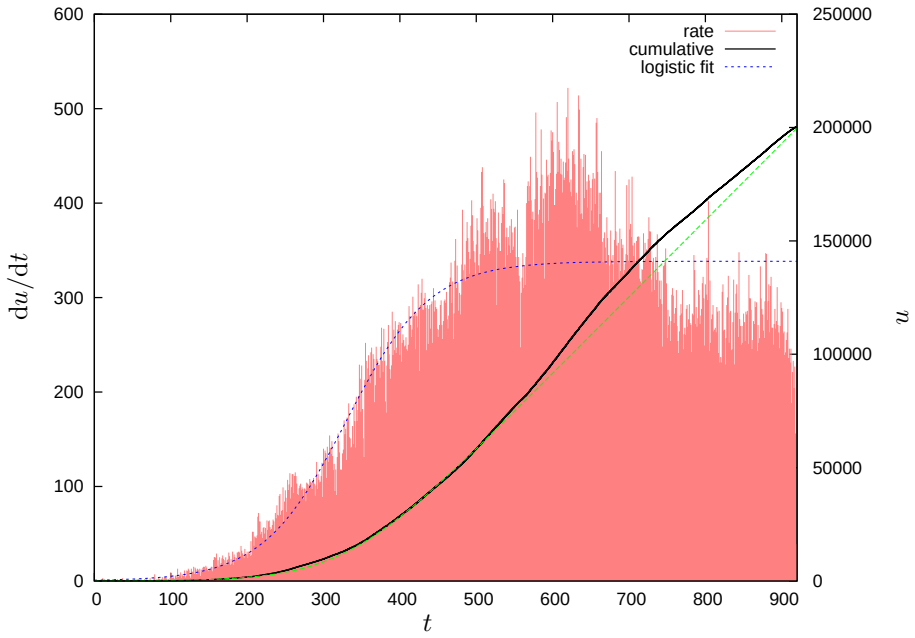

Figure 4.2: New items seem to be added at an increasing rate per day $\mathrm{d}u/\mathrm{d}t$, which is shown in the histogram using the left y-axis. The solid line displays the cumulative number of items u added before a certain time using the right y-axis. The solid black line shows the cumulative distribution, which almost seems to follow a linear trend from $t \approx 400$ onwards, suggesting a steady rate. The daily rate is fitted to a logistic curve $\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{a}{1+b^{-ct}}$ with estimates $a \approx 338$, $b \approx 412$ and $c \approx 0.0183$. The green line shows the integral of the logistic curve, adapted so that it equals zero at $t = 0$. The drop-off in the last 300 days or so might be due to the sampling method.

through the concept of 'related movies'. Possibly, more recently introduced movies are slightly worse connected in the network of related movies, and thus have a lower probability of turning up in the sample. The change in rate seems to be rather small however.

When we look at the movies added per day in figure 4.2 however, the rate seems to be increasing from time $t = 0$ until $t \approx 400$. This is in agreement with results from Cheng et al. (2007). They fit a power law to the rate at which movies are uploaded. They expect the rate at which movies are uploaded $\mathrm{d}u/\mathrm{d}t$ per day to increase with time t as $\mathrm{d}u/\mathrm{d}t = t^{\alpha}$ where $\alpha \approx 1.91$. The number of total movies u is thus $t^{\alpha+1}/\alpha$, or $u \sim t^{2.91}$.

As there is an upper limit to the number of people actively involved—all the people on the planet at most in any case—there is probably an upper limit for the rate at which movies are added. So, a power law does not seem very realistic for the rate of introduction. The rate of introduction will more likely

follow some sort of logistic curve, which follows largely an S shape. I have fitted a logistic curve of the form

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{a}{1 + be^{-ct}}$$

which approaches 0 if $t \to -\infty$ and approaches a if $t \to \infty$. So, a can be interpreted as the upper limit of the rate of introduction. The parameters b and c indicate how fast the growth is. The estimates of the parameters are $a \approx 338$, $b \approx 412$ and $c \approx 0.0183$ with $r^2 \approx 0.889$ and is shown in figure 4.2. According to the logistics function, the rate will thus stabilise at approximately 338 movies a day[4]. Of course, with the slight decrease in rate in mind after $t \approx 700$, this is no conclusive evidence whether the rate actually has stabilised.

Even though Cheng et al. (2007) fit a power law function, which keeps on increasing, it seems more appropriate to fit a logistic function, where the rate of introduction eventually stabilises. With an r^2 of 0.889 the logistic function performs quite well.

In summary, a uniform introduction rate is not apparent from the start, but it may show such behaviour after $t \approx 400$. The evidence seems to be inconclusive however. Perhaps this assumption is violated for the YouTube market, but this is uncertain.

4.1.3 Quality Distribution

The third question concerns the underlying quality distribution. Unfortunately, this is quite difficult to estimate. At first glance, the ratings might provide such an estimate, but this seems unreliable. Probably a lot of people rate a movie only if they actually like it. When people dislike the movie, or are just indifferent to it, they are presumable more likely to close the movie, than to actually rate it.

Furthermore, we could wonder whether 'intrinsic quality' is actually what people have in mind when they rate a movie. So, the rating could very well measure something different than what we would like to. In more technical terms, it might not be a valid indicator.

The cumulative distribution however is displayed in figure 4.3. It can be seen from this illustration that most of the ratings are actually below 4. This is consistent with findings from Gill et al. (2007).

The best solution in order to obtain a relatively autonomous estimate of the 'intrinsic quality' is to rule out any social influence. Of course, this is not easily done, since the view count does not only play a role on the website itself (it being displayed there), but it also plays a role in the contacts people have with each other. What it will probably do, at the very least, is to diminish social influence, since preferences can then only be transferred through social contacts.

[4]It should be kept in mind that I cannot provide an estimate of the rate at which movies are being uploaded, since this estimate would increase with sample size. The figure of 338 can therefore only be interpreted within the context of this thesis, and has no ramifications for other studies.

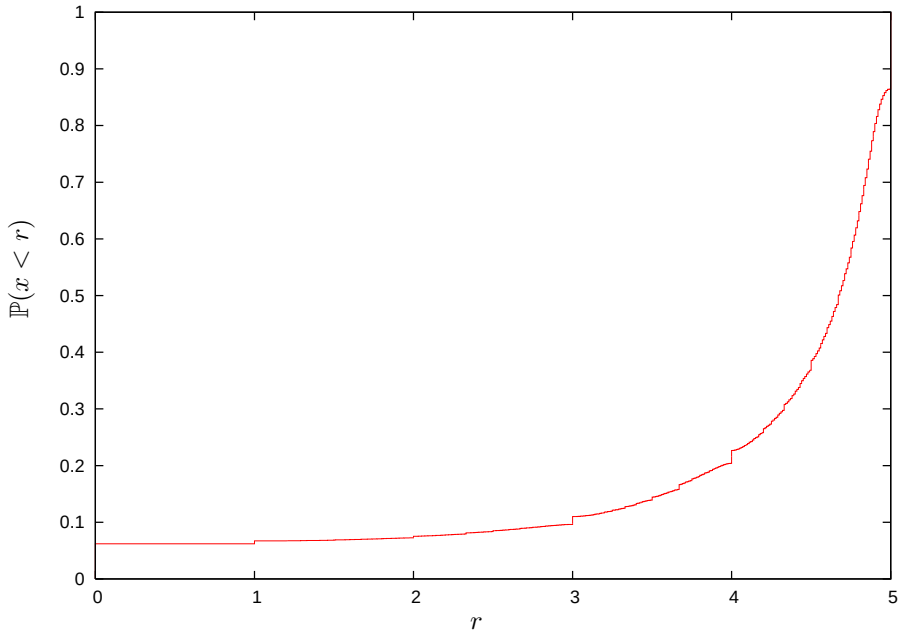

Figure 4.3: Empirical cumulative distribution function of the ratings r of the sample from YouTube. As can be seen, about 80% of the ratings are between 4 and 5, where 5 is the highest rating that can be given.

Data of this kind were actually collected by Salganik et al. (2006) in their experimental study of inequality in a cultural market[5]. The independent setting of their experiment has the lowest social influence. It does not report the number of downloads or the rating, and social influence is thus limited to social contacts. Of course, social influence may manifest itself in this way. Still it will be a relatively 'clean' environment. We could look at their data, and try to estimate a quality distribution. Unfortunately, at the time of writing, the data are not yet available.

So, as a quality distribution is not available to us empirically, I have analysed the model for all three quality distributions.

4.2 Model

There are two parameters to be estimated for the various models: the mean number of votes/views m and social influence λ. Because of computational limits I used a smaller sample of 9,997 for these procedures. The estimated average number of views is $\hat{m} \approx 339{,}548$. Since m does not depend on the

[5] We introduce their research shortly on page 23.

underlying quality distribution, I will use this estimate for all distributions.

It should be observed that our theoretical distributions are stationary distributions. That is, after a while they should change little. So we would also expect the distribution from YouTube to approximate this stationary distribution. Hence, regardless of the time at which we draw a sample the results should approximately be similar. This behaviour is actually observed by Cheng et al. (2007) who drew samples from YouTube at various instances in time.

It should be taken into account that the empirical standard error of the mean still is about 16,901. When social influence $\lambda \geq 1/2$ variance goes to infinity, and so will the standard error. Of course, a finite sample with finite values will always produce a finite estimate of the variance. But if social influence is indeed higher than 1/2, the estimated mean $\hat{m}$ can vary substantially from one sample to the next.

We ran the procedure for estimating λ for each distribution (Dirac, uniform and exponential). The estimates for the data collected from YouTube are displayed in table 4.1. The total social influence is estimated to be within the range of 88–94%. The theoretical and actual YouTube results can be seen in figure 4.4. It can be seen that for more skewed quality distributions, the social influence indeed does decline. Social influence remains high however.

This means that—if the model is correct of course—relatively much of the actual view count is accounted for by how many views a movie already got. Quality thus seems to be relatively unimportant. The amount of social influence is lower for more heterogeneous quality distributions, reaching 87.8% for the exponential case.

In Gill et al. (2007) the view count is reported to follow a power law, which they fitted with an exponent[6] of 2.79. Since all models follow a power law asymptotically $k^{-(1+1/\lambda)}$, this yields an estimate of social influence of 0.56. The data analysed by Gill et al. (2007) however is gathered only locally on a campus, so will probably differ from the global distribution. Their analysis still seems to indicate that also in local communities some social influence is present.

The presence of social influence is also supported by the difference between the growth in the unique number of movies viewed by the users on the campus and the total number of movies viewed on the campus. The unique number of movies viewed increases much slower than the total number of movies viewed over time. This indicates that movies are being recommended between friends, thus resulting in the slower increase in the unique number of movies viewed.

Next to the analysis of YouTube data, I also analysed data from the Hollywood movie industry. This makes possible not only an additional empirical analysis, but also a comparison between a traditional and an on-line market.

[6]They plot the rank R against the view count F, and fit a Pareto distribution, such that $F \sim R^{-\beta}$ using least squares regression analysis. So, $R \sim F^{-1/\beta}$. Since the rank is simply the number of items that are larger than F, the derivative of this gives the probability density function, hence $F^{-(1+1/\beta)}$. A good explanation of these relationships can be found at http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html. Estimating the exponent by least squares regression analysis has serious bias, and is therefore not recommended (Clauset et al. 2007). The exponent should be interpreted with caution.

Dataset	Distribution	$\hat{\lambda}$	Std. Error
Youtube	Dirac	0.938	$1.263 \cdot 10^{-3}$
	Uniform	0.908	$2.063 \cdot 10^{-3}$
	Exponential	0.878	$2.878 \cdot 10^{-3}$
Hollywood	Dirac	0.843	$6.916 \cdot 10^{-3}$
	Uniform	0.740	–
	Exponential	0.663	$1.726 \cdot 10^{-2}$

Table 4.1: Overview of the estimates for the various models for the YouTube data set and for the Hollywood movies data set. Only items with $k > 0$ were evaluated.

The Hollywood data was collected from the website *The Movie Times*[7]. I have not performed checks on the assumptions as I did for the YouTube case though.

The website reports gross income, which I use as an indicator for popularity. Although I would rather have the actual number of visitors, the gross income will be quite a good indicator for popularity. The gross income is reported in millions of dollars[8] in the USA.

We collected data for the years 2000 until 2007, totalling to $n = 2615$ movies. The minimum reported gross income is \$0.005 million dollars or about $5,000$ dollar for the movie *The Intruder*, which appeared in only one theatre. This contrasts with the maximum reported gross income of about \$438 million for the movie *Shrek 2*, which appeared in 4223 theatres. The mean gross income is $m \approx 26.97$ million dollars with an empirical standard error of about 0.98.

Now we can compare the relative volatility of the two markets. According to the the MPAA[9] the average ticket price between 2000 and 2007 is about \$6.11. As a rough estimate, the average number of views per movie is $\hat{m}/6.11 \approx$ 4,414,275. This yields a volatility[10] of $\hat{\nu} = 1/\hat{m} \approx 2.26 \cdot 10^{-7}$. The volatility for YouTube is $\hat{\nu} \approx 2.94 \cdot 10^{-6}$. Although it is relatively hard to interpret this figure on itself, we can at least compare the two. It indicates that YouTube is in the order of 10 times more volatile than the traditional Hollywood market.

Of course, there are several uncertainties with this comparison. First, the mean $\hat{m}$ is expected to vary substantially from sample to sample. Secondly, I only used an average ticket price to calculate the number of views for the Hollywood market. Thirdly, due to the sampling methods, I might have missed out on some of the lesser viewed movies for both YouTube and Hollywood. Including these less popular movies would lead to a lower estimate of both means (and thus a more volatile market). Still, with YouTube being more than 10 times more volatile than Hollywood, the difference is quite striking.

The estimate $\hat{\lambda}$ for the social influence is also substantially lower for the

[7] http://www.the-movie-times.com

[8] The scale of measure for popularity does not make a difference for the estimate of λ. Stated otherwise, any linear transformation of the gross income (multiply or divide by a constant) does not change the estimate.

[9] Motion Picture Association of America, see http://www.mpaa.org

[10] See section 2.3.1 on page 27 for details on volatility.

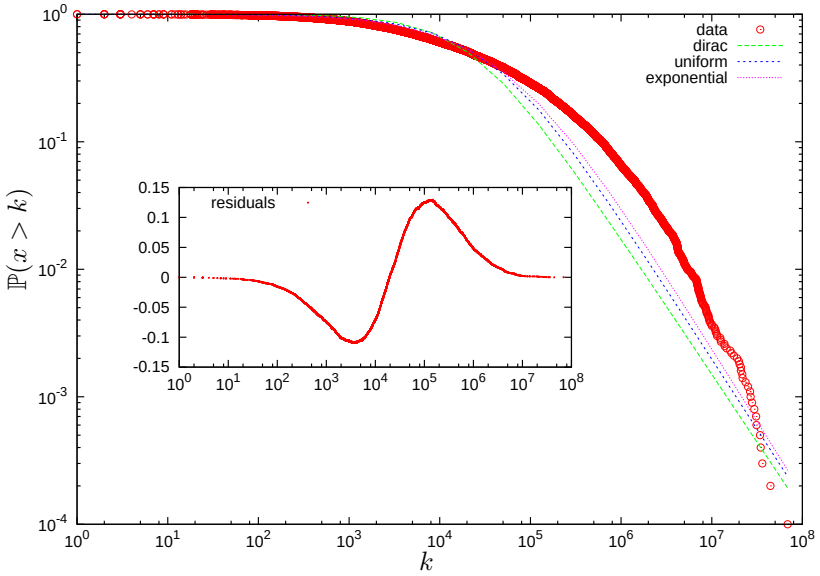

Figure 4.4: Empirical and theoretical results for various distributions for the YouTube market. The fitted parameters can be seen in table 4.1. The inset shows the residuals (i.e. the difference between the actual and predicted values) for the Dirac distribution, but note that only the x-axis is on a logarithmic scale. The residuals for the other distributions are alike.

Hollywood market than for the YouTube market, which can be seen in table 4.1. The results for the Hollywood market are displayed in figure 4.5. The estimate of social influence ranges from 84.3% for the Dirac quality distribution to 66.3% for the exponential distribution.

Since the asymptotic sampling distribution of any MLE is normally distributed, we can easily test whether the difference in social influence is significant. The difference between the YouTube and the Hollywood estimates of social influence has a standard error of $\sqrt{(1.762 \cdot 10^{-2})^2 + (2.878 \cdot 10^{-3})^2} \approx 0.0174983$ assuming an exponential quality distribution. The actual difference is 0.245, which is about 14 times the standard error of the difference, which implies a one-sided p-value in the order of $7 \cdot 10^{-45}$, which is highly significant. Similarly for the estimates of social influence assuming a Dirac quality distribution, we obtain a one-sided p-value of about $6 \cdot 10^{-42}$. We thus have statistical confirmation, that the social influence is significantly higher for the YouTube market than the Hollywood market.

The YouTube market shows an inequality Gini coefficient of 0.881, which is somewhat lower than a coefficient of 0.942 what would have been predicted based on my analysis for the Dirac quality distribution (cf. equation 2.15). The Hollywood market also shows a somewhat lower Gini coefficient of 0.744 than the prediction of 0.864. Hollywood does score lower than YouTube though.

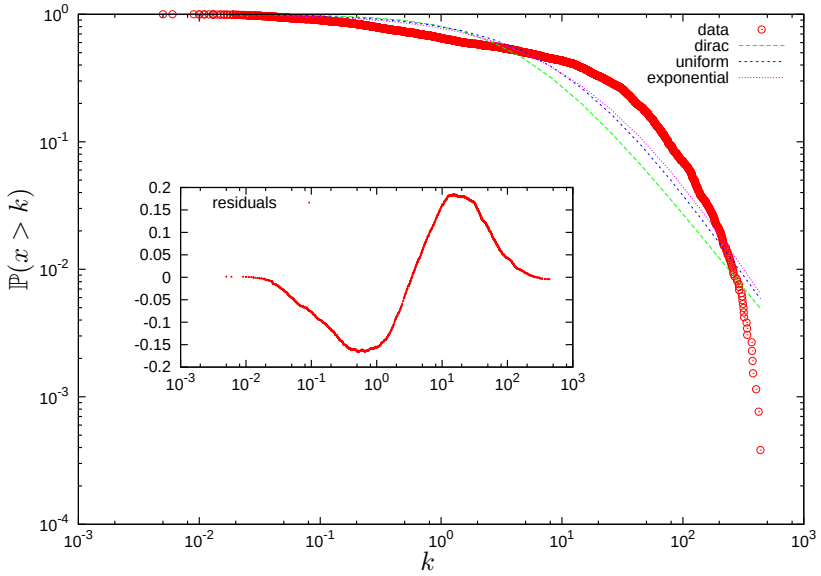

Figure 4.5: Empirical and theoretical results for various distributions for the Hollywood market. The fitted parameters can be seen in table 4.1. Observe that k is in millions of dollars. The inset shows the residuals (i.e. the difference between the actual and predicted values) for the Dirac distribution, but note that only the x-axis is on a logarithmic scale. The residuals for the other distributions are alike.

Hence, YouTube shows more inequality than Hollywood, in congruence with my predictions.

In the insets of figures 4.4 and 4.5 the deviation of the data from the theoretical distributions can be seen. The inset shows that for low k the theoretical model tends to overestimate the actual probabilities, and for somewhat higher k the theoretical model underestimates the actual probabilities. More specifically, the tails of the distributions do not seem to follow a power law. This clearly deviates from the theoretical distributions. Other distributions might provide a closer fit.

So, I also analysed the log-normal distribution. The parameters are estimated to be $\hat{\mu} = 9.95$ and $\hat{\sigma} = 2.56$ for the YouTube data and $\hat{\mu} = 1.27$ and $\hat{\sigma} = 2.55$ for the Hollywood data set. With a p-value of about[11] 0.506 for the KS-test, the result for the YouTube data is not significant, which indicates that the sample might be drawn from a log-normal distribution. The p-value for the Hollywood data set is about 0.505 which suggests that this sample also might have been drawn from a log-normal distribution.

The KS-test was only performed for the Dirac quality distribution, and according to the KS-test neither the YouTube sample nor the Hollywood sample

[11] Estimated using 10,000 samples.

was drawn from this distribution (both resulted in a p-value of almost 0).

To summarise, the Dirac model does not fit the data as good as a lognormal distribution. Assuming different quality distributions might provide a better fit, although I do not expect that. But since I couldn't perform the KS-test for the other distributions because of computational limits, I remain somewhat inconclusive on this point. Still, the distributions seem to capture the qualitative behaviour of both markets reasonably well, and is in agreement with the experiments from Salganik et al. (2006). Until some theoretical foundation is found for the log-normal distribution, incorporating social influence, and showing results in agreement with Salganik et al. (2006), the suggested distributions have a better theoretical foundation.

Chapter 5

Conclusion

The model considered in this thesis is based on a 'rich-get-richer' effect. Basically, this means that popular items tend to become increasingly popular. However, I allow for quality of items to play a role in popularity. The model thus incorporates both a 'rich-get-richer' as well as a 'good-get-richer' effect. The balance between these two effects is expressed as the amount of social influence. With more social influence, the first effect becomes more important, and with less social influence, the latter effect becomes more important.

The main findings can be summarised in four brief points:

- Uncertainty rises with quality ϕ and social influence λ.
- Inequality rises with social influence λ.
- Popularity follows, in theory, a power law of the form $\mathbb{P}(k) \sim k^{-(1+1/\lambda)}$ asymptotically. A lognormal distribution is a better fit for empirical data however.
- On-line markets seem more volatile, and show more social influence than traditional markets.

The last point is a broader interpretation of the difference between the YouTube and the Hollywood market. The first two points are confirmed empirically by Salganik et al. (2006), and have consequences for producers of books, songs and movies. The model considered here might serve as a theoretical justification of the results found by Salganik et al.. Based on a simple assumption it shows qualitatively similar behaviour, and ranges from an exponential distribution to a power law distribution, smoothly interpolating between the two, depending on the amount of social influence.

The model considered here also has a relevance for growing networks. It can be used as an interpolation between the two extremes suggested by Albert and Barabasi. The model considered here also might serve as an alternative to the competitive fitness model of Bianconi and Barabasi (2000). The model

of Pennock et al. (2002) is the same if we assume a Dirac quality distribution. Because of the social influence λ, vertices can attract edges at a higher rate if their quality is higher. This might lead to a network topology different from the ones generated by other models. Properties such as the clustering coefficient, average path length and the size of the largest connected component are of interest. These topics might be investigated in future research, and are not yet analysed by Pennock et al. (2002).

The theoretical distribution of popularity is not confirmed by the data however, and a lognormal distribution provides a better fit. One possibility for obtaining a lognormal distribution is through multiplicative growth (see Mitzenmacher (2003) for example). Incorporating the idea of a social influence parameter λ into such a model might yield a qualitatively similar result (increasing uncertainty and inequality with social influence), yet a better fit to the empirical data. Whether that is the case needs to be investigated.

We find a substantial difference in both volatility and social influence between YouTube and Hollywood. The YouTube market has more social influence and is more volatile than the Hollywood market. The difference might suggest a broader distinction between on-line markets and traditional markets. It would be interesting to see how other on-line markets such as Amazon.com[1], iTunes[2], Flickr[3] and SourceForge[4] would fit to this model. It would be especially interesting to compare them to their traditional counterparts, to see whether on-line markets in general are indeed more volatile and have a higher social influence.

The ease and speed of on-line communication—not only via e-mail or chatting, but also through the reported number of views—is probably partly responsible for the high social influence for the YouTube market. This sets it apart from the Hollywood movies, where individual preferences and quality seem to be somewhat more prevalent, although social influence is still quite high.

The prevalance of social influence contrasts with the interpretation given by Anderson (2007) in the popular business book *The Long Tail*. He suggests that individual users have thousands of 'niche' items available to them in on-line markets such as YouTube, iTunes or Amazon, which begets the possibility for unique preferences and taste. So people can express themselves more individually than ever before. Or so Anderson suggests.

The analysis done here, however, suggests the opposite. Indeed, the markets available on-line are far larger and more diverse than any traditional bookstore, music store or movie theatre. But internet users do not seem to follow their own preferences or individual taste, but are mostly guided by the choices of others. The internet does not seem to increase individualism, but to increase herding behaviour.

Still, the sheer amount of choice makes it *possible* for users to follow their

[1] http://www.amazon.com
[2] http://www.apple.com/itunes/
[3] http://www.flickr.com
[4] http://sourceforge.net/, see Hunt and Johnson (2002) for a quick overview of the download distribution. It is reported to follow a power law, but this is based on an analysis of only 30 days.

own choice, but it seems only few actually do. So when Anderson claims that 'going on-line' will "transform entire industries—and the culture—for decades to come" (Anderson 2007:26) he might have a point. But it is higher social influence and higher volatility, rather than a materialising wider range of choice as Anderson believes, that might warrant such a dramatic claim.

It is worth investigating why on-line settings might have a higher social influence. One of the reasons, no doubt, is that the number of downloads, views or sales is often presented on the webpage. That figure presents other users with the idea that a certain item is more popular than others. It informs them of the choices others have made. Secondly, fora and commentary inform users of what others think about an item. Thirdly, people can easily refer their friends or colleagues to a movie they've just seen, or a song they've just heard. In traditional markets this depends much more on face-to-face contact. What the magnitude of the effects of these various principles are, makes an interesting research topic.

When markets go on-line, producers should be prepared to take the increase in risk into account. Higher social influence produces higher inequality and higher uncertainty. As would be expected by most people, the first few views or sales thus might trigger a reinforcing process of increasing popularity. Some books might break all records, while others remain on the shelf, and it becomes harder to predict which books that will be.

Appendix A

Data Collection

We collected the YouTube data in two stages. First we collected data on 23 December 2007, and then on 30 December 2007 once again. This way we could estimate whether popular movies would have a higher download rate.

A.1 Technical Information

Collection of the data was done through the YouTube API[1]. They provide an XML response based on an HTTP request. For example, we send a regular HTTP/1.1 request to

```
http://gdata.youtube.com/feeds/api/videos/dMH0bHeiRNg
```

where `dMH0bHeiRNg` is the identification string of a YouTube movie. This identification string can consist of the characters `[A-Z]`, `[a-z]`, `[0-9]` and `_`. We then get an XML response which looks roughly as following:

```
<entry>
    <id>http://gdata.youtube.com/.../dMH0bHeiRNg</id>
    ...
    <published>2006-04-06T14:30:53.000-07:00</published>
    ...
    <yt:statistics viewCount="72109974"/>
    <gd:rating ... average="4.65"/>
    ...
</entry>
```

where we have included only the entries relevant to our inquiry. The related movies can be retrieved through a request to

```
http://gdata.youtube.com/.../dMH0bHeiRNg/related
```

[1] Application Programming Interface. This means that we have a number of functions available which we can query for information.

with a response similar to

```
<feed>
    <id>
        http://gdata.youtube.com/.../dMH0bHeiRNg/related
    </id>
    ...
    <entry>
        <id>
            http://gdata.youtube.com/.../QjA5faZF1A8
        </id>
    </entry>
    <entry>
        <id>
            http://gdata.youtube.com/.../vr3x_RRJdd4
        </id>
    <entry>
    ...
    <entry>
        <id>
            http://gdata.youtube.com/.../NI17GctQfWM
        </id>
    </entry>
</feed>
```

These related movies were then added to the queue. We picked items in front of the queue with multiple threads. So, information for multiple movies and related movies could be obtained simultaneously. This speeded the process up, and increased the rate of movie information retrieval from around 0.7 items per second to about 2 items per second. This program[2] was written in C#.

One week later we ran another program, this time revisiting the movies which we downloaded a week before. This is simply a replication of the first process, but then the queue is filled with the movies which we already obtained, instead of the queue being filled 'on the fly'. We also programmed this in C#.

A.2 Data set

The data set consist of eight columns in total

```
movieID viewcount rating age viewcount2 rating2 age2 differ
```

where the second variables (`viewcount2`, `rating2` and `age2`) contain the information from our second information retrieval. The age is a floating point number indicating the number of days between the date it was uploaded and the date we retrieved the information. The rating can vary from 0 to 5 and the

[2]Source code of the program is available upon request.

view count is simply the number of times the movie was viewed. The variable `differ` is simply the average number of downloads per day in between. So

```
differ = (viewcount - viewcount2)/(age2 - age)
```

This data is stored in a standard ASCII text file, so that it can be loaded into various programs, such as `R` for statistical analysis and `GNUPlot` for plotting. In total we collected information on $n = 200{,}201$ movies.

The data on Hollywood movies was collected from the website *The Movie Times*[3]. The data was transformed to a standard ASCII text file containing the title of the movie, the opening revenue, the total revenue, the theatres in which it played, the number of weeks in the top 60 and the studio. This totalled to $n = 2615$ movies.

[3] http://www.the-movie-times.com

Appendix B

Numerical Computations

Most of the equations considered here are quite complex. That is, we need to evaluate the incomplete Beta function B with non standard parameters and the confluent hypergeometric function of the second kind U. These are just the functions relevant to the probability density function (pdf), while we also need the cumulative distribution function (cdf) for the KS test, which involves other complex functions. Then, in order to estimate our parameters, we need to evaluate the log likelihood, and numerically obtain a maximum. Most of this work was done in the statistical program `R`.

B.1 Function Evaluation

In `R` we have available the package `gsl`, which is the GNU Scientific Library. In that package, several functions are available, among which are the incomplete Beta function and the hypergeometric function.

B.1.1 Incomplete Beta function

Unfortunately, the incomplete Beta function $B(x, a, b)$ only accepts parameters for which $0 \leq x \leq 1$, and in our case $x \leq 0$. So this function cannot be used straigth away. But, the incomplete beta function may also be written as

$$B(x, a, b) = a^{-1} x^a {}_2F_1(a, 1-b, a+1, x),$$

where ${}_2F_1$ is Gauss's hypergeometric equation(Abramowitz and Stegun 1970:eq. 6.6.8). The ${}_2F_1$ function is available from the `gsl` package as `hyperg_2F1`, which converges only for $|x| < 1$, or for $k > 2m(\lambda-1)/\lambda$. So, in order to obtain results for $k \leq 2m(\lambda-1)/\lambda$ we numerically integrated

$$\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t,$$

with the `myintegrate` function of the `elliptic` package, which can handle integration over a complex domain.

B.1.2 Hypergeometric Equation

The hypergeometric equation

$$U(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{b-a-1} \mathrm{d}t,$$

is implemented as `hyperg_U` in the `gsl` package. Again, evaluation is limited, and for too small z, this function does not compute. Since $b = 1$, we can use the approximation (Abramowitz and Stegun 1970:eq. 13.5.9)

$$U(a,1,z) = -\frac{1}{\Gamma(a)} \left[\log(z) + \psi(a)\right] + \mathcal{O}(|z \log z|),$$

where $\psi(a)$ is the digamma function which is the derivative of the logarithm of the gamma function

$$\psi(a) = \frac{\mathrm{d}}{\mathrm{d}a} \log \Gamma(a).$$

B.2 Numerical Maximisation

`R` provides a standard algorithm to maximise likelihood. This algorithm is the `mle` function. We used the algorithm for the boxed constraints developed by R. H. Byrd, P. Lu and Zhu (1995) and implemented in `mle`. This algorithm allows us to specify lower and upper boundaries for the parameters. We used this algorithm for estimating the parameters for all models. We used simplified log likelihood as stated in equations 3.1, 3.2 and 3.3 for maximising.

The biggest problem is that there's no guarantee that we have actually found a global minimum. So, we should interpret the estimated parameters with some caution.

B.3 Cumulative Distribution Functions

For all but the Dirac distribution we needed numerical integration. This was provided by the `integrate` function of the `stats` package. The cdf for the Dirac distribution can be written as

$$\mathbb{P}(K > k) = 1 - \left(\frac{m(1-\lambda)}{m(1-\lambda) + k\lambda} \right)^{\frac{1}{\lambda}},$$

but the other integrals cannot be simplified that much. Hence, we have numerically integrated the pdf's in order to obtain results for the cdf's.

Appendix C

Mathematics

Some more precise derivations of the results are given here. More specifically, we show how we got from the differential (equation 2.3) to the uncertainty distribution (equation 2.5) and how to derive the Lorenz curve and the Gini coefficient.

C.1 Differential Equation

We start off by taking equation 2.3, which is

$$\frac{\partial k_i}{\partial t} = \left[(1-\lambda)\frac{m\phi_i}{t\mu} + \lambda\frac{k_i}{t}\right]. \tag{C.1}$$

First we solve the homogeneous equation

$$\frac{\partial k_i}{\partial t} = \lambda\frac{k_i}{t},$$

we divide by k_i and multiply by ∂t and thus arrive at

$$\frac{1}{k_i}\partial k_i = \lambda\frac{1}{t}\partial t.$$

Now we integrate

$$\int \frac{1}{k_i}\mathrm{d}k_i = \lambda\int \frac{1}{t}\mathrm{d}t,$$

and obtain

$$\log k_i = \lambda \log t + C_0.$$

where C_0 is some constant. Taking the exponential we get

$$k_i = \exp(\lambda \log t + C_0) = t^\lambda C,$$

where $C = \exp C_0$ is some positive constant. This is the solution for the homogenous equation. Now instead of just having C as a constant, we take C to

be dependent on t and write $k_i = t^\lambda C(t)$. If we differentiate it should equal C.1. The derivative of k_i with respect to t is $\lambda t^{\lambda-1}C(t) + t^\lambda C'(t)$, so we obtain

$$\lambda t^{\lambda-1}C(t) + t^\lambda C'(t) = (1-\lambda)\frac{m\phi_i}{t\mu} + \lambda\frac{k_i}{t},$$

where $C'(t)$ denotes taking the derivative of $C(t)$ with respect to t. If we substitute in our solution $k_i = t^\lambda C(t)$ this becomes

$$\lambda t^{\lambda-1}C(t) + t^\lambda C'(t) = (1-\lambda)\frac{m\phi_i}{t\mu} + \lambda\frac{t^\lambda C(t)}{t},$$

or

$$\lambda t^{\lambda-1}C(t) + t^\lambda C'(t) = (1-\lambda)\frac{m\phi_i}{t\mu} + \lambda t^{\lambda-1}C(t).$$

Since both sides now have $\lambda t^{\lambda-1}C(t)$ we subtract this from both sides obtaining

$$t^\lambda C'(t) = (1-\lambda)\frac{m\phi_i}{t\mu},$$

solving for $C'(t)$ gives us

$$C'(t) = (1-\lambda)\frac{m\phi_i}{t^{\lambda+1}\mu},$$

which since $\int C'(t)\mathrm{d}t = C(t)$ we solve for $C(t)$ and hence integrate both sides, which results in

$$\begin{aligned} C(t) &= \int (1-\lambda)\frac{m\phi_i}{t^{\lambda+1}\mu} \\ &= (1-\lambda)\frac{m\phi_i}{\mu}\int t^{-(\lambda+1)}\mathrm{d}t \\ &= (1-\lambda)\frac{m\phi_i}{\mu}\frac{-t^{-\lambda}}{\lambda} + K, \end{aligned}$$

where K is some constant. Now we have obtained a solution for $C(t)$ which we can substitute in our equation $k_i = t^\lambda C(t)$ which becomes

$$\begin{aligned} k_i(t) &= t^\lambda\left[(1-\lambda)\frac{m\phi_i}{\mu}\frac{-t^{-\lambda}}{\lambda} + K\right] \\ &= (1-\lambda)\frac{-m\phi_i}{\mu\lambda} + Kt^\lambda. \end{aligned} \tag{C.2}$$

Since items are being introduced without any votes at their time of introduction t_i, we have $k_i(t_i) = 0$. We use this to solve for our constant K and obtain

$$(1-\lambda)\frac{-m\phi_i}{\mu\lambda} + Kt_i^\lambda = 0,$$

or

$$K = (1-\lambda)\frac{m\phi_i}{\mu\lambda t_i^{\lambda}},$$

which we substitute back into C.2 which results in our final solution

$$\begin{aligned} k_i(t) &= (1-\lambda)\frac{-m\phi_i}{\mu\lambda} + (1-\lambda)\frac{m\phi_i}{\mu\lambda t_i^{\lambda}} t^{\lambda} \\ &= (1-\lambda)\frac{-m\phi_i}{\mu\lambda} + (1-\lambda)\frac{m\phi_i}{\mu\lambda}\left(\frac{t}{t_i}\right)^{\lambda} \\ &= \left[\left(\frac{t}{t_i}\right)^{\lambda} - 1\right](1-\lambda)\frac{m\phi_i}{\mu\lambda} \end{aligned} \tag{C.3}$$

which yields equation 2.4.

C.2 Uncertainty Distribution

Let $X_{t,\phi}$ be the number of votes and $\tau_{t,\phi}$ be the time of introduction of a random item having quality ϕ after time t. So, after time t we draw from the population of items having quality ϕ one random item, and denote the number of votes of that item by $X_{t,\phi}$ and the time of introduction of that item by $X_{t,\phi}$. Then we are looking for the solution $\mathbb{P}(X_{t,\phi} < k)$.

Since we are considering items which have the same quality ϕ, all items grow in the same fashion. Items differ in the number of votes they have received, only because they were introduced at a different time. So the number of votes and the time of introduction are directly related to one another. Using solution C.3 we get that for an item that was introduced at time $\tau_{t,\phi}$ has

$$X_{t,\phi} = \left[\left(\frac{t}{\tau_{t,\phi}}\right)^{\lambda} - 1\right](1-\lambda)\frac{m\phi}{\mu\lambda}$$

votes. Hence we can rewrite the inequality as

$$\begin{aligned} X_{t,\phi} &< k \\ \left[\left(\frac{t}{\tau_{t,\phi}}\right)^{\lambda} - 1\right](1-\lambda)\frac{m\phi}{\mu\lambda} &< k \\ \left(\frac{t}{\tau_{t,\phi}}\right)^{\lambda} - 1 &< \frac{k\lambda\mu}{(1-\lambda)m\phi} \\ \left(\frac{t}{\tau_{t,\phi}}\right)^{\lambda} &< \frac{k\lambda\mu + (1-\lambda)m\phi}{(1-\lambda)m\phi} \\ \tau_{t,\phi}^{-\lambda} &< \frac{k\lambda\mu + (1-\lambda)m\phi}{(1-\lambda)m\phi} t^{-\lambda} \\ \tau_{t,\phi} &> \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}} t. \end{aligned} \tag{C.4}$$

Using this we can write

$$\mathbb{P}(X_{t,\phi} < k) = \mathbb{P}\left(\tau_{t,\phi} > \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}} t\right)$$

The introduction of new items happens at a uniform rate. That means after time t, items could have been introduced with equal probability from time 1 until t. After time t, $t+s$ items exist. The probability that an item with quality ϕ was introduced at a certain time is $1/(t+s)$ (proportional to the probability at which an introduced item has quality ϕ). The probability that an item was introduced before a time c, or $\mathbb{P}(\tau_{t,\phi} < c) = 1 - \frac{c}{t+s}$. We can thus write

$$\begin{aligned}\mathbb{P}(X_{t,\phi} < k) &= \mathbb{P}\left(\tau_{t,\phi} > \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}} t\right) \\ &= 1 - \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}} \frac{t}{t+s}.\end{aligned}$$

Now taking the limit $t \to \infty$ we obtain the stationary distribution

$$\lim_{t\to\infty} 1 - \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}} \frac{t}{t+s} = 1 - \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}},$$

since $\lim_{t\to\infty} \frac{t}{t+s} = 1$. This gives use $\mathbb{P}(X_\phi < k)$ where X_ϕ is the number of votes of a random item having quality ϕ after a long enough time period. Since this is the cumulative distribution function (cdf) and we would like to have the probability density function (pdf), we differentiate to k and get

$$\begin{aligned}&\frac{\partial}{\partial k}\left[1 - \left(\frac{(1-\lambda)m\phi}{k\lambda\mu + (1-\lambda)m\phi}\right)^{\frac{1}{\lambda}}\right] \\ &= -((1-\lambda)m\phi)^{\frac{1}{\lambda}}(-1/\lambda)(k\lambda\mu + (1-\lambda)m\phi)^{-1/\lambda-1}\lambda\mu \\ &= \mu((1-\lambda)m\phi)^{\frac{1}{\lambda}}(k\lambda\mu + (1-\lambda)m\phi)^{-1/\lambda-1},\end{aligned}$$

which gives equation 2.5.

C.3 Expectation

Let X_ϕ again be the number of votes for a random item having quality ϕ. The expected number of votes for a random item can be calculated as

$$\mathbb{E}(X_\phi) = \int_0^\infty k\mathbb{P}(X_\phi = k)\mathrm{d}k,$$

or

$$\begin{aligned}\mathbb{E}(X_\phi) &= \int_0^\infty k\mu((1-\lambda)m\phi)^{\frac{1}{\lambda}}(k\lambda\mu+(1-\lambda)m\phi)^{-1/\lambda-1}\mathrm{d}k \\ &= \mu((1-\lambda)m\phi)^{\frac{1}{\lambda}}\int_0^\infty k(k\lambda\mu+(1-\lambda)m\phi)^{-1/\lambda-1}\mathrm{d}k.\end{aligned}$$

The integral is solved through integration by parts. Using the substitution $v = k\lambda\mu$ we can write

$$\int (k\lambda\mu+(1-\lambda)m\phi)^{-\frac{1}{\lambda}-1}\mathrm{d}k = -\frac{(k\lambda\mu+(1-\lambda)m\phi)^{-\frac{1}{\lambda}}}{\mu}.$$

With integration of parts the integral thus becomes

$$\begin{aligned}\int_0^\infty k(k\lambda\mu+(1-\lambda)m\phi)^{-\frac{1}{\lambda}-1}\mathrm{d}k &= -\frac{k(k\lambda\mu+(1-\lambda)m\phi)^{-\frac{1}{\lambda}}}{\mu}\Bigg|_0^\infty \\ &\quad -\int_0^\infty \frac{(k\lambda\mu+(1-\lambda)m\phi)^{-\frac{1}{\lambda}}}{\mu}\mathrm{d}k,\end{aligned}$$

where the first part evaluates to 0. Evaluating the second part gives

$$-\frac{(k\lambda\mu+(1-\lambda)m\phi)^{-1/\lambda+1}}{\mu^2(\lambda-1)}\Bigg|_0^\infty,$$

using the substitution $v = k\lambda\mu$ again. Since $-1/\lambda+1<0$ for $0<\lambda<1$ this evaluates to

$$\frac{((1-\lambda)m\phi)^{-1/\lambda+1}}{\mu^2(1-\lambda)}.$$

Simplifying gives

$$\begin{aligned}&\frac{(1-\lambda)m\phi((1-\lambda)m\phi)^{-1/\lambda}}{\mu^2(1-\lambda)} \\ =\ &\frac{m\phi((1-\lambda)m\phi)^{-1/\lambda}}{\mu^2}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}(X_\phi) &= \mu((1-\lambda)m\phi)^{\frac{1}{\lambda}}\frac{m\phi((1-\lambda)m\phi)^{-1\lambda}}{\mu^2} \\ &= \frac{m\phi}{\mu}. \qquad \text{(C.5)}\end{aligned}$$

The results for the popularity distribution can be obtained as $\mathbb{E}(X) = \int \phi\mathbb{E}(X_\phi)\mathrm{d}\phi$. Simplifying this gives

$$\mathbb{E}(X) = \frac{m}{\mu}\int_{\phi_{\min}}^{\phi_{\max}} \phi\rho(\phi)\mathrm{d}\phi.$$

Since the integral simply gives the mean quality μ, this yields $\mathbb{E}(X) = m$.

C.4 Variance

The variance can be obtained using $\text{Var}(X_\phi) = \mathbb{E}(X_\phi^2) - \mathbb{E}(X_\phi)^2$ where $\mathbb{E}(X_\phi)^2$ can be obtained from equation C.5. So we will now calculate $\mathbb{E}(X_\phi^2)$. This is the integral

$$\mu((1-\lambda)m\phi)^{\frac{1}{\lambda}} \int_0^\infty k^2(k\lambda\mu + (1-\lambda)m\phi)^{-1/\lambda-1}\mathrm{d}k.$$

Again we will use integration by parts. Integration by parts once results in

$$\mu((1-\lambda)m\phi)^{\frac{1}{\lambda}} \int_0^\infty 2k\frac{(k\lambda\mu + (1-\lambda)m\phi)^{-1/\lambda}}{\mu}\mathrm{d}k.$$

By integrating by parts again we obtain for $\lambda < 1/2$

$$\mu((1-\lambda)m\phi)^{\frac{1}{\lambda}} \frac{2m^2\phi^2(1-\lambda)((1-\lambda)m\phi)^{-1/\lambda}}{\mu^3(1-2\lambda)}.$$

We simplify and obtain

$$\frac{2m^2\phi^2(1-\lambda)}{\mu^2(1-2\lambda)}. \tag{C.6}$$

Hence the variance is given by

$$\begin{aligned}\text{Var}(k) &= \frac{2m^2\phi^2(1-\lambda)}{\mu^2(1-2\lambda)} - \left(\frac{m\phi}{\mu}\right)^2 \\ &= \frac{m^2\phi^2}{\mu^2(1-2\lambda)}\end{aligned}$$

Again $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, where X is now the number of votes of a random item. Since $\mathbb{E}(X) = m$, we can write this as

$$\text{Var}(X) = \int_{\phi_{\min}}^{\phi_{\max}} \int_0^\infty \rho(\phi)k^2\mathbb{P}(X_\phi = k)\mathrm{d}k\mathrm{d}\phi - m^2.$$

We already obtained the solution of $\int k^2\mathbb{P}(X_\phi = k)$ in equation C.6. Hence, we can write this as

$$\begin{aligned}\text{Var}(X) &= \int_{\phi_{\min}}^{\phi_{\max}} \rho(\phi)\frac{2m^2\phi^2(1-\lambda)}{\mu^2(1-2\lambda)}\mathrm{d}\phi - m^2, \\ &= \frac{2m^2(1-\lambda)}{\mu^2(1-2\lambda)}\int_{\phi_{\min}}^{\phi_{\max}} \phi^2\rho(\phi)\mathrm{d}\phi - m^2.\end{aligned}$$

The integral $\int \phi^2\rho(\phi)\mathrm{d}\phi$ is simply the second moment of the quality distribution. The results for the various quality distribution can thus easily be obtained. A more insightful formulation can be given if we realise that the variance of quality, given by σ can be given by $\int \phi^2\rho(\phi)\mathrm{d}\phi - \mu^2$. Plugging this in, and simplifying gives

$$\text{Var}(X) = \frac{m^2(2\sigma(1-\lambda) + \mu^2)}{\mu^2(1-2\lambda)}, \tag{C.7}$$

where σ is the variance of the quality distribution.

C.5 Lorenz Curve and Gini Coefficient

I show for the uncertainty distribution how the Lorenz Curve and Gini coefficient are derived. If we take $\phi = \mu = 1$ the results for the Dirac quality distribution can be seen. First, we need to calculate the number of votes K at which the less popular p items are, or $\mathbb{P}(X < K) = p$. The cumulative distribution function (CDF) of equation 2.5 is

$$\mathbb{P}(X < K) = 1 - (m\phi(1-\lambda))^{\frac{1}{\lambda}}(K\lambda\mu + m\phi(1-\lambda))^{-\frac{1}{\lambda}}$$

So we need to solve

$$1 - (m\phi(1-\lambda))^{\frac{1}{\lambda}}(K\lambda\mu + m\phi(1-\lambda))^{-\frac{1}{\lambda}} = p$$

for K. Doing so yields

$$\begin{aligned}
(K\lambda\mu + m\phi(1-\lambda))^{-\frac{1}{\lambda}} &= (1-p)(m\phi(1-\lambda))^{-\frac{1}{\lambda}} \\
K\lambda\mu + m\phi(1-\lambda) &= (1-p)^{-\lambda}m\phi(1-\lambda) \\
K\lambda\mu &= (1-p)^{-\lambda}m\phi(1-\lambda) - m\phi(1-\lambda) \\
K\lambda\mu &= m\phi(1-\lambda)((1-p)^{-\lambda} - 1) \\
K &= \frac{m\phi(1-\lambda)((1-p)^{-\lambda} - 1)}{\lambda\mu}, \qquad \text{(C.8)}
\end{aligned}$$

The Lorenz curve is defined as

$$L(p) = \frac{\int_0^K k\mathbb{P}(X=k)\mathrm{d}k}{\int_0^\infty k\mathbb{P}(X=k)\mathrm{d}k}$$

The denominator is simply the mean, which is calculated in section C.3, and is $m\phi/\mu$. Using similar methods as in section C.3, $\int_0^K k\mathbb{P}(k)\mathrm{d}k$ can be written as

$$\frac{(1-\lambda)m\phi - ((1-\lambda)m\phi)^{\frac{1}{\lambda}}(K\lambda\mu + (1-\lambda)m\phi)^{-\frac{1}{\lambda}}(K\mu + (1-\lambda)m\phi)}{\mu(1-\lambda)}.$$

Substituting K as in equation C.8 and simplifying yields

$$\frac{m\phi}{\lambda\mu}(1 - (1-p)^{1-\lambda} - p(1-\lambda)),$$

which after dividing by the mean $m\phi/\mu$ yields the Lorenz curve

$$L(p) = \frac{(1 - (1-p)^{1-\lambda} - p(1-\lambda))}{\lambda}$$

The Lorenz curve is thus independent of quality ϕ and the mean number of votes m. The integral $\int_0^1 L(p)\mathrm{d}p$ is needed for the Gini coefficient and equals

$$\frac{\lambda - 1}{2(\lambda - 2)}.$$

Then the Gini coefficient, which is defined as

$$1 - 2\int_0^1 L(p)\mathrm{d}p,$$

becomes

$$1 - \frac{\lambda - 1}{\lambda - 2} = \frac{1}{2 - \lambda}.$$

This result is valid for both the uncertainty distribution as well as the inequality distribution assuming a Dirac quality distribution.

Bibliography

Abramowitz, M., and I. A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, 1970.

Albert, Reka, and Albert-Laszlo Barabasi. "Statistical mechanics of complex networks." *Review of Modern Physics* 74: (2002) 47–97.

Anderson, Chris. *The Long Tail*. Random House Business Books, 2007.

Barabási, Albert-László. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, 2003.

Barabasi, Albert-Laszlo, and Reka Albert. "Emergence of scaling in random networks." *Science* 286: (1999) 509–512.

Bianconi, Ginestra, and Albert-Laszlo Barabasi. "Competition and multiscaling in evolving networks." *Europhysics Letters* 54, 4: (2000) 436–442.

Boyd, Robert, and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, 1985.

Cheng, Xu, Cameron Dale, and Jiangchuan Liu. "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study." arXiv:cs.NI/0707:3670v1, 2007.

Chessa, Antonio G., and Jaap M.J. Murre. "A memory model for internet hits after media exposure." *Physica A Statistical Mechanics and its Applications* 333: (2004) 541–552.

Clauset, Aaron, Cosma R. Shalizi, and M. E. J. Newman. "Power-law distributions in empirical data." arXiv:physics.data-an/0706.1062, 2007.

Collins, Randall. *Macrohistory: Essays in Sociology of the Long Run*. Stanford University Press, 1999.

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

———. "General theory of bibliometric and other cumulative advantage processes." *Journal of the American Society for Information Science* 27, 5-6: (1976) 292–306.

Edling, Christofer R. "Mathematics in Sociology." *Annual Review of Sociology* 28: (2002) 192–220.

Friedkin, Noah E. "Theoretical Foundations for Centrality Measures." *The American Journal of Sociology* 96, 6: (1991) 1478–1504.

———. "Structural Bases of Interpersonal Influence in Groups: A Longitudinal Case Study." *American Sociological Review* 58, 6: (1993) 861–872.

———. *A Structural Theory of Social Influence.* Cambridge University Press, 1998.

———. "Choice Shift and Group Polarization." *American Sociological Review* 64, 6: (1999) 856–875.

———. "Norm formation in social influence networks." *Social Networks* 23: (2001) 167–189.

Gastwirth, Joseph L. "The Estimation of the Lorenz Curve and Gini Index." *The Review of Economics and Statistics* 54, 3: (1972) 306–316.

Gilbert, Nigel, and Klaus G. Troitzsch. *Simulation for the Social Scientist*. Open University Press, 2005.

Gill, Phillipa, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. "Youtube traffic characterization: a view from the edge." In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, 15–28.

Goldstone, Jack. *Revolution and Rebellion in the Early Modern World*. University of California Press, 1993.

Hunt, Francis, and Paul Johnson. "On the Pareto distribution of Sourceforge projects." In *Proceedings of the Open Source Software Development Workshop*. 2002.

Johansen, Anders. "Response time of internauts." *Physica A: Statistical Mechanics and its Applications* 296, 3-4: (2001) 539–546.

———. "Probing Human Response Times." *Physica A: Statistical Mechanics and its Applications* 338, 1-2: (2004) 286–291.

———. "Experiments on Internet Response." In *Eighth Granada Lectures Feb. 2005 AIP Conference Proccedings 779*. 2005.

McElreath, Richard, and Robert Boyd. *Mathematical Models of Social Evolution: A Guide for the Perplexed*. University of Chicago Press, 2007.

Merton, Robert K. "The Matthew Effect in Science." *Science* 159, 3810: (1968) 56–63.

Mitzenmacher, M. "A brief history of generative models for power law and lognormal distributions." *Internet Mathematics* 1, 2.

Myung, In Jae. "Tutorial on maximum likelihood estimate." *Journal of Mathematical Psychology* 47: (2002) 90–100.

Newman, M. E. J. "The structure and function of complex networks." *SIAM Review* 45: (2003) 167–256.

Pennock, David M., Gary W. Flake, Steve Lawrence, Eric J. Glover, and Lee C. Giles. "Winners don't take all: Characterizing the competition for links on the web." *Proceedings of National Academy of Sciences* 99, 8: (2002) 5207–5211.

R. H. Byrd, J. Nocedal, P. Lu, and C. Zhu. "A limited memory algorithm for bound constrained optimization." *Journal of Scientific Computing* 16: (1995) 1190–1208.

Salganik, M. J., P. S. Dodds, and D. J. Watts. "Experimental study of inequality and unpredictability in an artificial cultural market." *Science* 311, 5762: (2006) 854–856.

Simon, Herbert A. "On a Class of Skew Distribution Functions." *Biometrika* 42, 3/4: (1955) 425–440.

Skocpol, Theda. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge University Press, 1979.

Strogatz, Steven H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Perseus Books Group, 2001.

Turchin, Peter. *Historical Dynamics: Why States Rise and Fall*. Princeton University Press, 2003.

UNDP. *United Nations Development Report 2007/2008*. Palgrave MacMillan, 2007.

Watts, Duncan J. "A simple model of global cascades on random networks." *Proceedings of the National Academy of Sciences* 99, 9: (2002) 5766–5771.