POWER LAW DISTRIBUTIONS, SELF-ORGANIZING BEHAVIOR AND POPULARITY

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

DUYGU GÖZDE NASUHBEYOĞLU

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

SEPTEMBER 2015

Approval of the thesis:

POWER LAW DISTRIBUTIONS, SELF-ORGANIZING BEHAVIOR AND POPULARITY

submitted by **DUYGU GÖZDE NASUHBEYOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science, Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Director, **Graduate School of Informatics** ________________

Prof. Dr. Cem Bozşahin
Head of Department, **Cognitive Science Department, METU** ________________

Prof. Dr. Cem Bozşahin
Supervisor, **Cognitive Science Department, METU** ________________

Assist. Prof. Dr. Cengiz Acartürk
Co-supervisor, **Cognitive Science Department, METU** ________________

Examining Committee Members:

Prof. Dr. Deniz Zeyrek Bozşahin
Cognitive Science Department, METU ________________

Prof. Dr. Cem Bozşahin
Cognitive Science Department, METU ________________

Assist. Prof. Dr. Cengiz Acartürk
Cognitive Science Department, METU ________________

Assist. Prof. Dr. Murat Perit Çakır
Cognitive Science Department, METU ________________

Assist. Prof. Dr. Murat Ulubay
Management Department, YBU ________________

Date: ________________

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

Name, Last Name: DUYGU GÖZDE NASUHBEYOĞLU

Signature :

ABSTRACT

POWER LAW DISTRIBUTIONS, SELF-ORGANIZING BEHAVIOR AND POPULARITY

Nasuhbeyoğlu, Duygu Gözde
M.S., Department of Cognitive Science
Supervisor : Prof. Dr. Cem Bozşahin
Co-Supervisor : Assist. Prof. Dr. Cengiz Acartürk

September 2015, 26 pages

Zipf (1949) formulated Zipf's Law, which is based on mathematical statistics and is named after him, for many types of empirical data collected for analyzing physical or social sciences. Zipf's Law is a kind of Power Law probability distributions. One can do a Kolmogorov-Smirnov test to check whether there is goodness of fit to the hypothesized Power Law distribution and then exponential and lognormal distributions can be compared with the log likelihood ratio of the Power Law distribution to test whether an empirical data set applies to Zipf's law (Clauset et al., 2009). This method forms the basis of the method used in this thesis. The hypothesis of this thesis is "People's behavior's effect on popularity fits a Power Law distribution". To investigate this hypothesis, random top (popular) songs on iTunes have been selected and their YouTube ratings were collected every day for more than one month. Results were expected to show that the frequency of likes or dislikes increase every day and for songs that are liked or disliked more would be liked or disliked more. Also, in order to understand how something so random can become so popular, Internet Mahir's website's (Çağrı, 1999) Google search engine statistics have been analyzed. This thesis is to find out whether popularity is affected by other people's behaviour.

Keywords: Power Law Distribution, Self-Organizing Behavior, Popularity, Social Media Analytics

ÖZ

GÜÇ YASASI DAĞILIMLARI, ÖZ ÖRGÜTLEMELİ DAVRANIŞ VE POPÜLARİTE

Nasuhbeyoğlu, Duygu Gözde
Yüksek Lisans, Bilişsel Bilimler Programı
Tez Yöneticisi : Prof. Dr. Cem Bozşahin
Ortak Tez Yöneticisi : Yrd. Doç. Dr. Cengiz Acartürk

Eylül 2015 , 26 sayfa

Zipf (1949), adını kendisinden almış olan matematiksel istatistiğe dayanan Zipf Yasası'nı fizik ve sosyal bilimler için toplanmış ampirik verileri analiz etmek için formülize etmiştir. Zipf Yasası bir çeşit Güç Yasası olasılık dağılımıdır. Ampirik bir veri setinin Zipf Yasası'na uygulanabilirliğinin test edilmesi için, hipotez edilmiş Güç Yasası dağılımına uyum derecesini bulurken Kolmogorov-Smirnov testi yapılıp, daha sonra Güç Yasası dağılımının log olabilirlik oranı ile üstel ve lognormal dağılımların karşılaştırması yapılır (Clauset et al., 2009). Bu yöntem, bu tezde kullanılan metodun temelini oluşturmaktadır. Bu tezin hipotezi, "Popülaritenin oluşmasında insanların davranışlarının etkisi bir Güç Yasası dağılımı göstermektedir." olarak tanımlanmıştır. Bu hipotezi incelemek için, iTunes'dan rastlantısal olarak en çok beğenilen popüler şarkılar seçilmiş ve bu şarkıların YouTube reyting değerleri bir aydan fazla bir süre boyunca her gün toplanmıştır. Sonuçların, beğeni ve beğenmeme frekanslarının her gün yükseldiğini ve çok beğenilen ve beğenilmeyen şarkıların daha fazla beğenilip beğenilmediğini göstermesi beklenmiştir. Aynı zamanda, çok rastlantısal bir şeyin, bir anda çok popüler olmasını araştırmak için, İnternet Mahir'in sitesinin (Çağrı, 1999) Google arama motoru istatistikleri incelenmiştir. Bu tez, popüleritenin oluşmasında diğer insanların davranışlarının etkisi olup olmadığını araştırmaktadır.

Anahtar Kelimeler: Güç Yasası Dağılımı, Öz Örgütlemeli Davranış, Popülarite, Sosyal Medya Analizi

To my loving family ...

ACKNOWLEDGMENTS

I would like to express my gratitude and thanks to my advisor Prof. Dr. Cem Bozşahin who gave me the opportunity, invaluable guidance and motivation to complete this thesis with all his patience and understanding. I also thank my co-advisor Assist. Prof. Dr. Cengiz Acartürk for his invaluable insights. I thank all the members of Cognitive Science Department for the vast and high level knowledge that they provided me as a Master's student. I thank my parents, my brother Burak, my cousin Şenöm Yalçın, my dear Murat Tümer, my friends and my colleagues at HAVELSAN A.Ş for their tireless support during the tough days that I went through within the past three years. I also thank social media and especially YouTube for providing free access to its analytics data.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Power Law occurs when the frequency of some phenomenon varies as a power of its size. According to Clauset et al. (2009), the populations of cities, the intensities of earthquakes, the sizes of power outages are some examples that are thought to show Power Law Distributions. It is easy to see the Power Law Distribution in an event that has already occured whereas it is really hard to examine whether a newly emerged phenomenon will show a Power Law Distribution within the upcoming days.

It was Herbert A. Simon who analyzed the complexity architecture and proposed a an attachment mechanism which was important in making preferences to explain Power Law Distributions. According to Simon (1955) these Power Law Distributions can explain data taken from sociological, biological and economic phenomena. For example, he argues that in prose sentences, the distributions of words by their occurrence frequency can be explained by using Power Law Distributions. This law can be applied to population of cities, size of incomes, number of papers published by scientists, number of species in biological genera (Simon, 1955). More instances can be found.

In the case of frequency of occurrence of words in a text, Simon (1955) states that as the text is progressing, a context is formed and because of this context, some words are more likely to appear than others. According to the stochastic model that Simon proposed, the probability of a certain word to occur is proportional to the next occurrence of that word (Simon, 1955). Power Law Distributions explain the reason why cities with large populations continue expanding. The reason is that it affects the preferences of people. It can be said that as a city grows in population, people can have the intuition that the reason why people choose that city is that it gives more economical opportunities than other cities. Moreover, the lifestyles of people living there can be appealing. Considering Istanbul, it is possible to understand why people tend to live there even if there is traffic jam. There are lots of attributes of a growing city to be chosen as the city to set up a home.

Power Law can also be seen in social networks. Within the last few years, some things or people have become more popular than one could imagine due to sharings on the internet. They have even become phenomenal. Trying to find out whether a song, a tweet, a video or anything you can think of will show a power law distribution is not that easy. It is possible to collect analytics data from the internet; however, when it comes to analyzing this data one cannot capture a power law distribution attached to it while the phenomenon's popularity is developing. Manually doing this analysis is also impossible since there is a vast amount of data scattered around on various websites or social media platforms. Moreover, it is not only the internet that makes such phenomena go viral. It is possible to encounter regular people, even with no access to the internet, wear a t-shirt with the phenomenon's picture printed on it without even knowing what it stands for, but making the phenomenon become even more popular on the street. To see whether a phenomenon shows a power law distribution, one has to look into a lot of things.

One can possibly say that something has shown a power law after collecting data after it has happened. This post hoc analysis is not that interesting if the researcher does not show what happens after the thing has become phenomenal. A really good song may hit the top charts on its release day; but what if a song, which is sung by an unknown singer, which is not even included in an album becomes so popular that even TV shows want to put it on their show? What if someone makes the most simple website with pictures of him, saying "i kiss you" and gets millions of visits to the website in a single year? It is only then when people change their behavior accordingly and the power law phenomenon becomes something really exciting to look for.

The latin saying “*de gustibus et coloribus non est disputandum*" touches the very well known phenomenon that is “*tastes and colors are indisputable*" (Farnsworth, 1950). While one describes the color of a t-shirt as red, the other may call it vermillion. Taste is a phenomenon that can be experienced in almost anything. While one person likes jogging, the other may like skiing, and another may not like doing sports at all. While one person likes classical music, the other may enjoy listening to heavy metal .

“*Taste*" is formed by the character of a person as well as some external factors. Taste is based on a person’s urge to fit in with a certain society or culture, on his socio-economic background, on advertising, on forces in the market or on social media looking at today’s circumstances.

Social media provides a vast amount of data on tastes. People have a chance to express their ideas about a song, a photograph, a video, a status update, about anything one can think of on social media platforms. On social media people have the chance to encounter and “*follow*" a larger amount of people than they do with their “*physical being*". This gives people the freedom to fit in with a broader society that shares the same tastes, feelings or thoughts with them. Social media brings together people from different cultures, and it breaks the borders and burdens between them. Age, race, sex, religion, income, political ideology, body shape, health do not prevent people from having an identity on social media. Your identity is whatever you feel like expressing yourself as. One may become a top designer, a food addict, a pop icon, a phenomenon, most importantly a unique human being on social media and form constellations with like-minded people.

Today, a single word, a single snapshot, a simple video, a song can mean lots of things for people. By sharing something, people connect to each other on the basis of “*six degrees of separation*" and have the opportunity to indulge into conversations, to say something about a phenomenon and state that they have good and important ideas which are sometimes extraordinary, sometimes not.

On February, 26 2015, the discussion on the color of “*the dress*" became the center of attention of the world in almost minutes. It was either blue and black, or white and gold (Griffin, 2015). This made it very clear to people that colors are indisputable because there seems to be multiple answers to the question “*What color is this dress?*". The manufacturer of the dress may give the exact answer to the question, because the color codes that are used on the dress are definite. However, the emerging question is on how each person sees those color codes as.

The “*hashtags*", the number of “*likes*" or “*dislikes*" are some forms of grouping alike things or tastes on social media. With the help of social analytics applied on alike data, it is possible to measure tendencies of large groups of people towards anything.

Social networks are very powerful in the sense that they connect people who do not know each other yet but might later on. We have seen social networks’ power in 1999 when Mahir Çağrı launched a website “*www.ikissyou.org*" and suddenly became a phenomenon when 50000 people viewed the website in a minute, 800000 people viewed it in a few hours. Between November 1999 and April, 12 2000 3 million 173 thousand 973 people visited the website according to Guinness Book of World Records. Forbes magazine announced that he is the 100th most powerful person in the year 2000 (Çağrı, 2013). He became the center of attention immediately and Forbes magazine stated that 800 billion people know him worldwide (Arman, 2013). "*Internet Mahir*"'s website was possibly the beginning of a new era in which a phenomenon called social media rules the world. It shows that something so random can become something that changes people's behavior.

The study in this thesis is based on the idea that any popular phenomenon shows post-hoc a Power Law Distribution; however the difficulty is to foresee the development before it happens. This thesis aims to explore what makes phenomena like "*Internet Mahir*"' show power law distribution and more importantly, the effects of such phenomena on social behavior. Furthermore, it will be presenting a statistical analysis of some random YouTube songs' ratings on a daily basis in order to show that it is not that easy to capture power law distribution while it is happening by doing such an analysis on big data.

Cognitive scientists would be interested in powerlaws. The reason is that as Simon (1955) said, if it has such a distribution, then we don't have to posit complex mechanisms to explain what's happening. The data seems to self-organize. That avoids having to postulate complex mechanisms when there are simpler explanations. If the data fits lognormal distribution, then the data self-organizes towards popularity. Lognormal fit gives a finite mean/variance, whereas Power Law fit does not give finite mean/variance, it rather needs other mechanisms to become this way. So, when there is a Power Law in the data, then there might be some other factors that make this happen such as in the Internet Mahir case. Popularity of Internet Mahir rised because of the power of the world wide web in connecting networks of people, television programmes, celebrities attention, the large parties made for him and people's behavior.

CHAPTER 2

BACKGROUND ON THE PRINCIPLE OF LEAST EFFORT AND POWER LAW DISTRIBUTIONS

People, animals and machines show a tendency towards paying the minimum amount of effort while doing something. The French philosopher Ferrero (1894) was the first to discover the Principle of Least Effort. Later, Zipf (1949) studied this principle in his book by theorizing that the urge for efficient communication with least effort affects the distribution of word use under the name Zipf's Law.

This chapter will focus on Zipf's Law and the Principle of Least Effort. In Section 2.2 a comparison of Zipf's Law, Pareto Distributions, Gibrat's Law and Power Law Distributions will be made and in Section 2.3 Power Law and Simon's findings in his study (Simon, 1955) will be explained in detail. In the last Section 2.4 Self-Organization in human communities and popularity will be explained.

2.1 Zipf's Law: Human Behavior and the Principle of Least Effort

George Zipf (1949) claims that "every individual's movement, of whatever sort, will always be over paths and will always tend to be governed by one single primary principle which, for the want of a better term, we shall call The Principle of Least Effort". According to Zipf (1949), when an individual has a problem to solve, he will try to solve his it with minimized total work taking into consideration the immediate and future problems that he has estimated accordingly. Travelling Salesman Problem (TSP) is an example of such a problem. The problem is to find the shortest possible route between a given list of cities by passing each city and returning to the city that the salesman started his journey at. It is an NP-hard problem, which is basically a problem that is at least as hard as the hardest NP problem (van Leeuwen, 1998). The reason that it is called NP-hard is because we have to call the binary search routine exponential number of times. A person who has to drop by a few places in the same range of time or who has a routine plan of work does this kind of optimizations. The important thing is that he wants to put minimum amount of work or minimum amount of time for doing what he is doing.

According to Zipf, "singleness of a superlative" in a sentence, whether it is the minimum or the maximum, makes it easier for the listener to understand what the sentence actually means. As an example, Zipf states that when one utters "there will be a prize for the submarine commander who sinks the greatest number of ships in a given interval of time" or "there will be a prize for the submarine commander who sinks the a given number of ships in the shortest possible time", it is easy to understand what the prize stands for. However, if there were two superlatives in the sentence, it would be hard to comprehend who would be given the prize, i.e. "there will be a prize for the submarine commander who sinks the greatest number of ships in the shortest possible time".

Taking inheritence into consideration as well, Zipf argues that all varying conduct of an individual is governed by the invariable minimum of least effort. Considering a path between two cities, to minimize the effort to go to the destination, an individual may use all sorts of transportation. If we think of a mountain passing through the path between those two cities, it is perhaps best to build a tunnel to connect the two cities. Building the tunnel is a collective work, minimizing the effort that every single person would put by passing the mountain to go to the destination city without the tunnel according to Zipf (1949).

Zipf states that the frequency of words is inversely proportional to its rank in the frequency table within a given utterance corpus. This means that the most frequent word would occur 2 times more than the second most frequent word, 3 times more than the third most frequent word, and so on. "The" is found to be the most frequent word in Brown Corpus of American English text, found 7 percent times (69,971) in all word occurances. "Of" is the second most occurring word and it has 3,5 percent times (36,411) occurance which is half of "the". "And" has the third place in occurance frequency and it occurs (28,852) times. In half of the Brown Corpus 135 words are used as stated on the study (Fagan, 2010).

According to Clauset et al., it is possible to test whether Zipf's Law applies in a data set. It is possible by examining the goodness of fit of an emprical distribution to the hypothesized Power Law distribution by Kolmogorov-Smirnov test and comparing the Power Law distribution with exponential or lognormal distributions (Clauset et al., 2009). According to Eeckhout, when Zipf's Law is checked for cities, a better fit of b = 1.07 has been found which means that the n^{th} largest population is $frac1n^{1.07}$ times the size of the largest population. Zipf's Law holds for the upper tail of the distribution however the entire distribution is lognormal. It follows Gibrat's Law (Eeckhout, 2004). A lognormal tail can not be distinguished from a Pareto -Zipf- tail. Therefore, these two laws are both consistent. A comparison of these distributions will be made in Section 2.2.

2.2 Comparison of Zipf's Law, Pareto Distributions, Gibrat's Law and Power Law Distributions

Zipf, Pareto and Power Law distributions all describe large phenomena are rare whereas small phenomena are common. While there are only a few large earthquakes, there are many small ones. There are a few words such as "the", "or" and "and" which occur a lot; there are many words which do not occur as often as stated in the study (Adamic, 2000). These sorts of phenomena are described by a line that appears on a log-log plot.

To start with, Zipf's Law states that the size y of an occurrence of a phenomenon is inversely proportional to its rank r i.e $y \sim r^{-b}$, where b is close to unity (Adamic, 2000).

The word occurences in Moby Dick follow Zipf's Law according to Shalizi.

Pareto asked how many people had an income greater than x rather than asking what the r^{th} largest income is; i.e The Cumulative Distribution Function (CDF) $P[X > x] \sim x^{-k}$. It means that there are only a few people who are billionaires, whereas there are many people with small income (Adamic, 2000).

In 2003, the richest 400 in the US follow Pareto's Law.

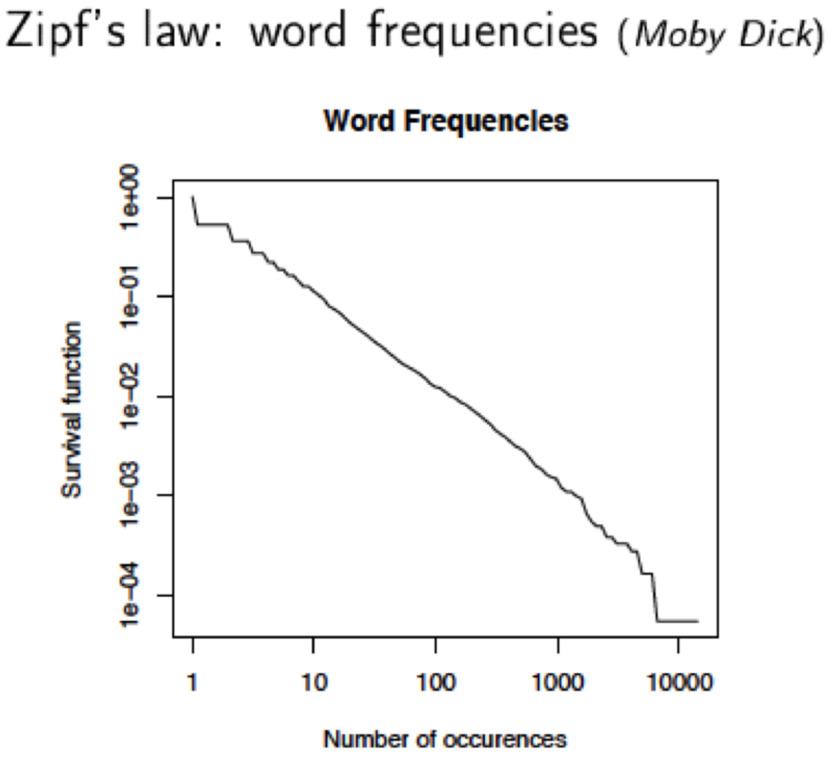

Figure 2.1: **Zipf's law: word frequencies Moby Dick.** Taken from Shalizi (2010)

Power Law distribution states the number of people whose income is exactly x; i.e the Probability Distribution Function (PDF): $P[X = x] \sim x^{-(k+1)} = x^{-a}$; where a = 1+k, where k is the Pareto distribution shape parameter (Adamic, 2000).

Gibrat's Law, the law of proportionate effect, states the size of a firm and its rate of growth are independent (Gibrat, 1931). This law gives a lognormal distribution. "While the city size distribution is often associated with Zipf's law, this holds only in the upper tail, because empirically the tail of a log-normal distribution cannot be distinguished from Zipf's law" (Eeckhout, 2004).

2.3 Power Law Distributions

Power Law distribution states the number of people whose income is exactly x; i.e the Probability Distribution Function (PDF):

$P[X = x]\ x^{-(k+1)} = x^{-a}$; where a = 1+k, where k is the Pareto distribution shape parameter (Adamic, 2000).

Some people think that it is really exciting to find out that the data that they have found fits a Power Law distribution. However, most of the time, people use bad methods to say that there is Power Law in the data. According to Shalizi, what has been found might not be a correct Power Law, because of using bad methods.

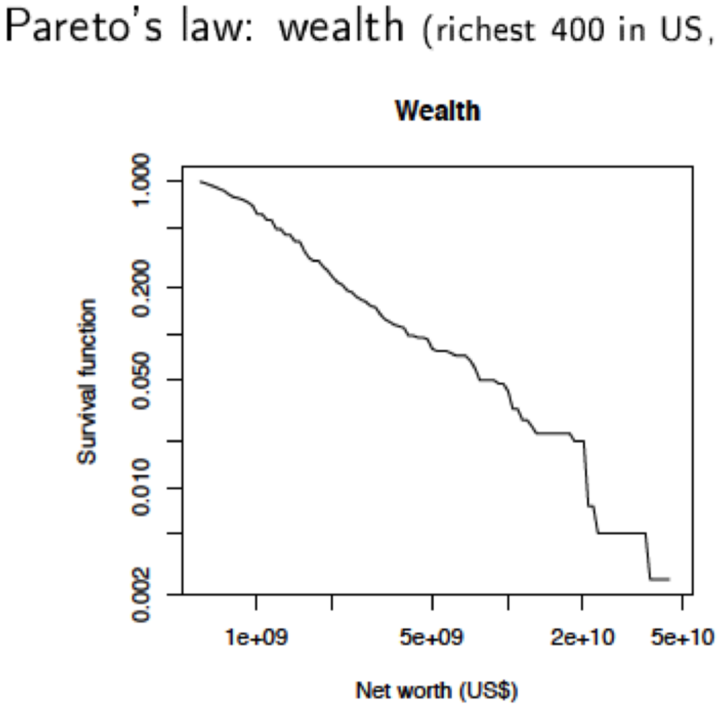

Figure 2.2: **Pareto's Law: wealth (richest 400 in US, 2003).** Taken from Shalizi (2010)

Some occurences of Power Laws are as follows;

"*Word frequency, protein interaction degree (yeast), metabolic network degree (E. coli), Internet autonomous system network, calls received, intensity of wars, terrorist attack fatalities, bytes per HTTP request, species per genus, number of sightings per bird species, population affected by blackouts, sales of bestsellers, population of US cities, area wildfires, solar flare intensity, earthquake magnitude, religious sect size, surname frequency, individual net worth, citation counts, number of papers authored, number of hits per URL, in-degree per URL, number of entries in e-mail address books, ...*" (Shalizi, 2010)

In actual fact, only few number of emprical data fit a Power Law fully. Mostly, the tail follows a Power Law rather than a perfect fit for all the values in the data set.

One has to distinguish a lognormal distribution from the Power Law distribution. "A nonnegative random variable X is said to have a Power Law distribution if

$Pr[X \geq x]\ cx^{-\alpha}$

for constants c > 0 and $\alpha > 0$. Here, $f(x)\ g(x)$ represents that the limit of the ratios goes to 1 as x grows large according to Mitzenmacher. If X has a power law distribution, then in a log-log plot of $Pr[X \geq x]$, also known as the complementary cumulative distribution function, asymptotically the behavior will be a straight line. This provides a simple empirical test for whether a random variable has a power law given an appropriate sample. For the specific case of a Pareto distribution, the behavior is exactly linear, as $\ln(Pr[X \geq x]) = -\alpha(\ln x - \ln k)$ (Mitzenmacher, 2004).

"*A random variable X has a lognormal distribution if the random variable Y = ln X has a normal (i.e.,*

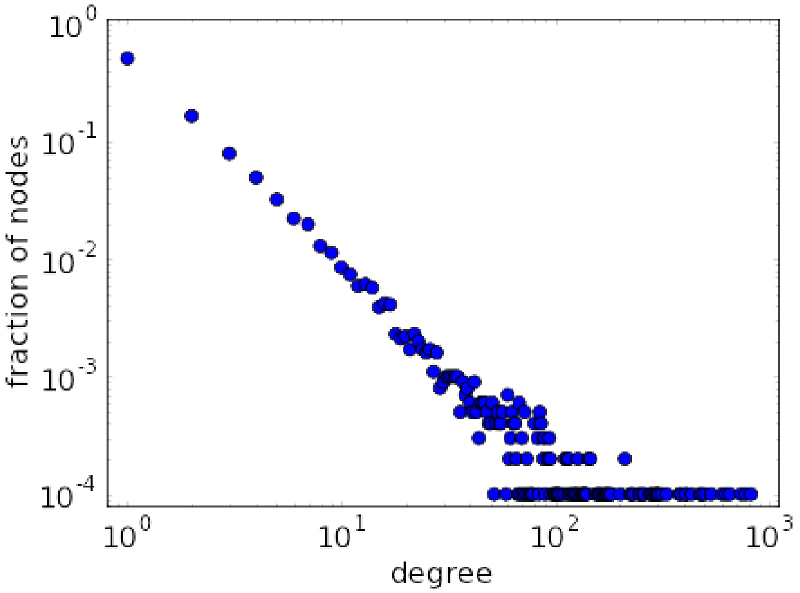

Figure 2.3: **Power Law Plot** Taken from Math Insight Nykamp (2015)

Gaussian) distribution" according to Mitzenmacher (2004).

$$\Pr[X \geq x] = \int_{z=x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma z} e^{-(\ln z-\mu)^2/2\sigma^2} dz.$$

Figure 2.4: **Complementary Cumulative Distribution for a Lognormal Distribution.** Taken from Mitzenmacher (2004)

New objects tend to attach to popular phenomena. This occurs in the World Wide Web and many other examples may be found. In the World Wide Web, newly created pages tend to address highly ranking webpages on search engines. We have found an occurrence of this for Internet Mahir's website. It will be explained in the next chapters.

Simon (1955), has given a much more clearer explanation for preferential attachment with a mathematical argument. He listed five occurences of preferential attachment that are distributions of word frequencies in documents, distributions of numbers of papers published by scientists, distribution of cities by population, distribution of incomes, and distribution of species among genera. In Simon's model, the system's dynamics is constantly growing when new elements are added while incrementing the counters at a rate which is proportional to their values. This means that new instances of a word occur, while incrementing new occurences of that word proportionally to their current values according to the study (Simon, 1955).

2.4 Self-Organizing Behavior and Popularity

Self-Organization is the capability of a system to arrange in a way that it does not need the help of any external agency when certain conditions are met.

According to Complexity Theory, organizations are "complex adaptive systems" which coevolve with the environment through the agents' self-organizing behavior. These agents navigate "fitness landscapes" of opportunities that the market provides and also of competitive dynamics (Kauffman, 1995; Coleman, 1999). Complexity theory suggests that self-organizing behavior naturally occurs without addressing the cause of it (Stacey, 1996; Coleman, 1999). Self-organizing behavior occurs when people can network with others freely to pursue their objectives (Coleman, 1999).

Today, many people are connected to each other through social network websites on the world wide web. People share and create knowledge every single moment through networking. People's behavior on social media is organized by certain phenomena like hashtags, like/dislike buttons and counters, status messages, sharings and many more. Who created this organization or how it is created is not that important for the society. What matters is that people have started to organize their behavior and create information through this way. Rather than who sent the first tweet on Twitter about a certain hashtag, the overall information stored under that hashtag is important. When a person tries to find pictures of cute cats, he just logs on to Instagram and makes a simple search under the hashtag *cutecats* and gets hold of the information that he wants to receive. Chaos of knowledge is prevented through such simple markings on social media. This is just like what complexity theory suggests. It is not the origin that matters, what matters is that people naturally begin to organize their behavior according to what they have learnt. For instance, people do not put a picture of a vase under the hashtag *cutecats*, they tend to preserve their rationality and validity on the Internet because what others think of it really matter for him since there is this phenomena of liking or disliking each post. The amount of data stored under these markings is also very important. This is because in order for someone or something to become popular, the statistics of it is what counts.

This is why social media analytics has become so important these days. Organizations, media, or simply put the *market* wants to know what people tend to do, or like. As mentioned in the Introduction of this thesis, the hashtag *thedress* has become so popular that companies used it for their advertisements on Twitter. This is the power of self-organizing behavior within society.

“*Popular*" comes from the Latin word *popularis* meaning “*belonging to people, general, common; devoted to or accepted by the people; democratic*" from *populus* which means *people* (Harper, 2015). Popularity occurs when there is a consensus about a certain thing or person. In their research, Cillessen and Rose have shown that there is “*sociometric popularity*" and there is “*perceived popularity* (Cillessen and A.J., 2012).

Sociometricaly popular people are mostly well liked people who have the empathy of others because of their good traits and because of their non-aggressive behavior whereas perceived to be popular people are mostly aggressive people who take advantage of social situations. Youth choose to find perceived popular people as “*popular*" rather than the people who they like. They want to listen to the music that the perceived popular people listen to, or dress and act like them even though they do not like them; because they think that other people like these kinds of people and they want to be part of this community according to the study (Cillessen and A.J., 2012).

‘*Internet Mahir* was an ordinary person before he created his website in 1999. He was well liked by his friends but when he entered the World Wide Web, he got to make new friends with people from other continents. There was a bit of positive aggressiveness in his website, because on his website he claimed to want to meet women with his phone number attached (he later claimed that his website got hacked in almost minutes and the sentence “*I like sex*" was added by the hacker). There were photos of him on the website, telling the world about him. Nobody seemed to believe that he was a good player of table tennis and that he was a gentleman; people thought that he was a joke. However, he was so charming that he made the sentence “*I kiss you*" become phenomenal. Within minutes, women from all around the world started calling him, to see whether such a person existed. The world liked him in minutes. Social networking made it possible for him to become popular. After becoming this popular, he did not become aggressive such that he did not want to get his share from this popularity. Rather, he attended social organizations to show that he is a good person. The reason we are interested in him in this thesis is that, he changed people’s behavior and became popular that way. When people saw him on the streets they sent him a kiss, celebrities sent him emails. He became a trend in a few hours and his website was probably the beginning of the social networking era.

CHAPTER 3

POWER LAW DISTRIBUTIONS IN EMPIRICAL DATA

Previous methods to claim that there is a Power Law in some empirical data seem to be mythical according to Shalizi (2010). As shown in the previous chapter, the data that was supposed to show Power Law are as follows:

"*Word frequency, protein interaction degree (yeast), metabolic network degree (E. coli), Internet autonomous system network, calls received, intensity of wars, terrorist attack fatalities, bytes per HTTP request, species per genus, number of sightings per bird species, population affected by blackouts, sales of bestsellers, population of US cities, area wildfires, solar flare intensity, earthquake magnitude, religious sect size, surname frequency, individual net worth, citation counts, number of papers authored, number of hits per URL, in-degree per URL, number of entries in e-mail address books, ...*" (Shalizi, 2010)

However, according to Shalizi, word frequency is the only data that follows Power Law distributions whereas *the rest are indistinguishable from lognormal and/or streched exponential and/or cut off significantly better than pure power law and/or goodness of fit is just horrible*. Shalizi gives an example of *animal foraging methods* and claims that scientists theorized for a dozen years that it follows a Power Law (Viswanathan et al., 1996), but later it has been found that it does not follow a Power Law (Edwards et al., 2007) (Shalizi, 2010).

This chapter will present the method in the study Clauset et al. (2009) for analyzing empirical data on Power Law distributions. Section 3.1 will present the methodology, Section 3.2 will elaborate their findings for certain emprical data.

3.1 Power Law Distributions Methodology

Clauset et al. claim that many empirical quantities dense around a typical value. The weights of bananas in a store can be given as an example. The values may vary somewhat, but their distributions place an acceptable amount of probability far from the average value, making the average value representative of most observations. "*Even the largest deviations are still only about a factor or two from the mean in either direction*". However, it can not be said that all distributions fit this pattern. While the ones that does not fit are often considered as defective, "*they are at the same time some of the most interesting of all scientific observations.* " "*The power law has attracted particular attention over the years for its mathematical properties, among such distributions. The intensities of earthquakes, for example, are thought to have power law distributions.*"

Power Law is as explained in the previous chapter, $p(x) = \alpha^{-a}$ "*where α is a constant constant pa-*

rameter of the distribution known as the exponent parameter. Even there can be exceptions the scaling parameter typically lies in the range of $2 < \alpha < 3$. *In practice, the power law applies only for values greater than some xmin. In such cases, it is said that the tail of the distribution follows a power law.*"

In this article, Clauset et al. try to answer the question of how to recognize a power law when they see one. In practice, they can rarely be certain, however. The most they can say is that their "*observations are consistent with the hypothesis that x is drawn from a distribution. Their goal here is to bring many methods, that have been discussed previously, together to create a complete procedure for the analysis of power law data.*" Clauset et al. state that Power law distributions are in two basic styles: "*continuous distributions governing continuous real numbers and discrete distributions where the quantity of interest can take only a discrete set of values, typically positive integers.*"

According to Clauset et al. there are 3 steps followed for the analysis of power law data:

- Applying power laws to empirical data: According to Clauset et al. For estimating α, a value is required for the lower bound xmin of power law behavior in the data. İt is assumed that α is known.

  Fitting power law distributions to observed data is called "*the method of maximum likelihood*". "*It provably gives accurate parameter estimates in the limit of large sample size. İf the data is drawn from a distribution that follows a power law exactly for* $x \geq$ *xmin, maximum likelihood estimators of the scaling parameter for both the discrete and the continuous cases can be derived.*"

  Before calculating α, all samples below xmin should be discarded; so, only the ones for which the power law model is a valid one would be left. Also, an accurate method for estimating xmin is needed for the accuracy of the estimation of α. In this thesis, the xmin that we have found for Internet Mahir's search engine statistics is 8.0 according to the Python output.

  According to Clauset et al. there are two methods that prevent noise or fluctuations in the tail of the distribution; one that is specific to discrete data and one that works for either discrete or continuous data. The first one is based on a so-called marginal likelihood, while the second one is based on minimizing the distance between the power law model and the empirical data. The first approach uses a generalized model to represent all of the observed data. It uses this model for both above and below xmin. The data is modeled by the standart discrete power law distribution. For above xmin; each of the xmin -1 discrete values of x are modeled by a seperate probability for below xmin. For the second approach, the value of xmin that makes the probability distributions of the measured data and the best-fit power law model is chosen as similar as possible above xmin.

  "Kolmogorov-Smirnov" or KS statistic is the commonest measure for quantifying the distance between two probability distributions which is used for non-normal data. It is the maximum distance between the Cumulative Distribution Functions of the data and the fitted model.

- Calculation of the goodness-of-fit between the data and the power law:

  Some way is needed to tell whether the fit of power law is a good match to the data. To analyze whether a data set follows a Power Law, basically many synthetic data sets from a true power law distribution need to be sampled. Then, how far they fluctuate from the power law form needs to be found and a comparison of the results with similar measurements on the empirical data needs to be done.

  Goodness-of-fit tests determine whether the hypothesis is a plausible one, given the data. They generate a p value that quantifies the plausibility of the hypothesis, based on measurement of the distance between the distribution of the empirical data and the hypothesized model. The p value is the fraction of the synthetic distances, larger than the empirical distance.If p is close to 1, it can be said that the model is plausible fit to the data; if it is small, it is not. In these calculations to quantif the distance between the two distributions, kolmogorov-smirnov statistic is used. There is one important point to be noted, that, xmin is the value that is used to decide how many data points to work with. Power law form is only fitted for values that are above the xmin value. According to Power law, there should be only a few data that fall above the xmin value in the data set. If there is a large value of xmin than the total value of n needed to reject the power law needs to be large as well.

- Alternative distributions: Even the methods described above provide a reliable result, an exponential or a log-normal distribution might also give a fit. This possibility can be eliminated by using a goodness-of-fit test. As an example, if p is reasonably large, the power law is not ruled out; if it is sufficiently small, exponential model can be ruled out.

  Many statistical tests can be used to determine whether some specific hypotheses can be ruled out or not, but it is up to the researcher to decide in the first place. The researcher should look for to rule out the candidate distributions first, if neither is ruled out, should examine which is a better fit. If the power law gets ruled out, then the work is done; but if it passes, the researcher should look for another distribution to see whether it might provide a better fit or not. In such cases, there is a method called the likelihood ratio test, which is considerably easier to implement to, than the KS test. What likelihood ratio test does is it computes the likelihood of the data under two competing distributions. One of them gets a higher likelihood, that means it is a better fit. Alternatively, the researcher can compare the results, which are negative or positive, by looking at the logarithm R of the ratio. (Clauset et al., 2009)

3.2 Findings

Power Law distributions study is done in many disciplines such as “*physics, biology, engineering, computer science, earth sciences, economics, political science, sociology and statistics.*" However, earlier studies were not well founded and the hypothesized distributions were not tested properly. When a straight line on a log-log scale has been achieved, one should not trust that it is a true power law. “*In studies of the Internet, for instance, the distributions of many quantities, such as file sizes, HTTP connections, node degrees, and so forth, have heavy tails and appear visually to follow a power law, but upon more careful analysis it proves impossible to make a strong case for the power-law hypothesis; typically the power-law distribution is not ruled out but competing distributions may offer a better fit to the data.*" (Clauset et al., 2009). According to Clauset et al. (2009) getting a power law is not that interesting as long as there is a plausible reason that describe the underlying mechanisms when “*observed quantity follows power law or some other form.*".

CHAPTER 4

METHODOLOGY AND EMPIRICAL WORK

In this chapter, the methodology and the empirical work done for this thesis will be explained in detail.

4.1 Model

Recently, scientists proposed many statistical methods to fit power law distributions to empirical data. Alstott et al. (2014) proposed a Python package to describe whether some phenomenon fit power law distributions. In this thesis, this package has been used and the Python code that has been provided by them has been changed to fit our purposes.

A random day was chosen to collect the top 100 songs on iTunes. On Sunday, November 16 2014, at 2:00 am, iTunes data were collected. A song appeared twice on the list that is why 99 songs were collected in the end. Then a search for these top 99 songs for their official or correct videos was made on YouTube. After collecting the video id's on YouTube for these songs, a random day was chosen (25.11.2014) to start collecting statistics data from YouTube. For YouTube videos, statistics data means view count, like count, dislike count, favorite count and comment count. In order to collect data from YouTube, YouTube's API for "*Videos: list*" has been used. The "*Try it!*" functionality of YouTube API, which is designed to "*use the API Explorer to call the method on live data and see the API request and response*" has been used. Every evening until 14.01.2015, this API has been called by setting "*part*" as statistics, "*id*" as the video ids separated by comma and then executing. YouTube's API returned 94 of these songs' ratings. After collecting the data, Python's Power Law package has been used to analyze the data. Alstott's power law package is "*a toolbox using the statistical methods developed by Clauset et al. 2007 and Klaus et al. 2011 to determine if a probability Distribution fits a Power Law*"

Another data that has been collected is Internet Mahir's Google Search statistics. The number of websites have been collected and the data has been analyzed.

4.2 Results

YouTube songs and Internet Mahir's Google search statistics have been collected and the following results were achieved.

Table 4.1: **YouTube and Internet Mahir's search engine statistics analysis**

Date and Data	ks distance	alpha error	xmin	p	R(PL vs Exp)	R(PL vs Logn)	Favored
25.11.2014 - YouTube	0.145	0.333	513379.0	0.282		-1.674	Lognormal
14.01.2015 - YouTube	0.153	0.118	165975.0	0.088		-4.324	Lognormal
14.03.2015 - YouTube	0.136	0.106	176957.0	0.020		-6.432	Lognormal
13.07.2015 - Internet Mahir	0.117	0.053	8.0	8.199	7.76		Power Law

'*The Kolmogorov-Smirnov distance is used to generate a p-value for an individual fit vs using loglikelihood ratios to identify which of the two fits is better*" (Clauset et al., 2009; Alstott et al., 2014). R is the loglikelihood ratio. When R is positive, the data is more likely to fit Power Law distribution. If R is negative, then the data is more likely to fit Lognormal distribution. In the powerlaw Python package, "*the Fit object retains information on all the xmins considered, along with their Ds (KS distance), alphas, and sigmas*" (Alstott et al., 2014). '*As a power law is fitted to data starting from different, the goodness of fit between the power law and the data is measured by the Kolmogorov-Smirnov distance, with the best minimizing this distance*" (Alstott et al., 2014). '*The optimal xmin is defined as the value that minimizes the Kolmogorov-Smirnov distance, D, between the empirical data and the fitted power law.*" (Alstott et al., 2014).

On November 25, 2014 for YouTube data, it cannot be said this data fits a Power Law Distribution as seen in figure 4.1.

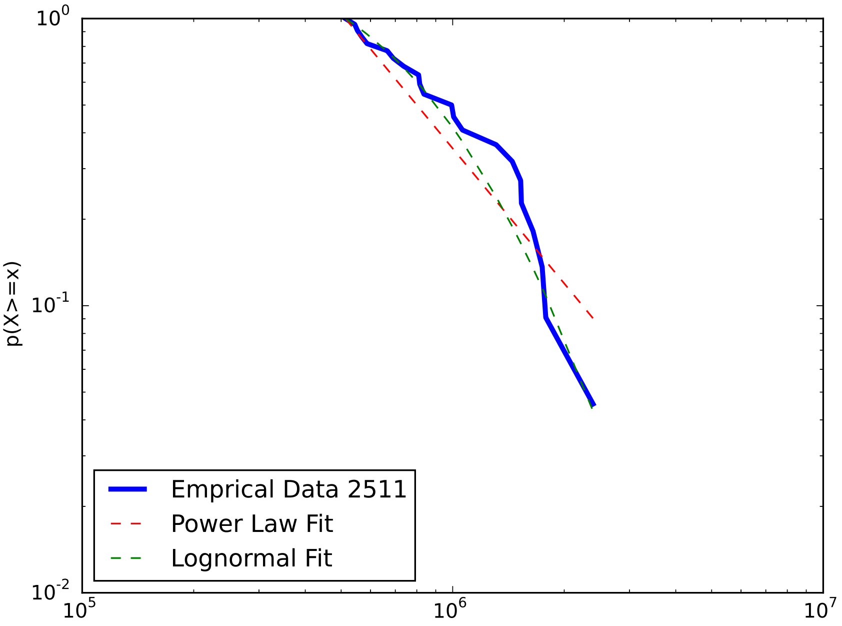

Figure 4.1: **YouTube Songs - Power Law vs Lognormal 25.11.2014.**

On January 14, 2015 for YouTube data; it cannot be said this data fits a Power Law Distribution as seen in figure 4.2.

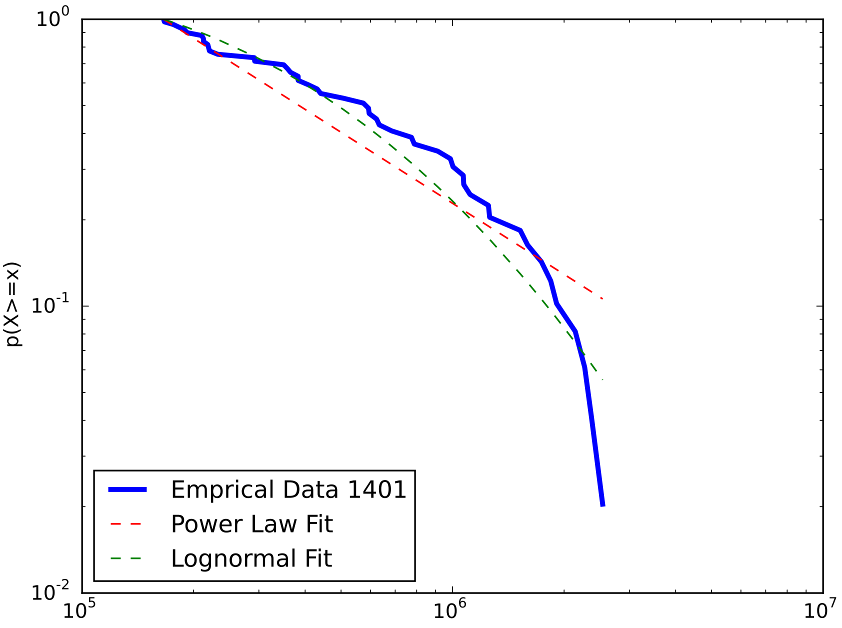

Figure 4.2: **YouTube Songs - Power Law vs Lognormal 14.01.2015.**

After two months, Power Law is still not achieved as seen in figure 4.3. Data collected on 14.03.2015 show that the like counts of songs that were liked by more people have not increased too much to show any Power Law.

The data is more likely to fit lognormal distribution rather than power law for YouTube songs.

For Internet Mahir, top ranking websites' number of hits on Google search has been found and this data has been the input for the Python code. The results in figure 4.4 show that this data fits power law distributions.

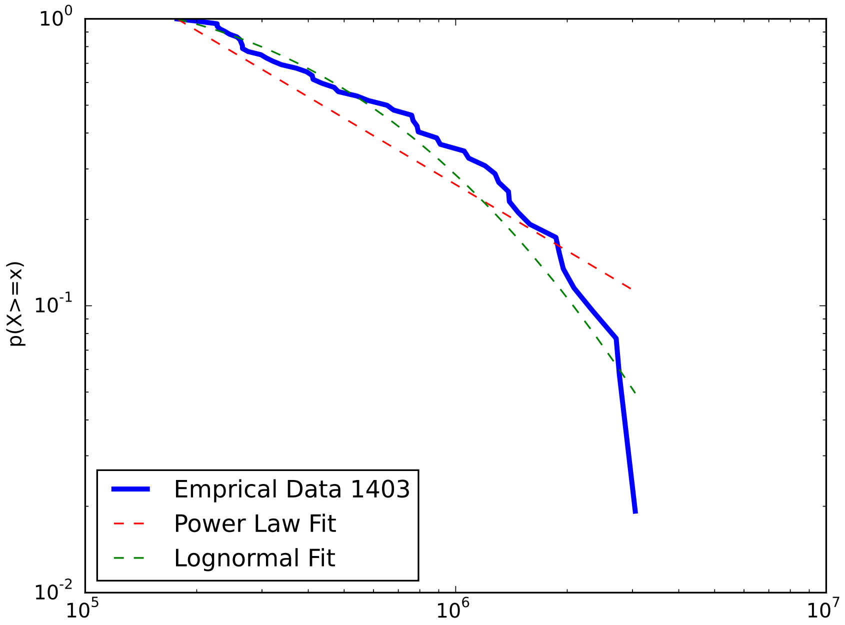

Figure 4.3: **YouTube Songs - Power Law vs Lognormal 14.03.2015.**

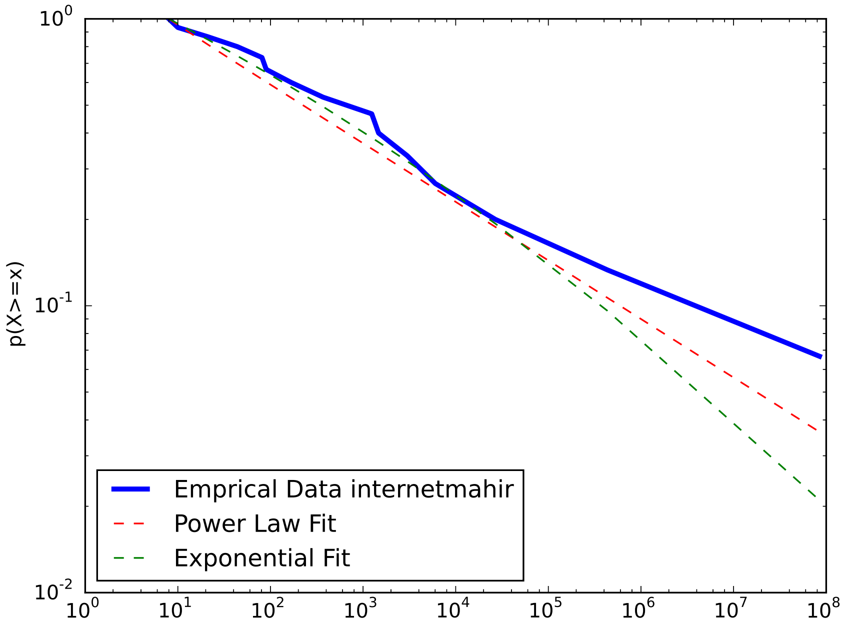

Figure 4.4: **Internet Mahir - Power Law vs Exponential.**

CHAPTER 5

CONCLUSION AND DISCUSSIONS

As seen in the results in the previous chapter, the YouTube songs' that have been analyzed in this thesis show a lognormal distribution whereas Internet Mahir's rise shows a Power Law distribution. According to Shalizi, one should not say that every popular phenomenon shows a Power Law. Rather, one should do a more detailed study to see whether the phenomenon is actually a Power Law (Shalizi, 2010; Clauset et al., 2009). Therefore, the analysis in this thesis has been done accordingly to show that some phenomenon might seem to show Power Law distributions whereas they do not.

This study shows that popularity of personal taste and popularity because of trend are two different kinds of phenomena. The popularity of a song like the ones that have been looked upon in this thesis, depends on people's real personel taste. Maybe they have not been advertised or have not become a trend on social media like Psy's song Gangnam Style. They were the outcome of iTunes' most downloaded song chart for a random day. People bought the song as personal taste and made it become top. Therefore, it was not possible to see a trending fashion in their daily YouTube statistics. For instance, if a new album of Barış Manço with songs that were never released by him before would become a trend on the day that it would be released because he was a really good musician. However, a new album of him with songs that were released earlier would not become so popular; and people would only buy it when they just feel like buying it. Henceforth, the songs that are listened to according to popularity of personal taste seem to show a lognormal distribution rather than a Power Law distribution.

The data collected for Internet Mahir shows a Power Law distribution. This is because he was a trending person related with the sentence "*i kiss you*" that many people, even celebrities, felt like sending a kiss back to him. After his rise in 1999, he tried to lead a simple life as he used to, however people still knew him. If he starred in a movie, maybe things would be different but he would soon be forgotten whereas he is still known by the world.

In this thesis, we tried to understand whether people's behavior has an effect on popularity or not by using Power Law distribution analysis. Power Law phenomena are usually the outcome of a trend like in Internet Mahir's case. We collected the data in 2015, whereas he put his website online in 1999. We analyzed that his website's statistics show a Power Law after it had already happened. However, if somebody like Internet Mahir would put a strange website online and we looked at it's statistics starting from the moment that the website went online, then, it would be really interesting but it is really hard to find out that something would show a Power Law in the upcoming days, since the data that would be looked at would be really small. Although it is hard to find out that it would show a Power Law in the upcoming days, if people's networking behavior would be analyzed as well, maybe things would differ. For example, in Internet Mahir's case, people not only referred to his website online, but they called him on the phone, invited him to their homes, stayed at his home, shouted "*i kiss you* when they saw him on the street, talked about him. If these would be analyzed, then an extreme popularity without substance would be easily seen.

It seems that only in the case of power law-distributed popular phenomena and not in log-normal or exponential phenomena when a social event becomes more than personal taste or show of quality and endurance. Only in these cases it seems that the mechanics of the connectivity kicks in to heighten and amplify the preferential attachment. Otherwise it's more of the same stuff, and does not need to look into network dynamics to explain the phenomenon.

The data that has been used is really very small compared to the vast resources on the Internet. If all the data on YouTube or the network of websites that have started with Internet Mahir's website (Çağrı, 1999) could have been collected and the results were calculated for each moment, which is very hard to do, a much precise result could have been achieved.

According to the results we got looking at a very small portion of YouTube data; it can be said that, one's music taste is not formed by what other people think of that music. The reason is that the amount of the data is really small to make a conclusion like this. There are various aspects that make a person like a song. It is important to distinguish personal taste from public taste. One can watch a music video on YouTube but it is not necessary that he will like the song based on other people's views; can listen to the same song for years and years and get the same taste whereas a song that took high ratings may appear to be disliked in the future. This can be seen in classical music versus pop music. A popular song can be forgotten in time whereas classical music can still be liked even if decades have passed.

What's most important is people's way of self-organizing towards some goal. The limitations of getting hold of data makes it hard to understand whether there is Power Law in people's behavior on popularity while it is happening. We can see Power Law in events that have already occurred such as İnternet Mahir's popularity. We looked at the highest ranked searches in Google for Internet Mahir. However, some pages get to enter search engines with highest rank by some marketing strategies other than only by visitor counts. Therefore, this makes a limitation for our research. If we could get hold of whole data, and found the xmin much more easily, would things be different? Probably slight changes in the findings could occur.

Briefly, the formalization presented in this thesis can be extended. Some of these possible proposals for future research can be listed as:

- Big Data project of collecting whole data on any sort of social media platform for each passing day and analyzing this data
- Collecting data for different genres of music and comparing them: Popular music or people tend to get more advertisement, therefore, it is important to see the rise of the popularity of a song or person without any advertisement.
- Emprical research for whether people really press the like button when they see huge numbers of like counts: A website can be designed and the behavior of people can be examined through an experiment conducted on that website.
- Sentiment Analysis and Opinion Mining on social media statistics: Social Networks can be analyzed using this method. A classifier can be trained using Machine Learning in order to categorize the texts on the comments or reviews shared on these websites. The classifier

assigns the data collected from the websites to one of the classes such as like or dislike in music sharing websites.

Bibliography

Adamic, L. A. (2000). Zipf, power laws and pareto. *HP Labs*.

Alstott, J., E. Bullmore, and D. Plenz (2014). powerlaw: a python package for analysis of heavy-tailed distributions. *PLoS ONE 9*(4).

Arman, A. (2013, March). Dikkat! internet mahir yeniden geliyoooor. *Hürriyet*.

Çağrı, M. (1999). www.ikissyou.org. www.ikissyou.org.

Çağrı, M. (2013). I kiss you. www.ikissyou.co.

Cillessen, A. and R. A.J. (2012). Understanding popularity in the peer system. *Current Directions In Psychological Science (American Psychological Society) 14*(2), 102–105.

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

Coleman, H. J. J. (1999). What Enables Self-Organizing Behavior in Businesses. *Emergence: Complexity and Organization 1*(1), 33–48.

Edwards, A., R. Phillips, N. Watkins, M. Freeman, E. Murphy, V. Afanasyev, S. Buldyrev, M. G. E. Luz, E. P. Raposo, H. Stanley, and G. Viswanathan (2007). Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature 449*, 1044–1048.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review 94*(5), 1429–1451.

Fagan, Stephen; Gençay, R. (2010). *Ullah, Aman; Giles, David E. A., Handbook of Empirical Economics and Finance*, Chapter An introduction to textual econometrics, pp. 133–153. CRC Press.

Farnsworth, P. (1950). *Musical Taste: Its Measurement and Cultural Nature*. Education-psychology ; v. 2, no. 1. Stanford University Press.

Ferrero, G. (1894). L'inertie mentale et la loi du moindre effort. *Revue Philosophique de la France et de l'Étranger 37*, 169–182.

Gibrat, R. (1931). *Les Inégalités économiques*. Paris, France.

Griffin, A. (2015, February). Blue and black or white and gold, how the dress colour you see says a lot about you. http://www.independent.co.uk/news/science/what-color-is-the-dress-blue-and-black-or-white-and-gold-whatever-you-see-says-a-lot-about-you-10074490.html.

Harper, D. (2015). Etymology online entry for popular. http://www.etymonline.com/index.php.

Kauffman, S. (1995). *At Home in the Universe: the Search for the Laws of Self-Organization and Complexity*. New York: Oxford University Press.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics I*(2), 226–251.

Nykamp, D. Q. (2015). "plot of power-law degree distribution on log-log scale.". http://mathinsight.org/image/power_law_degree_distribution_scatter.

Shalizi, C. (2010, October). "so, you think you have a power law, do you? well isn't that special?". http://www.stat.cmu.edu/ cshalizi/2010-10-18-Meetup.pdf.

Simon, H. A. (1955, Dec.). On a class of skew distribution functions. *Biometrika 42*(3/4), 425–440.

Stacey, R. D. (1996). *Complexity and Creativity in Organizations*. San Francisco, CA: Berrett-Koehler.

van Leeuwen, J. (1998). *Handbook of Theoretical Computer Science, Vol.A, Algorithms and complexity*. Amsterdam: Elsevier.

Viswanathan, G. M., V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley (1996). Lévy flight search patterns of wandering albatrosses. *Nature 381*, 413–415.

Zipf, G. (1949). *Human Behavior and The Principle of Least Effort, An Introduction to Human Ecology*. Addison-Wesley Press, Inc.