Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea Magistrale in Matematica

Stochastic models and graph theory for Zipf's law

Tesi di Laurea in Fisica Matematica

Relatore:
Chiar.mo Prof.
Marco Lenci

Presentata da:
Anna Di Natale

Correlatori:
Chiar.mi Proff.
Giampaolo Cristadoro
Vito D.P. Servedio
Vittorio Loreto

Sessione V
Anno Accademico 2017/2018

"La Rettorica ha detto tanto bene di Dante, che io ebbi la vaghezza di sapere che cosa ne pensasse l'Aritmetica, chiamata con verità dal Gibbon la nemica naturale della Rettorica. E l'Aritmetica ne dice meglio che mai; com'ebbi a ragionarne all'Accademia dei Lincei, e come più distesamente ne ho scritto in questo libretto, che forse non riuscirà disutile per la scienza e l'arte."

Filippo Mariotti, *Dante e la statistica delle lingue*

Contents

Abstract

In questo elaborato ci siamo occupati della legge di Zipf sia da un punto di vista applicativo che teorico. Tale legge empirica, osservata per la prima volta da Estoup ma formalizzata da Zipf nel 1936 [26], afferma che il rank in frequenza delle parole di un testo seguono una legge a potenza con esponente -1. Nello steso modo è stato osservato che molteplici altri fenomeni seguono la stessa legge, per esempio il numero di abitanti delle città o la magnitudine dei terremoti. Essendo così diffusa, tale legge potrebbe essere legata a qualche proprietà fondante della natura o al nostro modo di indagare i fenomeni che ci circondano.
Vengono richiamate alcune nozioni di probabilità e di teoria dei grafi necessarie alla comprensione della trattazione. Successivamente ci siamo occupati di introdurre le leggi di Zipf e di Heaps, due leggi legate tra loro ma che trattano di argomenti diversi. Infatti, mentre la legge di Zipf tratta del rank di frequenza la seconda si occupa della legge di crescita del vocabolario. In particolare, la legge di Heaps afferma che il numero di novità di alcuni sistemi cresce seguendo una seconda legge a potenza il cui esponente può essere legato a quello della power-law che approssima il rank in frequenza dello stesso sistema.
Abbiamo affrontato l'analisi di queste leggi richiamando anzitutto alcuni dei modelli che classicamente sono stati ideati per dare una spiegazione dell'onnipresenza delle leggi a potenza in una vasta classe di fenomeni naturali. In seguito abbiamo trattato in modo più approfondito due classi di modelli in grado di ricreare power-laws nella loro distribuzione di probabilità. La caratteristica principale di questi processi risiede nella caratteristica variazione nel tempo dello spazio dei possibili risultati. In particolare, abbiamo considerato delle generalizzazioni

delle urne di Polya il cui spazio si espande con il tempo e i sample space reducing (SSR) processes il cui spazio al contrario si riduce progressivamente con l'evoluzione del sistema. Di questi ultimi abbiamo dato una formalizzazione in termini di Markov chain. Infine abbiamo proposto un modello di dinamica delle popolazioni capace di unificare e riprodurre i risultati dei tre SSR presenti.
Successivamente siamo passati all'analisi quantitativa dell'andamento del rank in frequenza sulle parole di alcuni testi. Infatti in questo caso si osserva questo non segue una pura legge a potenza ma ha un duplice andamento che può essere rappresentato da una legge a potenza che cambia esponente. Il nostro intento era quello di legare l'analisi dell'andamento del rank in frequenza con le proprietà topologiche di un network. In particolare, a partire da un corpus di testi abbiamo costruito un network di adiacenza dove ogni parola era collegata tramite un link alla parola successiva. Infatti nella nostra interpretazione il grafo doveva avere una struttura topologica particolare che abbiamo chiamato daisy. Tale configurazione è formata da una parte centrale e tanti petali esterni e sarebbe legata alle capacità e limitazioni della mente umana. Infatti nella nostra ipotesi la parte centrale della daisy è popolata dalle parole più frequenti e comuni, mentre i petali constano di parole rare e specifiche raggruppate in base al contesto a cui si riferiscono. Questa struttura è suggerita da alcuni limiti della mente umana, infatti ci aspettiamo che gli esseri umani conoscano approfonditamente solo alcuni argomenti e quindi i loro scritti possono popolare solo alcuni petali della daisy. Al contrario, la parte centrale del network è presente in ogni testo perché è caratterizzata dalle parole più comuni e di significato generale. Di conseguenza, le componenti interna ed esterne sarebbero legate al cambiamento di slope del rank di potenza. Notiamo che uno shuffle dei testi originerebbe a una struttura del grafo che presenta ancora una parte centrale e una esterna, quindi la doppia power-law sarebbe conservata, ma non sarebbe caratterizzato da parti esterne più e meno dense, cioè i petali non sarebbero più osservabili. Perciò pensiamo che il network abbia una duplice struttura: una forte legata solo alla frequenza delle parole e una più debole legata anche al loro significato.
Inizialmente, un primo studio della topologia del network ci ha confermato che

la sua struttura topologica potesse essere legata in qualche modo alla frequenza delle parole e al cambio di pendenza del rank in frequenza. Ispirati da questi risultati, successivamente abbiamo introdotto un metodo nuovo per individuare alcune componenti del network in base alla densità dei loro link. Applicando questa tecnica al grafo semplice e confrontando i risultati con quelli ottenuti sul network costruito sullo shuffle dei testi abbiamo trovato che alcune delle componenti sono legate all'ordine delle parole nel testo mentre altre sembrano essere legate soltanto alla frequenza. Perciò l'ipotesi iniziale di una struttura del network a forma di daisy sembra essere plausibile.
In conclusione, in questo elaborato abbiamo analizzato alcuni modelli capaci di ricreare leggi a potenza con un ampio range di esponenti. Abbiamo formalizzato alcuni di questi e mostrato come le catene di Markov e il teorema di Perron-Frobenius siano utili al fine di studiare le loro proprietà. Successivamente abbiamo studiato un network e correlato le sue proprietà topologiche al cambiamento di pendenza della legge a potenza che descrive il rank in frequenza delle parole trovando un risultato degno di nota: il network sembra avere una struttura in qualche modo legata sia all'ordine delle parole che alla loro frequenza. Questo risultato può portare ad alcuni sviluppi nell'ambito dello studio del linguaggio e della mente umana. Inoltre, siccome la struttura del network presenterebbe alcune componenti che raggruppano parole in base al loro significato, un approfondimento di questo studio potrebbe condurre ad alcuni sviluppi nell'ambito del text mining.

Introduction

Quantitative linguistics is a field of linguistics that applies statistical methods to the study of texts. It originated in the 19th century, when the first scholars started to count language elements of texts. For example, in Italy Filippo Mariotti studied the most important Italian poem, the *Divina Commedia* by Dante Alighieri applying basic statistical methods [14]. In his works he deepened the knowledge of the Italian language with the purpose of improving the emergent method of stenography.
With the same aim of optimizing stenography, at the beginning of the 20th century stenographers dusted off the quantitative studies of language deepening it. Worth mentioning is J.B. Estoup who studied the relationship between frequency and rank of words finding the result today known as Zipf's law [13]. G.K. Zipf gave a mathematical formula for Estoup's discovery and an interpretation of it based on the principle of the least effort [26, 25]. More in detail, he stated that the frequency rank distribution of words in texts follows a power-law with exponent -1. Eventually, the validity of Zipf's law was deeply studied by Mandelbrot, who brought new interest on it [12]. This revived activity was conveyed in finding new fields were Zipf's law could hold. It was discovered that it could be applied not only to linguistic but also to the population of cities, magnitude of earthquakes, the peak of gamma-ray intensity of solar flares and many others [17].
Meanwhile, the techniques applied to the study of texts and language became more complex and the simple count of words was supported by more refined mathematical methods. Stochastic processes started to take part into the lin-

guistics analysis, driven by Yule that used them for explaining the origins of power-laws [22]. Markov himself applied the stochastic process he defined, the Markov chains, to the linguistic study of *Eugene Onegin* by Aleksander Puskin [15].

In the meantime computational linguistics was emerging as a new science. It originated in the 50s in the USA with the purpose of machine translation and in few years it spread all over the world. Computers were showing their fast calculation skills and it was thought that they could be used to study also language. Since the mechanical translations did not achieve the expected results, the computational linguistics left the translations for the quantitative study of language. Applied to the calculation of Zipf's law, the computational power of computers allowed a wider and deeper study of the frequency rank distribution on a large scale of texts. It was proved that Zipf's law holds for texts written in different languages and dealing with different topics. However, it was found that it does not hold on every text length. In fact, on long texts the power-law known as Zipf's law has a change in its slope [10]. This evidence is observable on corpus with more than 10^4 different words, reason why it has not been detected before the use of computers.

Based on the application of mathematical methods, quantitative linguistics has kept growing with the development and refinement of those tools. In recent years it benefited from the studies of graph theory. As many other fields in mathematics, graph theory was originated from a problem that was posed by Euler in 1735. It is known as the Königsberg bridges problem and concerns the possibility of finding a path through all the seven bridges of the town of Königsberg without crossing any bridge twice [6]. From that moment on, the newborn field of graph theory was expanded and many correlations with other branches of mathematics were discovered. For example, the topological representation of graphs and the study of their geometric properties led to the birth of the topological graph theory.

Born of a problem, graph theory has been used to solve other problems in many fields. Since texts and language in general establish relations among words and

topics, graph theory has been useful to represent those links and to study their properties. For example, the semantic relations between concepts can be represented using a network, i.e. the semantic network. The study of it can lead us to the deepening of natural language analysis and text mining, with a possible consequent development of artificial intelligence and other natural language applications [19]. A second relation that leads to the definition of a graph is the order in which words appear in texts, i.e. the co-occurrence of words. The study of this second type of graph can help us outline the dynamics that lay under sentences construction and communication. For example, if the network built basing on a text presents substantial differences from the average probably the author of the text suffers from a language disorder. Therefore, the study of this kind of graphs can help in understanding the origins of some deficits [8].
t For example, if substantial differences in the structure of a network are observed a network structurally different from the usual one means that the author of the text suffers from some language disorders. Therefore, the study of it can help in understanding the origins of those deficits [8].

This thesis deals with Zipf's law in the context of language. The first chapter is intended as a recall of the needed elements of elementary probability, Markov chains and graph theory. In the second chapter we introduce two of the main laws of quantitative linguistics: Zipf's and Heaps' laws. As already mentioned, Zipf's law deals with the frequency of words in a text and states that the frequency rank distribution of the words follows a power-law with exponent -1. The same relation is found considering the number of new words with respect to the length of a text and it is called Heaps' law [11]. We explain the relationship between those two laws and recall Zipf's explanation in order to justify their omnipresence in several fields of science. We then introduce a model for that law that was clearly inspired by Zipf's principle of least effort [7]. Subsequently, we present a classical and trivial stochastic model, Simon's model, that is able to recall both laws [18]. Moreover we introduce the evidence of the changing slope of Zipf's law and recall a model that has been displayed to explain better this

double nature [10].
In the third chapter we define and analyze different stochastic models for generating Zipf's law. We take into account models that present an expansion in their sample space (the Urn models [20]) and a reduction of it (the SSR models [3, 4]). We write those models in terms of Markov chains and study their properties. Moreover, for the SSR models we also give an interpretation in terms of linear operators and show that this way it is possible to unify the three SSR processes under the same model.
Finally in the fourth chapter we apply the network theory for studying texts in relation with Zipf's changing slope. We build a network of co-occurrence of words and analyze it with the help of topological quantities, centrality measures and components detection. At last we try to link the measured quantities to a particular structure of the network that is related to Zipf's changing slope and the dynamics of sentences building.

Chapter 1

Mathematical background

This chapter is intended as a recall of all the mathematical concepts and results that will be used later on in our dissertation. It is divided in two main topics: the theory of probability and the theory of graphs. Since we are interested in studying some phenomena from the point of view of stochastic processes, the first section of this chapter will deal with main ideas and results of probability theory. We will start from the definition of a probability space, define the concepts of random variables, stochastic processes, entropy and Markov chains. Subsequently, we will get to the heart of the concept of Markov chains, recalling their properties and establishing a connection between them and their representation as graphs.

1.1 Introduction to probability theory

In order to understand the following dissertation, we need to introduce some classical ideas and result from the probability theory. We will start from the definition of the environment which every of the following concepts belongs to, the probability space. This concept is useful for modeling a particular class of real-world processes (or experiments).

Definition 1.1 (σ-algebra)**.** Let Ω be a set. A σ-algebra $\mathcal{A}$ on Ω is a collection of subsets A, A$\in \Omega$ s.t.

- $\Omega \in \mathcal{A}$;
- if $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$;
- if $A_i \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Remark 1. It is easy to prove that the power set of Ω, $\mathcal{P}(\Omega)$, is a σ-algebra on Ω. If not specified, from now on we will use $\mathcal{P}(\Omega)$ as σ-algebra.

Definition 1.2 (Probability measure function)**.** Let $\mathcal{A}$ be a σ-algebra on Ω. A probability measure function P on $(\Omega, \mathcal{A})$ is a function P: $\mathcal{A} \longrightarrow [0,1]$ s.t.

- $P(A) \geq 0 \quad \forall A \in \mathcal{A}$;
- $P(\Omega) = 1$;
- $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i), \quad A_i \in \mathcal{A} \forall i \quad A_i \cap A_j = \emptyset$ if $i \neq j$.

Definition 1.3 (Probability space)**.** A probability space is a triple $(\Omega, \mathcal{A}, P)$ where Ω is a non-empty set called sample space, $\mathcal{A}$ is a σ-algebra on Ω and P a probability measure function on $(\Omega, \mathcal{A})$.

In other words, a probability space is constituted by a set Ω of all the possible outcomes of the experiment we want to model, a collection of subsets of Ω, the σ-algebra $\mathcal{A}$, that characterizes groups of outcomes and a function, P, that assigns to every outcome or group of outcomes a number, its probability. For example, with this construction it is easy to model the experiment of tossing a fair coin. In fact, in this case $\Omega = \{Head, Tail\}$ because those are all the possible outcomes. We take $\mathcal{A} = \mathcal{P}(\Omega)$, that has cardinality $|\mathcal{P}(\Omega)| = 2^{|\Omega|} = 2^2 = 4$. In fact, $\mathcal{A} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ where H is for Head and T is for Tail. Obviously, the function P is defined on the elements of $\mathcal{A}$ as follows: $P(\emptyset) = 0$, $P(\{H\}) = P(\{T\}) = \frac{1}{2}$, $P(\{H, T\}) = 1$.
Every element $\omega \in \Omega$ is called elementary event, while the elements $A \in \mathcal{A}$ are called events. As we already said, given a probability space $(\Omega, \mathcal{A}, P)$ it is possible to compute the probability of every event and elementary event. However, sometimes it could be useful to calculate the probability of an event knowing that

another event has happened. This probability is different from the probabilities of the events but depends on them.

Definition 1.4 (Conditional probability)**.** Let $\mathcal{A}$ be a σ-algebra on Ω, if $A, B \in \mathcal{A}$, $P(A) > 0$ we define the conditional probability of A given B:

$$P(B|A) = \frac{P(A \cup B)}{P(A)}$$

Remark 2. Note that under the hypothesis of the definition, i.e. $A \in \mathcal{A}$ and $P(A) > 0$, $P(\cdot|A)$ is a probability measure function.

Definition 1.5 (Independent events)**.** In $(\Omega, \mathcal{A}, P)$ two events $A, B \in \mathcal{A}$ are independent if $P(A \cup B) = P(A)P(B)$.

Remark 3. As we said before, the conditional probability is a way of measuring the correlation between two events. More precisely, it follows from definitions 1.4 and 1.5 that if A and B are independent, $P(B|A) = P(B)$ and $P(A|B) = P(A)$.

Proposition 1.1.1 (Bayes' formula)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space. If $A, B \in \mathcal{A}$, $P(A) > 0$*

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

Proof. The probability of the event $A \cap B$ can be written in two different ways: $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B) = P(B|A)P(A)$. Now, with a simple algebraic passage it is possible to find Bayes' formula. □

We defined the probability of an event in $\mathcal{A}$ using a function, the probability measure function. If Ω is numerable, we can denote the probability of the events in $\mathcal{A}$ using a vector that shows some properties:

Definition 1.6 (Probability vector)**.** Let Ω be a numerable set with elements $\omega \in \Omega$. A function $p : \Omega \longrightarrow [0, 1]$ s.t. $\sum_\omega p(\omega) = 1$ is a probability vector.

From now on we may consider only finite sample spaces Ω.
Notation-wise, we will write $\vec{p} = (p_1 \dots p_n) = (p(\omega_1) \dots p(\omega_n))$ where $\Omega = \{\omega_1 \dots \omega_n\}$.

Being only a change in notation, it is obvious that the definition of probability using the probability measure functions and using the probability vectors are the same. More formally, it is possible to construct a bijection between the former and the latter.

Proposition 1.1.2. *There is a bijection between probability measure functions on numerable sets and probability vectors, i.e. if Ω is a numerable set, $\vec{p}$ a probability vector on Ω and P a probability measure function on $(\Omega, \mathcal{P}(\Omega))$, then $\vec{p}$ and P are equivalent.*

Proof. Let ω be elements in a subset A of Ω, hence $P(A) =: \sum_{\omega \in A} p(\omega)$. Vice versa, given a probability space $(\Omega, \mathcal{A}, P)$ the probability vector $\vec{p}$ is defined as: $p(\omega) := P(\{\omega\})$. □

1.1.1 Random variables

Now we are introducing a key concept in our dissertation and more in general in the probability theory, the random variables. Random variables are functions which have the set of possible outcomes as domain. They are used to model several phenomena that are governed by probability laws.

Definition 1.7 (Random variable)**.** Let (Ω, $\mathcal{A}$, P) be a probability space, χ a non-empty set and $\mathcal{F}$ a σ-algebra on χ. The function X: $\Omega \longrightarrow \chi$ is a random variable if $\forall A \in \mathcal{F}$, $X^{-1}(A) \in \mathcal{A}$.

In general, given the function X, we take χ as its image.

Remark 4. A random variable X induces a probability measure function P_X on $(\chi, \mathcal{F})$: if $A \in \mathcal{F}$ $P_X(A) := P(\{\omega \in \Omega | X(\omega) \in A\}) = P(X^{-1}(A))$. Therefore $(\chi, \mathcal{F}, P_X)$ is a probability space and P_X is called law or distribution of the stochastic variable X. We call $\vec{p}_X$ the correspondent probability vector.

Notation-wise, sometimes we will write $\vec{p}_X = P_X$.
Recalling the example of the tossing of a coin, we can model it using a random variable. We take $\Omega = \{0, 1\}$ and P s.t. $P(\{0\}) = \frac{1}{2}$, $P(\{1\}) = \frac{1}{2}$. The random

variable that models the experiment is:

$$X(\omega) = \begin{cases} T & \text{if } \omega = 0 \\ H & \text{if } \omega = 1 \end{cases}$$

This random variable induces a probability on its image $\chi = \{T, H\}$: $P_X(T) = \frac{1}{2}$, $P_X(H) = \frac{1}{2}$.

Definition 1.8 (Identically distributed variables)**.** Let X,Y be random variables. If they have the same distribution they are identically distributed.

In the case the codomain of a random variable χ is a subset of $\mathbb{R}$ we can define the expected value of that variable. This concept can be interpreted as the average outcome of the variable.

Definition 1.9 (Expected value)**.** Let X be a real random variable, $X : \Omega \longmapsto \chi \subseteq \mathbb{R}$. The expected value of X is defined as follows:

$$\mathbb{E}(X) = \sum_{x \in \chi} p(x)x$$

It is important to note that the random variable X must have real values, otherwise the product $p(x)x$ is not defined. Using the notion of expected value we can define:

Definition 1.10 (Variance)**.** Let X be a real random variable, $X : \Omega \longmapsto \chi \subseteq \mathbb{R}$. We can define the variance of it as

$$var(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

In other words, the variance of a random variable is the expectation of the squared deviation from its mean.

Definition 1.11 (Joint probability)**.** Let $(\Omega, \mathcal{A}, P)$ be a probability space and X, Y random variables, $X : \Omega \longrightarrow \mathcal{X}$, $Y : \Omega \longrightarrow \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ finite sets. Then the

joint probability is

$$P_{(X,Y)}(x,y) := P(X^{-1}(x) \cap Y^{-1}(y)) \quad x \in \mathcal{X}, y \in \mathcal{Y}$$

We will use the notation $p(x,y) = P_{(X,Y)}(x,y)$.

Remark 5. Knowing the joint probability $P(X,Y)$ of two random variables X and Y it is possible to calculate the probability associated to each of them, $P(X)$ and $P(Y)$:

$$p(x) = \sum_{y \in \mathcal{Y}} p(x,y) \qquad p(y) = \sum_{x \in \mathcal{X}} p(x,y)$$

It is possible to define the concept of independence also on random variables:

Definition 1.12 (Independence)**.** Let X,Y be stochastic variables on Ω, $X : \Omega \longrightarrow \chi$ with σ-algebra $\mathcal{F}_X$, $Y : \Omega \longrightarrow \mathcal{Y}$ with σ-algebra $\mathcal{F}_Y$. X and Y are independent if $\forall A \in \mathcal{F}_X, \forall B \in \mathcal{F}_Y$

$$P(X^{-1}(A) \cap Y^{-1}(B)) = P(X^{-1}(A))P(Y^{-1}(B))$$

Remark 6. Due to remark 4 and definition 1.11 we can rewrite the condition above: $P_{(X,Y)}(x,y) = P_X(x)P_Y(y)$ or $p(x,y) = p(x)p(y)$ $\forall x, y$.

Definition 1.13 (iid variables)**.** If X and Y are two independent and identically distributed random variables, they are called iid variables.

We can now generalize the conditional probability to the case of two random variables.

Definition 1.14 (Conditional distribution)**.** Let X,Y be random variables on $(\Omega, \mathcal{A}, P)$ and $p(x,y)$ the joint distribution. Then we define the conditional distribution

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{P(X^{-1}(x) \cap Y^{-1}(y))}{P(Y = y)}$$

Remark 7. Using the definition 1.12, if X and Y are independent the conditional probability is:

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x)$$

1.2 Stochastic processes

We are now defining one of the key concepts of probability theory: the stochastic processes. The stochastic processes are a collection of random variables that are usually used to model phenomena that vary in a random manner. For example, if we index the variables by time, they could represent numerical values of some system randomly changing over time.

Definition 1.15 (Stochastic process)**.** A stochastic process is a sequence of random variables $\{X_i\}_{i\in\mathbb{N}}$, $X_i : \Omega \longrightarrow \chi$.

Notation-wise, we will write $p(x_1, x_2 \dots x_n) = P(X_{i_1} = x_1, X_{i_2} = x_2 \dots X_{i_n} = x_n) = P(X_1^{-1}(x_1) \cap X_2^{-1}(x_2) \cap \dots \cap X_n^{-1}(x_n))$ where $\{x_1, x_2 \dots x_n\} \in \chi^n \ \forall n$.

Definition 1.16. The conditional distribution of n random variables is:

$$p(x_n|x_1 \dots x_{n-1}) = \frac{p(x_1 \dots x_n)}{p(x_1 \dots x_{n-1})}$$

Remark 8. As a consequence of the previous definition,

$$p(x_1, x_2 \dots x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_n|x_1 \dots x_{n-1})$$

Since a stochastic process is defined giving $p(x_1, x_2 \dots x_n)$, from remark 8 follows that a stochastic process is fully defined knowing $p(x_1)$ and $p(x_j|x_1 \dots x_{j-1}) \quad \forall j$. Being a stochastic process a collection of random variables, all the results and definitions that hold for random variables can be extended to stochastic processes.

Definition 1.17. Let $\{X_n\}_{n\in\mathbb{N}}$ be a stochastic process. The process is identically distributed if X_n are identically distributed $\forall n$.

Definition 1.18 (iid process)**.** If $\{X_n\}_{n\in\mathbb{N}}$ is a stochastic process s.t. the random variables X_n are independent and identically distributed $\forall n$, $\{X_n\}_{n\in\mathbb{N}}$ is called iid process.

Definition 1.19 (Stationarity)**.** A stochastic process is stationary if it is invariant under translations of time:

$$P(X_{i_1} = x_1, X_{i_2} = x_2 \dots X_{i_n} = x_n) = P(X_{i_1+h} = x_1, X_{i_2+h} = x_2 \dots X_{i_n+h} = x_n)$$

$\forall n \ \forall i_1 \dots i_n \ \forall x_1 \dots x_n \ \forall h.$

In other words, a stochastic process is stationary if it models a sequence of repeatable experiments. In particular, from definition 1.19 follows that $p(X_1 = x_1) = p(X_{1+h} = x_{1+h})$, that means that the probability of obtaining a particular outcome does not depend on the time .

Remark 9. Let $\{X_i\}_{i\in\mathbb{N}}$ stochastic process iid with probability vector $\vec{p}$, hence:

$$P(X_{i_1} = x_1, X_{i_2} = x_2 \dots X_{i_n} = x_n) = P(X_{i_1} = x_1)P(X_{i_2} = x_2) \dots P(X_{i_n} = x_n) =$$
$$= p(x_1)p(x_2) \dots p(x_n).$$

Therefore if we have an iid stochastic process, we only need the probability vector $\vec{p}$ to define the process.

1.2.1 Markov chains

A particular type of stochastic process is the Markov chain. It is a stochastic model that describes a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

Definition 1.20 (Markov chain)**.** A stochastic process defined by the probability $\vec{p} = p(x_1 \dots x_n) \, \forall n$ is a Markov chain if $p(x_j|x_1 \dots x_{j-1}) = p(x_j|x_{j-1}) \quad \forall j.$

Remark 10. Since a stochastic process is fully defined knowing $p(x_1)$ and $p(x_j|x_1, x_2 \dots x_{j-1})$ a Markov chain is fully defined by $p(x_1)$ and $p(x_j|x_{j-1}) \quad \forall j.$

Definition 1.21 (homogeneity)**.** A Markov chain is (temporally) homogeneous if

$$P(X_j = a|X_{j-1} = b) = P(X_2 = a|X_1 = b) \quad \forall j \qquad a, b \in \chi$$

. That is, $P(X_j = a|X_{j-1} = b)$ does not depend on $j \ \forall j$.

In other words, the probability of obtaining a particular outcome depends only on the previous outcome.
Notation-wise, we will write: $p(a|b) = P(X_j = a|X_{j-1} = b)$.
Note that every iid process is a Markov chain.

Proof. From definitions 1.16 and 1.18 follows that $p(x_j) = p(x_j|x_1 \dots x_{j-1})$.
It is easy to prove by induction that $p(x_j|x_{j-1}) = p(x_j)$, therefore the process is a Markov chain. □

In the following we shall always consider homogeneous Markov chains.
It is important to note that a Markov chain could be expressed in terms of matrices and vectors. These algebraic entities can be easily represented using graphs. In this sense, the algebraic representation of a Markov chain connects probability theory with graph theory.

1.2.2 Graphs and Markov chains

As already advanced, it is possible to convert all the features of Markov chains in terms of matrices and vectors. This is useful because all the characteristic elements of a Markov chain are written using algebraic entities that are easy to work with.
Due to remark 10, for identifying a Markov chain we need to know $p(x_1)$ and $p(x_j|x_{j-1}) \quad \forall j > 2$. Therefore in order to give an algebraic transcription of a Markov process we have to rewrite those quantities in terms of matrices and vectors.
Starting with $p(x_1)$, if we define $\mu_j = p(X_1 = j) \quad \forall j$ then obviously $\vec{\mu} = (\mu_1 \dots \mu_n)$ is a probability vector.
Notation-wise, given the set of outcomes $\chi = \{x_1 \dots x_n\}$ we identify it with the set of the first n natural numbers $\{1, 2 \dots n\}$. Therefore, with this notation $P(X_k = x_{i_k})$ is $P(X_k = i_k)$. For the first random variable, we will write $p(j)$ instead of $P(X_1 = j)$. Moreover, we put $P(X_n = j|X_{n-1} = i) = p(j|i) = p_{ij}$.

Remark 11. Supposing to have an homogeneous Markov chain, hence:

$$P(X_n = i_n | X_{n-1} = i_{n-1}) = p(i_n|i_{n-1}) = \sum_{i_1 \dots i_{n-2}} p(i_n|i_1 \dots i_{n-1}) =$$
$$= \sum_{i_1 \dots i_{n-2}} p(i_1)p(i_2|i_1) \dots p(i_n|i_1 \dots i_{n-1}) = \sum_{i_1 \dots i_{n-2}} p(i_1)p(i_2|i_1) \dots p(i_n|i_{n-1}) =$$
$$= \sum_{i_1 \dots i_{n-2}} \mu_{i_1} p(i_2|i_1) \dots p(i_n|i_{n-1}) = \sum_{i_1 \dots i_{n-2}} \mu_{i_1} p_{i_1 i_2} \dots p_{i_n i_{n-1}}$$

Therefore from remarks 10 and 11 follows that the Markov chain represented by the graph is fully defined knowing $\vec{\mu}$ and the conditional probabilities $p(i_j|i_{j-1}) \quad \forall j$.
Now we try to represent the conditional probabilities in terms of matrices.

Definition 1.22 (Stochastic matrix)**.** A stochastic matrix $\{P_{ij}\}_{i,j}$ is a matrix s.t. $P_{ij} \geq 0 \quad \forall i, j$ and $\sum_j P_{ij} = 1$.

An example of stochastic matrix is the transition probability matrix. We will use it to represent the transition probabilities of the Markov chain.

Definition 1.23 (Transition probability matrix)**.** Let $\{X_i\}_{i \in \mathbb{N}}$ be a Markov chain on χ with probability vector $\vec{p}$. Suppose $|\chi|$=n. The transition probability matrix of that Markov chain is a nxn matrix

$$P = \begin{pmatrix} & p(j|i) & \end{pmatrix} \longleftarrow i\text{-th row}$$
$$\uparrow$$
$$j\text{-th column}$$

Note that with the notation $p(j|i) = p_{ij}$, p_{ij} is exactly the element P_{ij} of the matrix.
As a consequence of remark 11, using the probability vector $\vec{\mu}$ and the transition probability matrix P we can represent the Markov chain in terms of algebraic

entities.
Now we introduce the concept of graph and reconnect it to the probability vectors and matrices just defined. This way we can see the properties of the Markov chain as properties of the graph and we can have a graphic representation of the process.

Definition 1.24 (Graph)**.** A graph is an ordered pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of vertices or nodes and $\mathcal{E}$ is a set of links, $\mathcal{E} = \{(i, j), i, j \in \mathcal{V}, i \text{ is connected to } j\}$. If the set $\mathcal{E}$ is constituted by ordered pairs the graph is called directed, i.e. the links have directions. Graphically, the links of this kind of graph are drawn as arrows that indicate the only direction allowed. If the pairs of $\mathcal{E}$ are not ordered, the graph is called undirected and every link allows moves in both directions.

Definition 1.25 (Strong connection)**.** A directed graph is strongly connected if there is a path between all the pair of vertices.
Every strongly connected component of a graph is called strongly connected component or SCC.

Definition 1.26 (Neighbours)**.** Considering the vertex i, its neighbours are all the vertex that are connected to i by an edge, regardless of the direction of the links (if they have one). In the case of a directed graph, we can divide the neighbours in in-neighbours and out-neighbours. The in-neighbours are the vertices that are connected to i with an exiting link, i.e. the link enters in i, and the out-neighbours are the vertices that are connected to i with an entering link, i.e. the link exits from i.

For example, considering the directed graph of figure 1.1, the neighbours of the vertex 2 are the vertices 1, 3, 4 and 5. While 3, 4 and 5 are its out-neighbours, 1 is its only in-neighbour.

We will use weighted directed graphs to represent Markov chains. The set of the nodes $\mathcal{V}$ is the set of the values in $\chi = \{x_1, x_2 \dots x_n\}$ and the set of the links is $\mathcal{E} = \{(x_i, x_j), x_i, x_j \in \chi, p(x_j|x_i) \neq 0\}$. The weight of the link that connects the vertex x_{j-1} with the vertex x_j is the transition probability $p(x_j|x_{j-1})$.
Note that in general if $p(x_j|x_{j-1}) \neq 0$ it is not guaranteed that $p(x_{j-1}|x_j) \neq 0$,

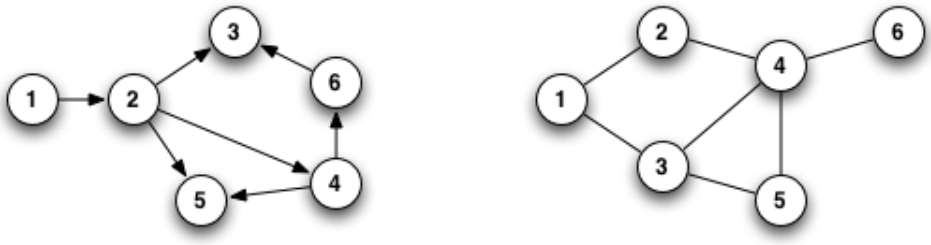


Figure 1.1: Right panel: example of an undirected graph. The lines represent the links and the numbers represent the vertex. It is allowed to move in both the directions of the links. For example, it is permitted to go from vertex 2 to vertex 4 and vice versa. Left panel: example of a directed graph. The links are represented by arrows with a direction. For example, in this graph it is allowed to move from vertex 2 to vertex 4 but not vice versa. In the case of the representation of a Markov chain with a directed graph, the arrows connect the previous result of the process with the current outcome.

i.e. it is allowed to move from vertex x_{j-1} to vertex x_j but not in the opposite direction. As a consequence, the graph that represent a Markov chain has to be directed.

For example, the graph in figure 1.2 is the graphic representation of a Markov

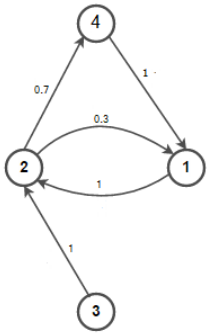


Figure 1.2: Example of a weighted directed graph. The labels on the arrows represent the weight of every link.

chain with $|\chi| = 4$ and with transition probability matrix:

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.3 & 0 & 0 & 0.7 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Note that being conditional probabilities the sum of the terms in every row of the conditional probability matrix has to be 1. This means that the sum of the weights of the links that start from a certain vertex has to be 1.
A path on a graph is a possible realization of the Markov chain represented by the graph: a n-steps path on the graph corresponds to a sequence of n+1 vertices that correspond to a realization of the Markov chain. In other words, considering the graph a sequence of vertices $(x_{i_1}, x_{i_2} \dots x_{i_n})$ represents the path that starts from the vertex x_{i_1}, passes trough the vertices x_{i_2}, x_{i_3} ...and ends in the vertex x_{i_n}. This corresponds to a realization of the Markov chain with probability $p(x_{i_1}, x_{i_2} \dots x_{i_n}) = P(X_1 = x_{i_1}, X_2 = x_{i_2} \dots X_n = x_{i_n})$.

Proposition 1.2.1. *P is a stochastic matrix $\Leftrightarrow \forall \vec{\mu}$ stochastic vector, $\vec{\mu}P$ is a stochastic vector.*

Proof.
$\boxed{\Rightarrow}$ The elements of the vector $\vec{\mu}P$ are ≥ 0 because they are obtained multiplying two numbers that are ≥ 0.
Now we have to prove the second condition, that is $\sum_{i=1}^{n} (\vec{\mu}P)_i = 1$:

$$\sum_{i=1}^{n} (\vec{\mu}P)_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \mu_j P_{ji} \right) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \mu_j P_{ji} \right) = \sum_{j=1}^{n} \mu_j \left(\sum_{i=1}^{n} P_{ji} \right) =$$
$$= \sum_{j=1}^{n} \mu_j = 1$$

Therefore $\vec{\mu}P$ is a stochastic vector.
$\boxed{\Leftarrow}$ We choose $\vec{\mu} = \delta^k$ where $\delta^k = (\delta_{1k} \dots \delta_{nk}) = (0 \dots 1 \dots 0)$ is the stochastic

vector with 1 in the k-th position and 0 elsewhere. In other words

$$\mu_j = \delta_{jk} = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{else} \end{cases}$$

Therefore $(\vec{\mu}P)_i = \left(\sum_j \mu_j P_{ji}\right) = \left(\sum_j \delta_{jk} P_{ji}\right) = P_{ki} \geq 0 \quad \forall k, i$ because $\vec{\mu}P$ stochastic vector.
Now we have to prove the second condition, so that $\sum_i P_{ki} = 1$. For hypothesis, $\sum_i (\mu P)_i = 1$, therefore $\sum_i P_{ki} = \sum_i \left(\sum_j \delta_{jk} P_{ji}\right) = \sum_i \left(\sum_j \mu_j P_{ji}\right) = \sum_i (\mu P)_i = 1$. □

Applying several time the proposition 1.2.1 we find the following

Corollary 1.2.2. *$\vec{\mu}P^2, \vec{\mu}P^3 \ldots \vec{\mu}P^k \quad \forall k$ are stochastic vectors.*

Proposition 1.2.3. *Given an homogeneous Markov chain identified by $(\vec{\mu}, P)$, it holds:*

1. $P(X_k = j) = (\vec{\mu}P^{k-1})_j \quad \forall k > 1$
2. $P(X_k = j | X_1 = i) = (P^{k-1})_{ij} \quad \forall k > 1$
3. $P(X_{k+h-1} = j | X_h = i) = (P^{k-1})_{ij} \quad \forall k > 1$

Proof.

1. $P(X_k = j)$ is the probability of being in the j-th vertex at time k.

$$P(X_k = j) = \sum_{i_1, i_2 \ldots i_{k-1}} P(X_1 = i_1, X_2 = i_2 \ldots X_k = j) =$$
$$= \sum_{i_1, i_2 \ldots i_{k-1}} P(X_1 = i_1) P(X_2 = i_2 | X_1 = i_1) P(X_3 = i_3 | X_2 = i_2) \ldots$$
$$\ldots P(X_k = j | X_{k-1} = i_{k-1}) = \sum_{i_1, i_2 \ldots i_{k-1}} \mu_{i_1} P_{i_1 i_2} P_{i_2 i_3} \ldots P_{i_{k-1} j} = \ldots$$
$$\ldots = \sum_{i_1} \mu_{i_1} (P^{k-1})_{i_1 j} = (\mu_1 \ldots \mu_n)(P^{k-1}) = (\vec{\mu}P^{k-1})_j$$

2. $P(X_k = j|X_1 = i)$ is the probability of arriving in the j-th vertex in exactly $k-1$ steps, starting from the i-th vertex.

$$P(X_k = j|X_1 = i) = \sum_{i_2 \dots i_{k-1}} P(X_2 = i_2|X_1 = i)P(X_3 = i_3|X_2 = i_2)\dots$$
$$\dots P(X_k = j|X_{k-1} = i_{k-1}) = \sum_{i_2 \dots i_{k-1}} P_{ii_2}P_{i_2i_3}\dots P_{i_{k-1}j} = (P^{k-1})_{ij}$$

   Where in the last equivalence we followed the procedure of proof 1.

3. $P(X_{k+h-1} = j|X_h = i)$ is the probability of leaving from the i-th vertex at time h and arriving in the j-th vertex after k steps.

$$P(X_{k+h-1} = j|X_h = i) = P(X_k = j|X_1 = i) = (P^{k-1})_{ij}$$

   Where in the first equivalence we used the definition of homogeneity and in the second equivalence the previous proof.

□

Proposition 1.2.4. *Let $(\vec{\mu}, P)$ be an homogeneous Markov chain. It is stationary if and only if $\vec{\mu} = \vec{\mu}P$*

This means that $\vec{\mu}$ is a left eigenvector of the matrix P with eigenvalue 1.

Proof.
$\boxed{\Rightarrow}$ Suppose that the Markov chain is stationary. Then

$$(\vec{\mu}P)_j = P(X_2 = j) = P(X_1 = j) = \vec{\mu}_j \quad \forall j \Rightarrow \vec{\mu} = \vec{\mu}P$$

where we applied the first point of theorem 1.2.3 and the definition of homogeneity.

$\boxed{\Leftarrow}$ We suppose that $\vec{\mu} = \vec{\mu}P$ and we prove that the Markov chain is stationary

$$
\begin{aligned}
&P(X_{1+k} = i_1, X_{2+k} = i_2 \dots X_{n+k} = i_n) = \\
&= \sum_{j_1, j_2 \dots j_k} P(X_1 = j_1, X_2 = j_2 \dots X_k = j_k, X_{1+k} = i_1, X_{2+k} = i_2 \dots X_{n+k} = i_n) = \\
&= \sum_{j_1, j_2 \dots j_k} \mu_{j_1} P_{j_1 j_2} P_{j_2 j_3} P_{j_{k-1} j_k} P_{j_k i_1} P_{i_1 i_2} \dots P_{i_{n-1} i_n} = \\
&= \left(\sum_{j_1} \mu_{j_1} \left(P^{k-1} \right)_{j_1 i_1} \right) P_{i_1 i_2} \dots P_{i_{n-1} i_n} = \\
&= \left(\vec{\mu} P^{k-1} \right)_{i_1} P_{i_1 i_2} P_{i_2 i_3} \dots P_{i_{n-1} i_n} \widehat{=} \mu_{i_1} P_{i_1 i_2} P_{i_2 i_3} \dots P_{i_{n-1} i_n} = \\
&= P(X_1 = i_1, X_2 = i_2 \dots X_n = i_n)
\end{aligned}
$$

where the hypothesis $\vec{\mu} = \vec{\mu}P$ was used in the equivalence marked with $\widehat{\ }$. $\square$

For example, we observed that an iid process is a Markov chain, therefore we can represent it with a graph or with the pair $(\vec{\mu}, P)$. $\vec{\mu}$ is the probability vector of the random variables X_i. It is well defined because the variables of the process are identically distributed.
Moreover, since the process is independent $P(X = i|Y = j) = P(X = i)$ i.e. $P_{ij} = \mu_i$. Therefore the transition probability matrix P has this form:

$$
P = \begin{pmatrix} | & | & | \\ \vec{\mu} & \vec{\mu} & \vec{\mu} \\ | & | & | \end{pmatrix}
$$

Is this process stationary? Using proposition 1.2.4 we need to calculate $\vec{\mu}P$:

$$
\begin{aligned}
&\vec{\mu}P = \vec{\mu} \begin{pmatrix} | & | & | \\ \vec{\mu} & \vec{\mu} & \vec{\mu} \\ | & | & | \end{pmatrix} = \mu_1 \vec{\mu} + \mu_2 \vec{\mu} + \ldots + \mu_n \vec{\mu} = (\mu_1 + \mu_2 + \ldots + \mu_n) \vec{\mu} = \\
&= 1 \cdot \vec{\mu} = \vec{\mu}
\end{aligned}
$$

Therefore we proved that an iid process is an homogeneous and stationary

Markov chain.

Definition 1.27. A non-negative matrix P is irreducible if $\forall i, j \quad \exists k$ s.t. $\left(P^k\right)_{i,j} \neq 0$.

Remark 12. We define the adjacency matrix $M = (M_{ij})_{i,j=1\dots N}$ where

$$M_{ij} = \begin{cases} 1 & \text{if } P_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Since P is non-negative, P is irreducible if and only if M is irreducible.
We can associate to M a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2 \dots N\}$ and $\mathcal{E} = \{(i, j), i, j \in \mathcal{V}, M_{ij} \neq 0\}$. M is irreducible if and only if exists always a path long k steps that connects every pair of vertices of the graph.

Note that if P is a stochastic matrix it is irreducible if and only if exists always a path long k steps that connects every pair of vertices of the graph associated with P.

Definition 1.28. Let P be a $n \times n$ matrix, $P \geq 0$. The period of a state i is defined as the GCD$\{k \in \mathbb{N} : (P^k)_{ii} > 0\}$. If the period is 1 then A is called aperiodic, otherwise it is periodic.

Theorem 1.2.5 (Perron-Frobenius for regular stochastic matrices)**.** *If P is a regular stochastic matrix* $\exists! \quad \vec{\pi}$,

1. $\vec{\pi} = \vec{\pi} P$
2. $\forall i = 1 \dots m \quad \lim_{n\to\infty} \left(P^n\right)_{ij} = \pi_j$

That is, exists only one set of initial conditions that makes the Markov chain defined by $(\vec{\mu}, P)$ stationary. The second point states that if the transition probability matrix of a Markov chain is regular, the process converges to its stationary distribution.
To prove this theorem we need to define a distance on the stochastic vectors space

$$\mathcal{S}^n = \left\{ \vec{\mu} = (\mu_1, \mu_2 \dots \mu_n), \mu_i \geq 0, \sum_i \mu_i = 1 \right\}$$

.

Definition 1.29. Let $\vec{\mu}$, $\vec{\mu}'$ be vectors in $\mathcal{S}^n$, we define the distance

$$d(\vec{\mu}, \vec{\mu}') = \frac{1}{2} \sum_{i=1}^{n} |\mu_i - \mu_i'|$$

Now we prove that the previous is a distance:

Proof. 1. $d(\vec{\mu}, \vec{\mu}') \geq 0$ because $|\mu_i - \mu_i'| \geq 0 \quad \forall i$ and the sum of non-negative quantities is non-negative.

2. $d(\vec{\mu}, \vec{\mu}') = d(\vec{\mu}', \vec{\mu})$. In fact $d(\vec{\mu}, \vec{\mu}') = \frac{1}{2} \sum_{i=1}^{n} |\mu_i - \mu_i'| = \frac{1}{2} \sum_{i=1}^{n} |\mu_i' - \mu_i| = d(\vec{\mu}', \vec{\mu})$.

3. $d(\vec{\mu}, \vec{\mu}') \leq d(\vec{\mu}, \vec{\nu}) + d(\vec{\nu}, \vec{\mu}')$. In fact since $|\mu_i - \mu_i'| \leq |\mu_i - \nu_i| + |\nu_i - \mu_i'|$, $d(\vec{\mu}, \vec{\mu}') = \frac{1}{2} \sum_{i=1}^{n} |\mu_i - \mu_i'| \leq \frac{1}{2} \sum_{i=1}^{n} |\mu_i - \nu_i| + \frac{1}{2} \sum_{i=1}^{n} |\nu_i - \mu_i'| = d(\vec{\mu}, \vec{\nu}) + d(\vec{\nu}, \vec{\mu}')$.

□

Remark 13. Note that $0 \leq d(\vec{\mu}, \vec{\mu}') \leq 1$

Proof. $d(\vec{\mu}, \vec{\mu}') \geq 0$ because it is a distance and $d(\vec{\mu}, \vec{\mu}') \leq 1$ because $\sum_i \mu_i = \sum_i \mu_i' = 1$. □

Proposition 1.2.6. *Let $\sum_i^+ \alpha_i$ be the sum of all the positive α_i. Then $d(\vec{\mu}, \vec{\mu}') = \sum_i^+ (\mu_i - \mu_i')$*

Proof. It is possible to prove that the sum of the positive addends is equal to the sum of the negative addends: $0 = 1 - 1 = \sum_i \mu_i - \sum_i \mu_i' = \sum_i (\mu_i - \mu_i') = \sum_i^+ (\mu_i - \mu_i') - \sum_i^+ (\mu_i' - \mu_i) \Leftrightarrow \sum_i^+ (\mu_i - \mu_i') = \sum_i^+ (\mu_i' - \mu_i)$.
Then $d(\vec{\mu}, \vec{\mu}') = \frac{1}{2} \sum_{i=1}^{n} |\mu_i - \mu_i'| = \frac{1}{2} \sum_i^+ (\mu_i - \mu_i') + \frac{1}{2} \sum_i^+ (\mu_i' - \mu_i) = \sum_i^+ (\mu_i - \mu_i')$. □

Lemma 1.2.7. *Let P be a stochastic matrix and $\vec{\mu}, \vec{\mu}' \in \mathcal{S}^n$, then:*

1. $d(\vec{\mu}P, \vec{\mu}'P) \leq d(\vec{\mu}, \vec{\mu}')$

2. If $\exists\alpha$ *s.t.* $P_{ij} \geq \alpha \quad \forall i,j \Rightarrow d(\vec{\mu}P, \vec{\mu}'P) \leq (1-\alpha)d(\vec{\mu}, \vec{\mu}')$

Proof. 1. Using the proposition 1.2.6 we have:

$$d(\vec{\mu}P, \vec{\mu}'P) = \sum_j{}^{+} ((\vec{\mu}P)_j - (\vec{\mu}'P)_j) = \sum_j{}^{+} \left(\sum_i \mu_i P_{ij} - \sum_i \mu'_i P_{ij} \right) =$$

$$= \sum_j{}^{+} \left(\sum_i (\mu_i - \mu'_i) P_{ij} \right) \leq \sum_j{}^{+} \sum_i{}^{+} (\mu_i - \mu'_i) P_{ij} = \sum_i{}^{+} (\mu_i - \mu'_i) \sum_j{}^{+} P_{ij}.$$

Since $\sum_j{}^{+} P_{ij} \leq 1$,

$$d(\vec{\mu}P, \vec{\mu}'P) \leq \sum_i{}^{+} (\mu_i - \mu'_i) \sum_j{}^{+} P_{ij} \leq \sum_i{}^{+} (\mu_i - \mu'_i) = d(\vec{\mu}, \vec{\mu}')$$

.

2. We just proved that $d(\vec{\mu}P, \vec{\mu}'P) \leq \sum_i{}^{+} (\mu_i - \mu'_i) \sum_j{}^{+} P_{ij}$. Now we observe that in this formula $\sum_j \neq \sum_j{}^{+}$, then $\exists j_0$ such that $\sum_j{}^{+}$ excludes this index. $P_{ij} \geq \alpha \quad \forall i,j \Rightarrow \sum_j{}^{+} P_{ij} \leq 1-\alpha$. Then $d(\vec{\mu}P, \vec{\mu}'P) \leq \sum_i{}^{+} (\mu_i - \mu'_i) \sum_j{}^{+} P_{ij} \leq (1-\alpha) d(\vec{\mu}, \vec{\mu}')$.

□

Remark 14. If P and Q are two stochastic matrices then $P \cdot Q$ is a stochastic matrix.

Proof. Obviously $(P \cdot Q)_{ij} = P_{ik} Q_{kj} \geq 0 \, \forall i,j$ because $P_{ik} \geq 0 \, \forall i,k, Q_{kj} \geq 0 \, \forall j,k$.

$$\sum_j (P \cdot Q)_{kj} = \sum_j \left(\sum_i P_{ki} Q_{ij} \right) = \sum_i P_{ki} \left(\sum_j Q_{ij} \right) = \sum_i P_{ki} = 1$$

□

Now we are ready to prove the Perron-Frobenius theorem.

Proof. Supposing P is a regular matrix with k=1, we prove the first point. Let's put $\vec{\mu}^{(n)} = \vec{\mu}P^n$, consequently $\vec{\mu}^{(n+l)} = \vec{\mu}P^{n+l} = \vec{\mu}^{(l)}P^n$. We want to prove that $\left\{\vec{\mu}^{(n)}\right\}_n$ is a Cauchy sequence with respect to the distance $d(\cdot,\cdot)$ defined

above. Since P is a regular matrix with k=1, we can use the lemma 1.2.7 with $\alpha = \min_{i,j} P_{ij} \geq 0$.

$$d(\vec{\mu}^{(n+l)}, \vec{\mu}^{(n)}) = d(\vec{\mu}^{(l)} P^n, \vec{\mu} P^n) \leq (1-\alpha)^n \, d(\vec{\mu}^{(l)}, \vec{\mu}) \leq (1-\alpha)^n \xrightarrow[n\to\infty]{} 0$$

Where in the last inequality we used the fact that $d(\cdot|\cdot) \leq 1$. Observe that we could use lemma 1.2.7 because as a consequence of remark 14, if P is a stochastic matrix then also P^n is a stochastic matrix.
We proved that $\left\{\vec{\mu}^{(n)}\right\}_n$ is a Cauchy sequence, then it has a limit, i.e. $\exists \vec{\pi} \in \mathcal{S}^n$ s.t. $\left\{\vec{\mu}^{(n)}\right\} \xrightarrow[n\to\infty]{d} \vec{\pi}$.
We observe that $\vec{\mu}^{(n+1)} = (\vec{\mu} P^n) P \xrightarrow[n\to\infty]{} \vec{\pi}$ but also $\vec{\mu}^{(n+1)} = (\vec{\mu} P^n) P \xrightarrow[n\to\infty]{} \vec{\pi} P$. Therefore $\vec{\pi} P = \vec{\pi}$.
Now we have to prove that $\vec{\pi}$ is unique. If $\exists \vec{\pi}_1, \exists \vec{\pi}_2$ s.t. $\vec{\pi}_1 = \vec{\pi}_1 P, \vec{\pi}_2 = \vec{\pi}_2 P$ then

$$d(\vec{\pi}_1, \vec{\pi}_2) = d(\vec{\pi}_1 P, \vec{\pi}_2 P) \leq (1-\alpha) d(\vec{\pi}_1, \vec{\pi}_2)$$

Since P is regular, $\alpha > 0$. Then $d(\vec{\pi}_1, \vec{\pi}_2) < d(\vec{\pi}_1, \vec{\pi}_2)$ that is absurd. For the prove of the second point we observe that $d(\vec{\mu} P^n, \vec{\pi} P^n) \leq (1-\alpha)^n \, d(\vec{\mu}, \vec{\pi}) \leq (1-\alpha)^n$. In particular $d(\vec{\mu} P^n, \vec{\pi}) \leq (1-\alpha)^n$ because $\vec{\pi}$ is a left eigenvector, i.e. $\vec{\pi} P^n = \vec{\pi}$.
Now we take $\vec{\mu} = (0 \ldots 1 \ldots 0)$ i.e. $\vec{\mu}_i = \begin{cases} 0 \text{ if } i \neq k \\ 1 \text{ if } i = k \end{cases}$. Then

$$(\vec{\mu} P^n)_j = \sum_i \mu_i P^n_{ij} = \sum_i \delta_{ik} P^n_{ij} = P^n_{kj} \xrightarrow[n\to\infty]{} \pi_j$$

Where we used the fact that $d(\vec{\mu} P^n, \vec{\pi}) \to 0$.
Now we suppose $k \neq 1$. As a consequence of remark 14, $Q = P^k$ is a regular stochastic matrix. We need to prove that $\vec{\mu}^{(n)}$ is a Cauchy succession and then we can use the previous procedure for finishing the proof.
We use two indices n_1, n_2 s.t. $n_2 \geq n_1 = qk + r$.

$$d(\vec{\mu}^{(n_2)}, \vec{\mu}^{(n_1)}) = d(\vec{\mu} P^{n_2 - n_1 + kq + r}, \vec{\mu} P^{kq+r}) = d(\vec{\mu}^{(n_2 - n_1 + r)} P^{kq}, \vec{\mu}^{(r)} P^{kq}) =$$
$$= d(\vec{\mu}^{(n_2 - n_1 + r)} Q^q, \vec{\mu}^{(r)} Q^q)$$

Since we defined $Q = P^k$, Q is regular with $k = 1$. We define $\alpha = \min_{i,j} Q_{ij} \geq 0$ and using the previous proof we find that $d(\vec{\mu}^{(n_2)}, \vec{\mu}^{(n_1)}) = d(\vec{\mu}^{(n_2-n_1+r)}Q^q, \vec{\mu}^{(r)}Q^q) \leq (1-\alpha)^q$.
We observe that if $n_1 \to \infty \Rightarrow n_2 \to \infty$. $n_1 \to \infty \Leftrightarrow q \to \infty \Rightarrow (1-\alpha)^q \to 0$. Therefore $\vec{\mu}^{(n)}$ is a Cauchy succession and we can prove the theorem as we did in the previous case. □

The vector $\vec{\pi}$ defined in the previous theorem has a peculiar shape in the case of bistochastic matrices.

Definition 1.30. A matrix P is bistochastic if it is stochastic and

$$\sum_i P_{ij} = 1$$

.

Proposition 1.2.8. *The matrix P is bistochastic if and only if $\vec{\pi}$ s.t. $\vec{\pi} = \vec{\pi}P$ is a uniform vector i.e. $\vec{\pi} = \left(\frac{1}{n}, \frac{1}{n} \dots \frac{1}{n}\right)$.*

Proof.
$\boxed{\Rightarrow}$ Suppose P is bistochastic. We want to prove that $\vec{\pi}P = \vec{\pi}$, i.e. $(\vec{\pi}P)_i = \vec{\pi}_i \quad \forall i$.

$$(\vec{\pi}P)_i = \sum_j \pi_j P_{ji} = \frac{1}{n}\sum_j P_{ji} = \frac{1}{n} = \vec{\pi}_i \quad \forall i$$

$\boxed{\Leftarrow}$ Let $\vec{\pi}$ be a uniform vector $\vec{\pi} = \left(\frac{1}{n}, \frac{1}{n} \dots \frac{1}{n}\right)$, we want to prove that P is bistochastic.
Since $\vec{\pi}_i = (\vec{\pi}P)_i$ we have:

$$\frac{1}{n} = \sum_j \pi_j P_{ji} = \frac{1}{n}\sum_j P_{ji} \Rightarrow \sum_j Pji = 1 \quad \forall i$$

□

The Perron-Frobenius theorem can be formulate in a more general way. In particular, we report the version that applies to irreducible non-negative matrices (see [16] for other versions of the theorem and their proofs).

Theorem 1.2.9 (Perron-Frobenius theorem for irreducible non-negative matrices)**.** *Let $P = (a_{ij}) \geq 0$ be a irreducible $n \times n$ matrix with period h. We write the set of the eigenvalues of P as $\Lambda := \lambda_1, \ldots, \lambda_n$ such that $\rho(P) = |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Then*

1. *$\lambda_1 \in \mathbb{R}^+$ (Perron root)*
2. *$\lambda_1, \lambda_2 = \lambda_1 e^{2\pi i \frac{1}{h}}, \lambda_3 = \lambda_1 e^{2\pi i \frac{2}{h}} \ldots \lambda_h = \lambda_1 e^{2\pi i \frac{h-1}{h}} \in \Lambda$; they all have algebraic multiplicity 1 and, obviously, $\lambda_1 = |\lambda_i| \, \forall i = 1 \ldots h$.*
3. *$\exists \, \vec{v} > 0$ and $\vec{w} > 0$ such that $A\vec{v} = \lambda_1 \vec{v}$ and $\vec{w}^T A = \lambda_1 \vec{w}^T$.*
4. *If $\vec{v}_i$ (or equivalently $\vec{w}_i$) is a right (left) eigenvector associated to λ_i and $\vec{v}_i > 0$ ($\vec{w}_i > 0$) $\Rightarrow i \in 1 \ldots h$.*
5. *If $h = 1$ then $\lim_{k\to\infty} \frac{P^k}{\lambda_1^k} = \vec{v}\vec{w}^T =: \mu$ where $\vec{v}$ and $\vec{w}$ are normalized such that $\vec{w}^T\vec{v} = 1$.*

Theorem 1.2.10. *Let $P = (a_{ij}) \geq 0$ be a irreducible $n \times n$ matrix and $\vec{v}$ be a left eigenvector of P. $\vec{v} \geq 0$ if and only if it is the eigenvector associated to the maximum eigenvalue.*

1.2.3 Shannon's entropy

We now introduce a function that is strictly related to information: the entropy. Shannon's entropy could be seen as the average rate at which information is produced by a stochastic source of data. For instance, when a low-probability event occurs it carries more information than an high-probability event. The reason lies in the fact that its happening is more surprisingly.
Shannon's entropy could be defined using the following

Theorem 1.2.11 (Shannon's entropy)**.** *Let X be a discrete random variable on $\chi = \{x_1, x_2 \ldots x_M\}$ with distribution $\vec{p} = (p_1, p_2 \ldots p_M)$. Exists a function unique up to a multiplicative constant that satisfies:*

1. *Monotonicity, i.e. $f(M) = H_M\left(\frac{1}{M}, \frac{1}{M} \ldots \frac{1}{M}\right)$ is increasing;*

2. *Extensivity, i.e.* $\forall L, M \geq 1 \qquad f(LM) = f(L) + f(M)$;

3. *Decomposition, i.e. let* $\vec{q} = (q_A, q_B)$ *be a probability vector,* $q_A = \sum_{i=1}^{r} p_i \quad q_b = \sum_{i=r+1}^{M} p_i \quad r < M$,

$$H_M(\vec{p}) = H_2(\vec{q}) + q_A H\left(\frac{p_2}{q_A}, \frac{p_2}{q_A} \dots \frac{p_r}{q_A}\right) + q_B H\left(\frac{p_r+1}{q_B}, \frac{p_r+2}{q_B} \dots \frac{p_M}{q_B}\right);$$

4. *Continuity.*

The function that satisfies the previous properties is:

$$H(\vec{p}) = -c \sum_{i=1}^{M} p_i \log p_i \tag{1.1}$$

where c is a constant. It is called Shannon's entropy.

From now on we assume that $0 \log 0 = 0$.

Proof.
We put $f(M) = H_M\left(\frac{1}{M}, \frac{1}{M} \dots \frac{1}{M}\right)$.

1. Monotonicity:

$$f(M) = H_M\left(\frac{1}{M}, \frac{1}{M} \dots \frac{1}{M}\right) = -c \sum_{i=1}^{M} \frac{1}{M} \log \frac{1}{M} = -c \log \frac{1}{M} = c \log M$$

It is increasing because $\log M$ is increasing as a function of M.

2. Extensivity:

$$f(LM) = H_{LM}\left(\frac{1}{LM}, \frac{1}{LM} \dots \frac{1}{LM}\right) = -c \sum_{i=1}^{LM} \frac{1}{LM} \log \frac{1}{LM} =$$
$$= -c \log \frac{1}{LM} = c \log(LM) = c(\log(L) + \log(M)) = H_M + H_L =$$
$$= f(L) + f(M)$$

3. Decomposition

$$H_M(\vec{p}) = -c\sum_{i=1}^{M} p_i \log p_i$$
$$H_2(\vec{q}) + q_A H\left(\frac{p_2}{q_A}, \frac{p_2}{q_A} \dots \frac{p_r}{q_A}\right) + q_B H\left(\frac{p_r+1}{q_B}, \frac{p_r+2}{q_B} \dots \frac{p_M}{q_B}\right) =$$
$$= q_A \log q_A + q_B \log q_B + q_a \sum_{i=1}^{r} \frac{p_i}{q_A} \log \frac{p_i}{q_A} + q_b \sum_{i=r+1}^{M} \frac{p_i}{q_B} \log \frac{p_i}{q_B}$$

And with some easy algebraic passages it is possible to pass from the first to the second equation and vice versa.

4. It is continuous because the logarithmic function is continuous.

Now we prove the uniqueness of the function. $\forall M, k \geq 1$ it holds

$$f(M^k) = f(M \cdot M^{k-1}) = f(M) + f(M^{k-1})$$

where we used the property *2.* Recursively we obtain:

$$f(M^k) = kf(M) \tag{1.2}$$

Now we prove by induction that $\forall M \geq 1$ integer, $\exists C$ s.t. $f(M) = C \log M$:

$$\text{If } M = 1 \quad f(1) = f(1 \cdot 1) = f(1) + f(1) = 2f(1) \Rightarrow f(1) = 0 = C \log 1$$

Now we need to prove it for $M > 1$. Notice that, since M^k is increasing as a function of k,

$$\forall r \geq 1 \qquad \exists k \text{ s.t. } M^k \leq 2^r \leq M^{k+1} \tag{1.3}$$

Since the function $f(M)$ is monotonic, using property *2.* we have:

$$f(M^k) \leq f(2^r) \leq f(M^{k+1}) \Rightarrow k \cdot f(M) \leq r \cdot f(2) \leq (k+1) \cdot f(M)$$

Therefore

$$\frac{k}{r} \leq \frac{f(2)}{f(M)} \leq \frac{k+1}{r} \tag{1.4}$$

Applying the logarithmic function to equation 1.3 we also find

$$M^k \leq 2^r \leq M^{k+1} \Rightarrow k \log M \leq r \log 2 \leq (k+1) \log M$$

Therefore

$$\frac{k}{r} \leq \frac{\log 2}{\log M} \leq \frac{k+1}{r} \tag{1.5}$$

From equations 1.4 and 1.5 we have:

$$\left| \frac{f(2)}{f(M)} - \frac{\log 2}{\log M} \right| \leq \frac{1}{r} \tag{1.6}$$

In the limit $r \to 0$,

$$\frac{f(2)}{f(M)} = \frac{\log 2}{\log M}$$

Therefore

$$f(M) = \frac{f(2)}{\log 2} \log M = C \log M \tag{1.7}$$

The next passage is to extend the previous result to a general probability vector. Let $\vec{p}$ be a rational probability vector, $p_i = \frac{r_i}{M} \quad i = 1 \dots N$. We consider the uniform probability vector $\left(\frac{1}{M}, \frac{1}{M} \dots \frac{1}{M}\right)$ and divide it into subvectors of length $r_i \quad i = 1 \dots N$. Using the property *3.*,

$$\begin{aligned} f(M) &= H\left(\frac{1}{M}, \frac{1}{M} \dots \frac{1}{M}\right) = H(p_1, p_2 \dots p_N) + \sum_i p_i H_{r_i}\left(\frac{1}{r_i} \dots \frac{1}{r_i}\right) = \\ &= H(p_1, p_2 \dots p_N) + \sum_i p_i f(r_i) \end{aligned} \tag{1.8}$$

From equation 1.7 we have that $f(M) = C \log M$ and $f(r_i) = C \log(r_i)$. There-

fore equation 1.8 becomes:

$$H(p_1 \dots p_N) = C\left(\log M - \sum_i p_i \log(r_i)\right) = C\left(-\sum_{i=1}^{N} p_i \log \frac{r_i}{M}\right) = \\ = -C\sum_{i=1}^{N} p_i \log p_i \tag{1.9}$$

Therefore we proved the theorem for uniform and rational probability vectors. Because of the continuity we can extend the result also to irrational probability vectors. □

Remark 15. In the definition of Shannon's entropy (theorem 1.2.11) the base of the logarithm is not specified. The reason lies in the fact that it is possible to use every base consequently adjusting the constant C. When it is not specified we are considering $\log_2$ and $C = 1$.

Since Shannon's entropy is a measurement for information, it is important to know its lower and upper bounds, that is the entropy associated to the most rare and surprisingly event and to the most common event.

Proposition 1.2.12. *$H(\vec{p})$ reaches its maximum value when $\vec{p}$ is the uniform vector.*

Proof. Let h be the function $h : \vec{p} \longrightarrow H(\vec{p}) = -\sum_{i=1}^{M} p_i \log p_i$, where $\vec{p}$ is a probability vector, $\vec{p} = (p_1, p_2 \dots p_M)$.
Since $\vec{p}$ is a probability vector, it must satisfy $\phi(\vec{p}) = \sum_{i=1}^{M} p_i - 1 = 0$. Using the method of Lagrange multiplier,

$$\partial_{p_j}(H(\vec{p} + \lambda\phi(\vec{\pi})) = 0 \Rightarrow -\log p_j - \frac{1}{ln2} + \lambda = 0 \Rightarrow -\log p_j = \lambda + \ln 2 \tag{1.10}$$

Therefore $\forall j \quad p_j$ constant because it does not depend on j. Since it has to be a probability vector, $\sum_j p_j = 1 \Rightarrow p_j = \frac{1}{M}$. □

Remark 16. Let $\vec{p} = (p_1, p_2 \dots p_M)$ be the uniform probability vector. We want to calculate the entropy of this vector, that is the maximum value that entropy

can reach.

$$H(\vec{p}) = -c\sum_{i=1}^{M} p_i \log p_i = -c\sum_{i=1}^{M} \frac{1}{M}\log\frac{1}{M} = \log(M)$$

It is important to note that the maximum of Shannon's entropy is a function that varies with the cardinality of the probability space Ω.

Proposition 1.2.13. *$H(\vec{p})$ reaches its minimum value when $\vec{p} = (p_i)_{i=1\dots M}$ is s.t. $\exists! j, \quad j = 1 \dots M$*

$$\begin{cases} p_j = 1 \\ p_i = 0 \quad \forall i = 1 \dots M, i \neq j \end{cases}$$

Proof. We note that $p_i \geq 0 \quad \forall i = 1 \dots M$ because $\vec{p}$ is a probability vector. From the definition of Shannon's entropy (theorem 1.2.11) follows that $H(\vec{p}) \geq 0 \quad \forall \vec{p}$. If $\vec{p}$ has one entry 1 and the others 0, $H(\vec{p}) = 0\log 0 + \dots + 1\log 1 + \dots + 0\log 0 = 0$, therefore this is the minimum. □

The two previous results are in line with intuition: the most uncertain case is when all the outcomes are equiprobable, while the event that carries the least surprise is a certain event.

Using the concept of entropy, we can introduce a new function, the Kullback-Leibler divergence. It is also called Kullback-Leibler distance because it is a measure of how one probability distribution is different from a second, reference probability distribution.

Definition 1.31 (Kullback-Leibler divergence)**.** Let $\vec{p}$ and $\vec{q}$ be probability vectors, we define the Kullback-Leibler distance as:

$$D_{KL}(\vec{p}||\vec{q}) = -\sum_x p(x)\log q(x) + \sum_x p(x)\log p(x) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

With the conventions $0\log\frac{0}{q} = 0$ if $q \neq 0$, $0\log\frac{0}{0} = 0$, $0\log\frac{p}{0} = \infty$.
It is important to remark that in spite of the name this is not a distance because

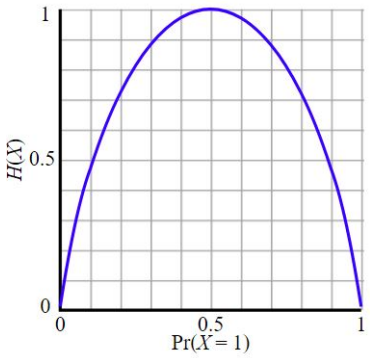


Figure 1.3: Entropy $H(X)$ of a coin flip. $Pr(X = 1)$ represent the probability of obtaining head. It is clear that the maximum is obtained when $Pr(X = 1) = \frac{1}{2}$ that is when the probability vector is uniform. The uniform vector describe the case of a fair coin and in this case the maximum is $H(X) = \log |\Omega| = \log 2 = 1$.

it is not symmetric and does not satisfy the triangle inequality.

Now we introduce a classical result of probability theory: Jensen's inequality. We will use it to prove some fundamental properties of Kullback-Leibler divergence.

Definition 1.32. A function $f : \mathbb{R} \longmapsto \mathbb{R}$ is convex if $\forall \lambda$, $0 \leq \lambda \leq 1$, $\forall x_1, x_2 \in \mathbb{R} \quad f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$

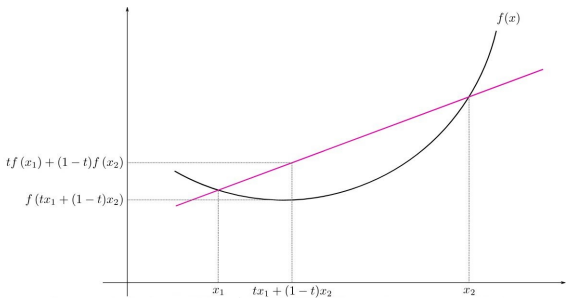


Figure 1.4: An example of convex function. In this representation, t has the same role of λ in the definition.

Proposition 1.2.14 (Jensen's inequality)**.** *Let $f : \mathbb{R} \longmapsto \mathbb{R}$ be a convex function and X a random variable. It holds that $\mathbb{E}(f(X)) = f(\mathbb{E}(X))$.*

Where the expected value is calculated on the random variable $f(X) = f \circ X$.

Proof. We proceed by induction on the codomain of X, $|\chi|$. If $|\chi| = 2$, $\vec{p}_X = (p_1, p_2) = (\lambda, 1-\lambda)$ where we put $p_1 = \lambda$. Being f convex we have:

$$\mathbb{E}(f(X)) = p_1 f(x_1) + p_2 f(x_2) = \lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2) = f(\mathbb{E}(X))$$

Now we suppose the Jensen's inequality to be true for every $\vec{p} = (p_1, p_2 \dots p_{n-1})$ and we prove it for $\vec{p} = (p_1, p_2 \dots p_n)$. Putting $\lambda = p_n$ we find:
If $\lambda = 1$, $X(\omega) = x_1$ or $X(\omega) = x_2$ with probability 1 that is X is the trivial random variable. In this case we have

$$\mathbb{E}(f(X)) = \sum_{i=1}^{n} p_i f(x_i) = p_n f(x_n) = \lambda f(x_n) = f(x_n) = f(\mathbb{E}(X))$$

If $\lambda \neq 1$,

$$\mathbb{E}(f(X)) = p_n f(x_n) + \sum_{i=1}^{n-1} p_i f(x_i) = \lambda f(x_n) + (1-\lambda) \sum_{i=1}^{n-1} p'_i f(x_i)$$

where $p'_i = \frac{p_i}{1-\lambda}$. $\sum_i p'_i = \sum_i \frac{p_i}{1-\lambda} = 1$, therefore $(p'_1, p'_2 \dots p'_n)$ is a probability vector. Because of the induction hypohtesis, $f(\mathbb{E}_{k-1}(X)) \leq \mathbb{E}_{k-1}(f(X))$, therefore:

$$\mathbb{E}(f(X)) = \lambda f(x_n) + (1-\lambda) \sum_{i=1}^{n-1} p'_i f(x_i) \geq \lambda f(x_n) + (1-\lambda) f\left(\sum_{i=1}^{n-1} p'_i x_i \right) \geq$$
$$\geq f\left(\lambda x_n + (1-\lambda) \sum_{i=1}^{n-1} p'_i x_i \right) = f\left(\lambda x_n + \sum_{i=1}^{k-1} p_i x_i \right) = f(\mathbb{E}(X))$$

□

Now we we can prove some of the most important properties of the Kullback-Leibler divergence:

Proposition 1.2.15.

1. $D_{KL}(\vec{p}||\vec{q}) \geq 0 \, \forall \vec{p} \forall \vec{q}$
2. $D_{KL}(\vec{p}||\vec{q}) = 0 \Leftrightarrow \vec{p} = \vec{q}$

Proof.

1. We define the set $A = \{x \in \chi | p(x) > 0\}$. Since the logarithm is a convex function, using Jensen's inequality we find

$$- D_{KL}(\vec{p}||\vec{q}) = -\sum_{x\in\chi} p(x) \log \frac{p(x)}{q(x)} = -\sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x\in A} p(x) \log \frac{q(x)}{p(x)} \leq$$
$$\leq \log\left(\sum_{x\in A} p(x)\frac{q(x)}{p(x)}\right) = \log\left(\sum_{x\in A} q(x)\right) \leq \log\left(\sum_{x\in\chi} q(x)\right) = 0$$

   Hence, $-D_{KL}(\vec{p}||\vec{q}) \leq 0 \Rightarrow D_{KL}(\vec{p}||\vec{q}) \geq 0$

2. $$D_{KL}(\vec{p}||\vec{q}) = 0 \Leftrightarrow \sum_x \log \frac{p(x)}{q(x)} \Leftrightarrow \frac{p(x)}{q(x)} = 1 \Leftrightarrow p(x) = q(x)$$

□

Using the Kullback-Leibler divergence it is possible to define a new quantity connected to information, the mutual information.

Definition 1.33 (Mutual information)**.** Let X, Y be random variables on Ω and $p(x, y)$ their joint probability. We put $q(x, y) = p(x)p(y)$, so that $q(x, y)$ would be the joint probability if the random variables X, Y were independent. The mutual information of the two variables X and Y is defined as

$$I(X;Y) = D_{KL}(p(x,y)||q(x,y)) = D_{KL}(p(x,y)||p(x)p(y))$$

.

In other words, the mutual information measures the distance between the true law of the two variables and the law they would have if they were independent. As an automatic consequence we have that $I(X;Y) = 0 \Leftrightarrow X, Y$ are independent. This is one of the most important properties of the mutual information:

Proposition 1.2.16. *The mutual information has the following properties:*

1. $I(X;Y) = 0 \Leftrightarrow X, Y$ *are independent;*
2. $I(X;Y) = I(Y;X)$*;*
3. $I(X;X) = H(X)$*.*

Proof.

1. Using the properties of the Kullback-Leibler divergence,

$$I(X;Y) = 0 \Leftrightarrow D_{KL}(p(x,y)||p(x)p(y)) = 0 \Leftrightarrow \vec{p} = \vec{q} \Leftrightarrow p(x,y) = p(x)p(y)$$

   And by definition this happens if and only if X and Y are independent.

2. $I(X;Y) = D_{KL}(p(x,y)||q(x,y)) =_{KL} (q(x,y)||p(x)p(y)) = D_{KL}(q(x,y)||p(x,y)) = I(Y;X)$

3. Let's suppose $Y = X$. Hence

$$p(x,y) = \begin{cases} p(x) & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

   Therefore

$$I(X;X) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \sum_{x} p(x) \log\left(\frac{p(x)}{p(x)p(x)}\right) =$$
$$= \sum_{x} p(x) \log \frac{1}{p(x)} = -\sum_{x} p(x) \log p(x) = H(X)$$

□

Chapter 2

Word-frequency laws and the classical models

We are now introducing the most famous and classical word frequency laws, Zipf's and Heaps' laws. Both are empirical laws that have been discovered while investigating texts features, but they can be applied to several fields of knowledge, such as population analysis, biology, music.

2.1 Zipf's law

Zipf's law is an empirical law that historically deals with the frequency of words in the written language. In the first formulation of his law, Zipf considered the number of words that occurred exactly n times in a text, $N(n)$ and found:

$$N(n) \sim n^{-\varsigma} \tag{2.1}$$

Where exponent ς varies from text to text but it is approximately 2. Equation 2.1 points out that the number of words that occur n times in a text is related to the number of its occurrences in the same sample.
A second formulation of his law involves the occurrence ranking of words: suppose to rank the words in the sample in decreasing order by their number of occurrences, so that the most frequent word has rank 1, the second most fre-

quent has rank 2 and so on. Let r be the rank of a word and n the number of its occurrences, then

$$n(r) \sim r^{-z} \tag{2.2}$$

where usually $z \approx 1$. That is, the relation between the occurrence of a word and its rank follows with a power-law with exponent -1.

Remark 17. The latter version of Zipf's law can be reformulated in terms of the frequency of words as a function of rank, $f(r) = \frac{n(r)}{N}$ where N is the total number of words of the sample. Having divided $n(r)$ by a constant, relation 2.2 still holds:

$$f(r) \sim r^{-z} \tag{2.3}$$

Since $z \approx 1$, we can write $f(r) \sim \frac{1}{r}$. This means that the frequency of any word is inversely proportional to its rank in frequency, i.e. the first ranked word occurs approximately twice as often as the second ranked word that occurs three times more than the third, etc.

It is interesting to note that Zipf's law holds in texts written in different language. Zipf himself gave examples of his law in English, Latin and Peiping Chinese dialect.

Zipf gave an explanation of this law applying the principle of least effort to the effort the speaker and the hearer put into a conversation in order to communicate efficiently. From the speaker's perspective the most efficient vocabulary consists of only one word that covers all the possible meanings, but it could be impossible for the hearer to determinate the particular meaning of the word in that specific context. The struggle between the speaker's inclination to reduce the vocabulary and the hearer's tendency to expand it ends with the development of a vocabulary where a few words are used very frequently while most words occur just a few times.

For better understanding the mechanisms that lead to Zipf's law, several researchers have proposed models for it. One of them has a peculiarity: it was inspired by Zipf's least effort interpretation.

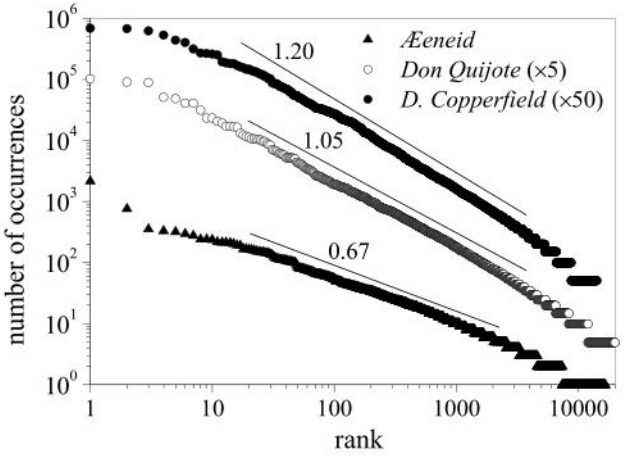


Figure 2.1: Zipf's law for the words of *Aeneid* (in Latin), *Don Quijote* (in Spanish) and *David Copperfield* (in English). In the figure, straight lines have the slopes that fit each data set and the slopes are indicated in the labels. A relation of the type $\frac{1}{r^z}$ is called power-law and in a log-log plot is represented by a straight line. Being an empirical law, the power-law dependence develops for intermediate values of r, as seen in the figure.

2.1.1 Ferrer i Cancho and Solé's model

Ferrer i Cancho and Solé used Zipf's least effort interpretation to build a mathematical model for Zipf's law. They tried to quantify the process by which a vocabulary diversifies as communication evolves under the pressure of the principle of least effort on both speaker and hearer. In their model, the process of communication implies the exchange of information about a collection of k objects $\{m_1, m_2 \dots m_k\}$, the meanings, and a set of l words $\{w_1, w_2 \dots w_l\}$. A binary $l \times k$ matrix $A = \{a_{ij}\}_{ij}$ is created as follows:

$$a_{ij} = \begin{cases} 1 & \text{if word } w_i \text{ is used to refer to meaning } m_j \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

It establishes the connections between words and meaning. We now observe that it is allowed to have more than one word referring to the same meaning, so that there may be several $a_{ij} = 1$ for the same value of j. The sum $\sigma_j =$

$\sum_i a_{ij}$ is the number of synonyms referring to meaning m_j. Let $p(w_i, m_j)$ be the joint probability that word w_i is used when the communication is referring to the meaning m_j, and assume that all meanings are referred to with the same probability, i.e. $p(m_j) = \frac{1}{k} \quad \forall j$. The conditional probability to see the word w_i knowing that the meaning of the communication is m_j is $p(w_i|m_j) = \frac{a_{ij}}{\sigma_j}$. Using Bayes formula we have:

$$p(w_i) = \sum_j p(w_i, m_j) = \sum_j p(m_j)p(w_i|m_j) = \frac{1}{k}\sum_j \frac{a_{ij}}{\sigma_j} \tag{2.5}$$

The entropy associated to this probability,

$$H_{speaker} = -\sum_{i=1}^{l} p(w_i)\log_l p(w_i) \tag{2.6}$$

is a suitable definition for the speaker's communication effort because its minimum $H_{speaker} = 0$ is reached on a single-word vocabulary, i.e. $\vec{\mu} = (p(w_1), p(w_2)\ldots p(w_l)) = (1, 0\ldots 0)$ and its maximum $H_{speaker} = 1$ is reached when every word is equiprobable, i.e. $\vec{\mu} = \left(\frac{1}{l}, \frac{1}{l}\ldots\frac{1}{l}\right)$.
The conditional probability $p(m_j|w_i)$ is the probability that a person that hears the word w_i will infer the meaning m_j, $p(m_j|w_i) = \frac{p(w_i,m_j)}{p(w_i)}$. The weighted sum of the entropies associated with the distribution $p(m_j|w_i)$ over all the words heard is an indicator of the hearer's effort:

$$H_{hearer} = -\sum_{i=1}^{l} p(w_i)\sum_{j=1}^{k} p(m_j|w_i)\log_k p(m_j|w_i) \tag{2.7}$$

H_{hearer} varies between zero and one and it can be interpreted as a measure of the average noise (or indeterminacy) with which information reaches the hearer. Now we define the total cost of communication as the weighted sum of both the efforts:

$$\Omega(\lambda) = \lambda H_{hearer} + (1-\lambda)H_{speaker} \tag{2.8}$$

where $\lambda \in (0, 1)$ is a parameter.
If Zipf's hypothesis were valid, the probabilities $p(w_i)$ would converge to a distribution compatible with the inverse relation between frequency and rank for some intermediate value of λ.
The mutual information between the probability distributions of words and meanings could be a measure for communication accuracy:

$$I(w, m) = \sum_{j=1}^{k} p(m_j) \sum_{i=1}^{l} p(w_i|m_j) \log_l p(w_i|m_j) - \sum_{i=1}^{l} p(w_i) \log_l p(w_i) \qquad (2.9)$$

as well as the relative lexicon size, L, defined as the ratio between the number of effectively used words and the total number of available words l. In figure 2.2

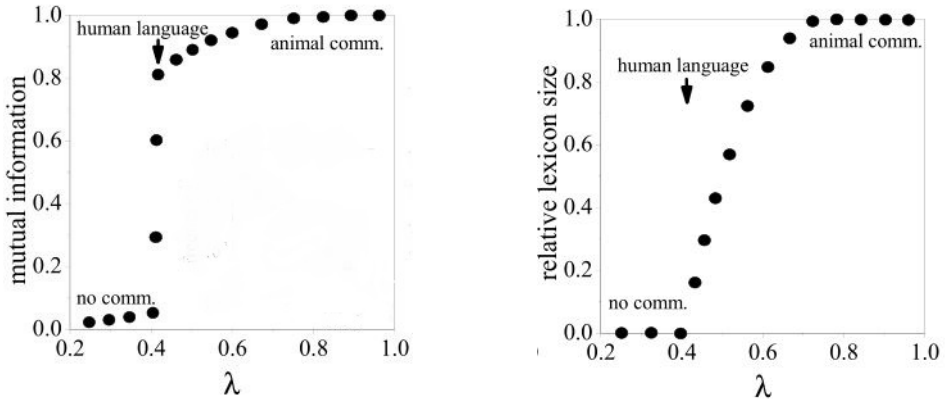


Figure 2.2: Left panel: mutual information $I(w, m)$ as a function of λ. Right panel: relative lexicon size L as a function of λ. Two distinct regimes are clearly identified, separated by a sharp transition at $\lambda \approx 0.41$.

it is possible to see that both the indicators change when λ changes. In particular, there is a sharp transition in correspondence of the value of λ, $\lambda^* = 0.41$. For $\lambda < \lambda^*$ there is practically no informational correlation between words and meanings, therefore communication fails. Accordingly, the relative lexicon size L is zero. For $\lambda > \lambda^*$ both $I(w, m)$ and L have significant levels, and approach their maximal values for $\lambda \to 1$. As a consequence, human language appears to have been tuned by the principle of least effort at the edge of the transition

between unworkable and feasible communication.

As already said, several models of Zipf's law have been created in order to understand the reasons that lie under this evidence. Some of them are based on mathematical assumptions and rely on the theory of stochastic processes. One of the first models of this type was created in the 1950s by Herbert Simon, a sociologist and economist. He proposed a mathematical model that conceives text production as a stochastic process. It is based on a few simple dynamical rules that explain the appearance of algebraic relation such as Zipf's law in many other phenomena.

2.1.2 Simon's model for Zipf's law

For his model, Simon considered the process of text generation as a sequence of events where one word is added at every step. Let $N_t(n)$ be the number of different words that appear exactly n times when the text has reached the length of t words. For example, if there are 354 words that have occurred exactly once each in the first t words $N_t(1) = 354$. The subsequent step follows these rules:

1. Let α be a constant, $0 \leq \alpha \leq 1$. With probability $1 - \alpha$ at step $t + 1$ a word that already appeared in the text is added. The word is chosen with a probability proportional to $nN_t(n)$, that is the total number of occurrences of all the words that have appeared exactly n times.

2. With probability α at step $t + 1$ is added a new word.

The second rule leaves open the possibility that, among the words that occurred exactly n times, the probability of recurrences of some words may be higher than some others. These rules describe a stochastic process in which the probability of writing one word at the next step depends on the probability of the words previously written and on a constant α. From 1. follows that

$$\mathbb{E}\left[N_{t+1}(n)\right] - N_t(n) = C(t)\left[(n-1)N_t(n-1) - nN_t(n)\right] \quad \forall n = 2 \ldots t+1 \quad (2.10)$$

because if the $(t+1)$-th word is chosen among the words that occurred $n-1$ times, then $N_{t+1}(n)$ will increase and the probability of this choice is proportional to the number of words that occurred exactly $n-1$ times, $(n-1)N_t(n-1)$. Instead if the $(t+1)$-th word is chosen among the words that occurred n times, i.e. $N_{t+1}(n)$ will decrease. This choice will happen with probability $nN_t(n)$. In the other cases, $N_{t+1}(n) = N_t(n)$.
Similarly, for n=1 we find:

$$\mathbb{E}\left[N_{t+1}(1)\right] - N_t(1) = \alpha - C(t)N_t(1) \tag{2.11}$$

Approximating $\mathbb{E}\left[N_t(n)\right]$ by $N_t(n)$, the previous equations become:

$$N_{t+1}(n) - N_t(n) = C(t)\left[(n-1)N_t(n-1) - nN_t(n)\right] \quad \forall n = 2 \dots t+1 \tag{2.12}$$

$$N_{t+1}(1) - N_t(1) = \alpha - C(t)N_t(1) \tag{2.13}$$

Now we want to evaluate the factor of proportionality $C(t)$. Since $C(t)nN_t(n)$ is the probability that the $(t+1)$-th word is one that previously occurred n times, we have:

$$\sum_{n=1}^{t} C(t)nN_t(n) = C(t)\sum_{n=1}^{t} nN_t(n) = 1 - \alpha$$

Now we observe that $\sum_{n=1}^{t} nN_t(n) = t$ because is the total number of words at step t, hence

$$C(t) = \frac{1-\alpha}{t} \tag{2.14}$$

Therefore the recursive relation is:

$$N_{t+1}(1) - N_t(1) = \alpha - \frac{1-\alpha}{t}N_t(1) \tag{2.15}$$

$$N_{t+1}(n) - N_t(n) = \frac{1-\alpha}{t}\left[(n-1)N_t(n-1) - nN_t(n)\right] \quad \forall n = 2 \dots t+1 \tag{2.16}$$

Equations 2.15 and 2.16 do not have an asymptotic, t-independent solution. However, a steady-state solution can be found assuming that for large t holds $\frac{N_{t+1}(n)}{N_t(n)} = \frac{t+1}{t} \quad \forall n, t$, that is all the frequencies grow proportionately with t.

Under this hypothesis it is possible to find a stationary profile $P(n)$ for $N_t(n)$ such that $N_t(n) = tP(n)$.

$$P(n) = \frac{\alpha}{1-\alpha} B(n, \zeta) \tag{2.17}$$

Where $B(n, \zeta)$ is the Beta function and $\zeta = 1 + (1-\alpha)^{-1}$.
For small values of α ($\lesssim 0.1$) and for all $n \geq 1$ the solution for the profile $P(n)$ is well approximated by the power-law function

$$P(n) \approx \frac{\alpha}{1-\alpha} \Gamma(\zeta) n^{-\zeta} \tag{2.18}$$

where $\Gamma(\zeta)$ is the Gamma function. Then $N(t)$ has the form of Zipf's law as written in equation 2.1 or in equation 2.2 with $z = 1 - \alpha$.
Since the probability of appearance of new words must be larger than 0, $z < 1$ and the characteristic value $z = 1$ is obtained for $\alpha \to 1$, i.e. when the appearance of new words becomes extremely rare. In real texts this condition happens when texts are long. However, there are some samples of natural language where the best fitting of the frequency-rank relation yields $z > 1$, for example *Don Quijote* and *David Copperfield*, as shown in Fig. 2.1. In the original form Simon's model is not able to explain power-law exponents z larger than one, but extensions of this model that work for larger values of z have been proposed.

Remark 18. Note that $N_t(n) = tP(n)$ with $P(n)$ given by equation 2.18 is an exact solution of Simon's model equations 2.15 and 2.16 with the initial condition $N_{t_0}(n) = t_0 P(n)$, but it is not the general solution.

The constitutive equations of Simon's model, equations 2.15 and 2.16, can be seen as the average evolution law deriving from a special case of a very general additive-multiplicative stochastic process

$$n_{t+1} - n_t = a_t + b_t n_t \tag{2.19}$$

where n_t is the number of occurrences of a word at step t and a_t and b_t are random variables drawn, at each step, from suitably chosen probability distributions $f(a)$ and $g(b)$. In fact, rule 2. defines an additive process by which the number of words with $n = 1$ grows stochastically, at a constant average rate α. Rule 1. describes a stochastic reinforcement in the occurrence of words: words that have already appeared a large number of times are more likely to be used again than those that are rarer. Let $p_t(n)$ be the probability that at step t the stochastic variable has value n which, in our context, is proportional to the number of words with exactly n occurrences. Supposing $f(a) \neq 0$, for large t and for a wide range of values of n, we have:

$$p_t(n) \sim n^{-1-\gamma}$$

where γ is determined by

$$\int g(b)(1+b)^{\gamma} db = 1$$

This points out that power-law distributions are inherent to generic additive-multiplicative stochastic processes, hence is not so surprisingly to find them in a large variety of disparate systems, as long as they are driven by random events. The process of creation any meaningful text is obviously not a sequential random choice of words, but a long chain of words, even though grammatically correct, is comprehensible only knowing the context which the sentence is referring to. Context emerges whit the growth of the text: as words are successively added to the text, a context is built up which favors the later appearance of some words, in particular those that have already been used and inhibits the use of others. This behaviour is expressed through Simon's model equations 2.15 and 2.16. Therefore even if the process of text creation is is not driven by random choices, Simon's model is able to capture its main features in order to represent it.

2.2 Heaps' law

Heaps' law is one of the most famous word frequency laws and deals with the dependence of the number of different words on the length of a given text. It states that if V is the number of different words and T the length of the text, then

$$V \sim T^{\nu} \tag{2.20}$$

with $0 \leq \nu \leq 1$. Therefore this relation is still a power-law relation and its representation on a log-log plot is a straight line. This law could be extended to the way in which the total number of different words grows as a text progresses. Within this interpretation, Heaps' law states that the rate of appearance of new words α decays with the text length

$$\alpha(t) = \alpha_0 t^{\nu-1} \tag{2.21}$$

where $\alpha_0 < 1$ is a constant.

Intuitively there is a correlation between Zipf's law exponent z and Heaps' law

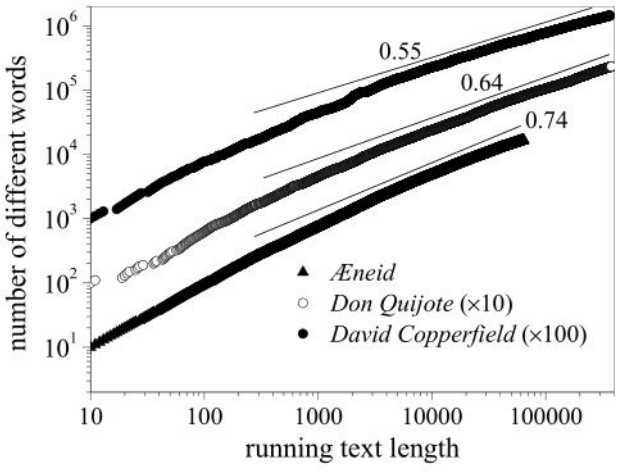


Figure 2.3: Heaps' law for the words of *Aeneid*, *Don Quijote* and *David Copperfield*. Straight lines have the slopes that fit each data set and the slopes are indicated in the labels.

exponent ν: large values of z correspond to small ν and vice versa. For instance in a Latin text there will be much more different words than in an English text of the same length, because Latin has noun declension and different verb conjugations while in English the same verb form is used for several forms and tenses, the same noun for different cases and so on. Considering for example *AEneid* and *David Copperfield* we note that while the lexicon of the Latin poem is larger that that of the English poem by 15%, the latter work is almost six times longer than the former. Due to the structure of the language, it is expected that the number of different words grows faster than in English, so Heaps's law exponent ν will be larger. On the other hand, the total number of word of the Latin poem is distributed among a larger number of different words and, therefore, the frequency-rank distribution has a flatter profile. Hence, Zipf's law exponent z should be smaller.
This evidence could lead us to think that Zipf's and Heaps' laws are strictly correlated. Actually, under some hypothesis of random-sampling it is possible to derive Heaps' law from Zipf's law.

2.2.1 Correlation between Zipf's and Heaps' laws

Supposing that the frequency-rank distribution has a strict power-law behaviour $f(r) \sim r^{-z}$, we can derive Heaps' law knowing only Zipf's law. For this purpose we construct a sequence of elements by randomly sampling from this Zipf distribution $f(r)$. We can calculate $f(r)$ from the following approximated integral:

$$\int_0^{r_{max}} f(\tilde{r})d\tilde{r} = 1 \tag{2.22}$$

Distinguishing two cases basing on the values of z we have:

$$f(r) = \frac{1-z}{r_{max}^{1-z}-1} r^{-z} \text{ if } z \neq 1 \tag{2.23}$$

and

$$f(r) = \frac{1}{\log(r_{max})} r^{-1} \text{ if } z = 1 \tag{2.24}$$

Considering equation 2.23, if $z > 1$ we can neglect the term r_{max}^{1-z} and if $z < 1$ we can write $r_{max}^{1-z} - 1 \sim r_{max}^{1-z}$. Therefore:

$$\begin{aligned} \text{if } z > 1, \qquad & f(r) \simeq (z-1)r^{-z} \\ \text{if } z = 1, \qquad & f(r) \simeq \frac{r^{-1}}{ln(r_{max})} \\ \text{if } 0 < z < 1, \qquad & f(r) \simeq (1-z)\frac{r^{-z}}{r_{max}^{1-z}} \end{aligned} \tag{2.25}$$

For deriving Heaps' law we have to estimate the number of distinct elements V appearing in the sequence as a function of its length T. Let's suppose that after the entrance in the sequence of a new element (never appeared before) the number of distinct elements is V. This new element will have a rank $r_{max} = T$ and frequency $f(r_{max}) = \frac{1}{V}$. From equation 2.25 we have:

$$\begin{aligned} \text{if } z > 1, \qquad & f(V) \simeq (z-1)V^{-z} = \frac{1}{T} \\ \text{if } z = 1, \qquad & f(V) \simeq \frac{1}{VlnV} = \frac{1}{T} \\ \text{if } 0 < z < 1, \qquad & f(V) \simeq \frac{1-z}{V^{1-z}}V^{-z} = \frac{1}{T} \end{aligned} \tag{2.26}$$

Inverting these relations we find:

$$\begin{aligned} \text{if } z > 1, \qquad & V \simeq T^{\nu} \text{ with } \nu = \frac{1}{z} \\ \text{if } z = 1, \qquad & V \simeq \frac{T}{lnT} \text{ with } \nu \simeq 1 \\ \text{if } 0 < z < 1, \qquad & V \simeq T \text{ with } \nu = 1 \end{aligned} \tag{2.27}$$

Therefore, supposing $f(r) \simeq r^{-z}$ we find Heaps' law $V(T) \sim T^{\nu}$ with the following relation between the exponents z and ν:

$$\begin{aligned} \text{if } z > 1, \qquad & \nu = \frac{1}{z} \\ \text{if } 0 < z \leq 1, \qquad & \nu = 1 \end{aligned} \tag{2.28}$$

It is interesting to notice that also a generalized version of Simon's model is

capable to explain this inverse relation between the exponents ν and z.

2.2.2 Simon's model for Heaps' law

Generalizing Simon's model it is possible to establish a relation between the Zipf and the Heaps exponents, finding the intuitive result just explained. Moreover, admitting that the probability of occurrence of new words can vary along the text, i.e. $\alpha = \alpha(t)$ the model can also predict that the Zipf exponent z could be larger than 1.
Considering the variables t and n of equations 2.15 and 2.16 as continuous quantities and denoting $N_t(n) = N(n,t)$, we have:

$$\frac{\partial N}{\partial t}(1,t) = \alpha(t) - \frac{1-\alpha(t)}{t} N(1,t) \tag{2.29}$$

$$\frac{\partial N}{\partial t}(n,t) = -\frac{1-\alpha(t)}{t}\frac{\partial}{\partial n}\left[nN(n,t)\right] \quad \forall n = 2 \ldots t+1 \tag{2.30}$$

The solution of the first equation is:

$$N(1,t) = N(1,t_0)\epsilon(t) + \epsilon(t)\int_{t_0}^{t} \frac{\alpha(t')}{\epsilon(t')}dt' \tag{2.31}$$

where

$$\epsilon(t) = e^{-\int_{t_0}^{t} \frac{1-\alpha(t')}{t'}dt'}$$

Assuming that the probability of the occurrence of new words is given by the Heap's law, as in equation 2.21, then α decays following a power law. Hence, for large values of t we have $1-\alpha(t) \approx 1$. In this limit, the general solution for equation 2.29 is

$$N(1,t) = At^{-1} + \frac{\alpha_0}{\nu+1}t^{\nu} \tag{2.32}$$

with A an arbitrary constant. Therefore the dominant contribution is a growing power of t, $N(1,t) \approx \frac{\alpha_0}{1+\nu}t^{\nu}$.

The general solution of equation 2.30 is

$$N(n,t) = \frac{1}{n} H\left(\frac{n}{t}\right) \tag{2.33}$$

where $H(\frac{n}{t})$ is an arbitrary function. We can use the solution of the first equation, equation 2.31, as a boundary condition for the solution of the second, in the limit $n \to 1$. Within this limit, the solution of equation 2.30 is:

$$N(n,t) = \frac{\alpha_0}{\nu+1} t^{\nu} n^{-1-\nu} \tag{2.34}$$

Taking into account that $N(n,t)$ and $n(r)$ are related according to $r = \int_n^{\infty} N(y,t)dy$, the Zipf exponent resulting from equation 2.34 are $\zeta = 1 + \nu$ and $z = \frac{1}{\nu}$. Consequently, the frequency-rank Zipf exponents z is larger than 1 and exhibits a simple inverse relation with the Heaps exponent ν.
It is important to note that the same stochastic dynamical rules may be useful to portray many of the phenomena which display power-law distributions in their statistical properties. In fact, from an abstract perspective Simon's model describes the growth in size of certain object classes, with a growth rate proportional to the class size itself. Therefore it could be applied to every field that have a sort of vocabulary that can be divided in classes. For instance the vocabulary could correspond to people and the class could be the country of origin of every person. The stochastic multiplicative growth is added with a random process by which new classes are created at a fixed rate.

2.3 The Zipf changing slope

As already mentioned, deviations of Heaps' and Zipf's laws are observed in the tails of Heaps' and Zipf's plots (respectively for large T and r). For example, with regard to Zipf's law formulated in equation 2.3, the scaling of this law has to break for large r because of the divergence of the harmonic series. In fact, if r is large enough $\sum_{r=1}^{N} f(r) > 1$, but the sum of the frequency can never be larger than one.

Several models have been created for fitting also the tails of the plots and two researchers, Gerlach and Altmann proved that for English the distribution with two power laws is the best fit for all databases with more than 10^9 words. Moreover they proposed a stochastic growth model for fitting those power-laws that needs only two free parameters. A particular feature of the model is that its parameters depend only on the language, therefore the universality of the original Zipf's law is preserved. In fact several models have been developed for fitting particular databases but they have parameters that depend on some peculiarity of the texts, as the topic, the size or date of publication.

2.3.1 Altmann and Gerlach's model

As already said, the best fit for English is the distribution with two power-laws (double power-law, dp):

$$f_{dp}(r;\gamma,b) = C \begin{cases} r^{-1} & \text{if } r \leq b \\ c(b,\gamma) r^{-\gamma} & \text{if } r > b \end{cases} \tag{2.35}$$

where $C = C(b,\gamma)$ is the normalization constant and b, γ are free parameters. The critical rank $r = b$ determines a transition from Zipf's original law to a second power law with exponent γ.

As we showed in the previous section, the Zipf and the Heaps exponent are inverse. We can use this correlation to adjust equation 2.35 for fitting Heaps' law:

$$V_{dp}(T;\gamma,b) = C_n \begin{cases} T & \text{if } T \ll T_b \\ T_b^{1-\frac{1}{\gamma}} T^{\frac{1}{\gamma}} & \text{if } T \gg T_b \end{cases} \tag{2.36}$$

where T_b is the number of words such that $V(T_b) = b$ and C_n is the scaling constant, $C_n = \frac{C}{n}$, $C \approx f(1)$ is the frequency of the most common word. The previous equations 2.35 and 2.36 where inferred from empirical analysis. At a later time, Gerlach and Altmann created a generative model for giving an interpretation of their empirical findings. Firstly, they divided words in two classes, core and non-core vocabularies. The total number of word is $V = N_c + N_{\bar{c}}$,

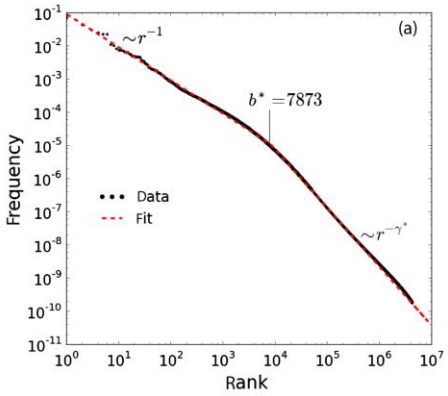


Figure 2.4: The double scaling behaviour of the rank-frequency distribution on the Google Ngram database in English. In red, the best fitting distribution with two power laws. b^* is the critical rank that determines the transition from the first power-law to the second.

where N_c is the number of core words and $N_{\bar{c}}$ is the number of non-core words. At every step a word (i.e. word token) is drawn and attributed to one of the distinct words (i.e. word type) depending on a probability. With probability p_{new} the word token is a new type and with probability $1 - p_{new}$ is an already existing type (see figure 2.6). In the latter case, a previously used word type is attributed to the word token at random with a probability proportional to the number of times this word type has occurred before. In the former case, with probability p_c the new word type originate from a finite set of N_c^{max} core words and with probability $1 - p_c$ can come from a potentially infinite set of non-core words.

In a first approximation, we consider p_c constant, $p_c \lesssim 1$, $p_c = 0$ only if all non-core words were drawn ($N_c = N_c^{max}$):

$$p_c(N_c) = \begin{cases} p_c^0 & \text{if } N_c < N_c^{max} \\ 0 & \text{if } N_c = N_c^{max} \end{cases} \tag{2.37}$$

We also choose p_{new} and p_c to depend on N_c and $N_{\bar{c}}$ (and therefore on V) because an increase in V necessarily reflects that fewer undiscovered words exist. On the contrary an increase in T is strongly affected by repetitions of frequently used

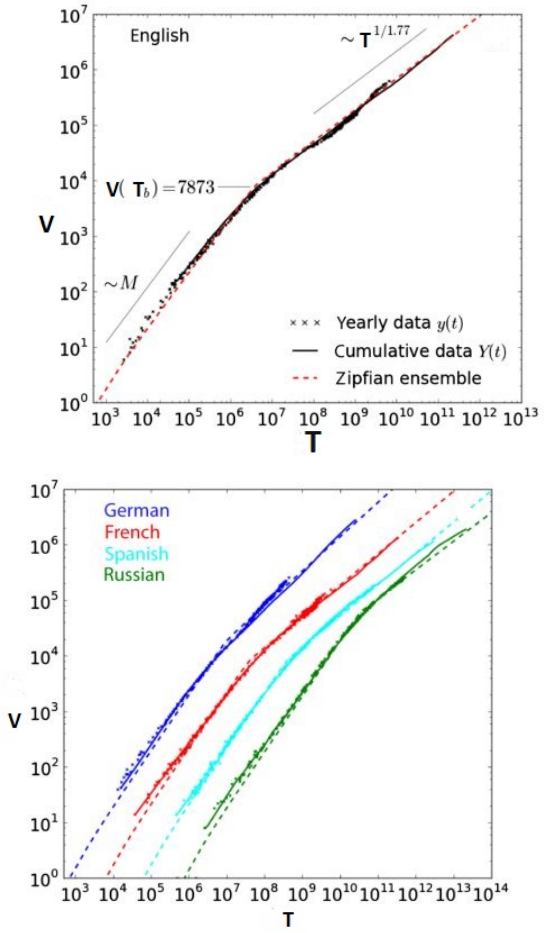


Figure 2.5: Vocabulary as a function of database size (Heaps' plot) on the Google Ngram database in English (left panel) and in other four languages (right panel). It is important to remark that the model is valid for databases that change in time (left panel) and for different languages (right panel).

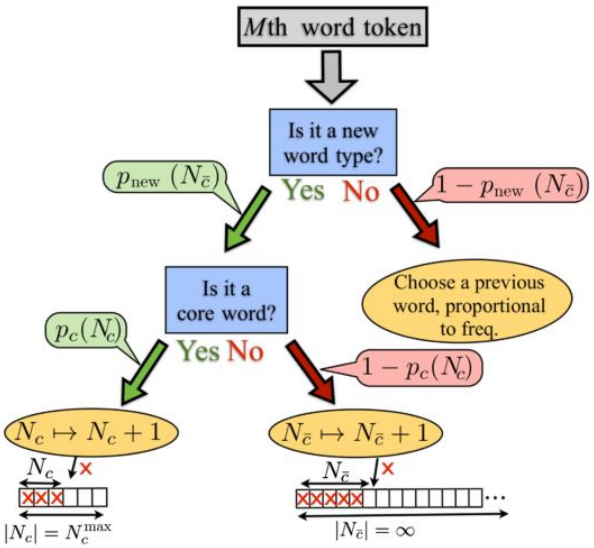


Figure 2.6: Scheme of the Gerlach and Altmann's model.

words. Moreover we put $p_{new} = p_{new}(N_{\bar{c}})$ because by definition core words are necessary in the creation of any text. Therefore the usage of a new core word should be expected and should not affect the probability of using a new (non-core) word type in the future. On the other hand, if a non-core word is used for the first time, the combination of this word and the previously used words leads to a combinatorial increase in possibilities of expression of new ideas with the already used vocabulary and thus to a decrease need for additional new words, i.e. p_{new} should decrease with $N_{\bar{c}}$.

After a new occurrence of a new non-core word we update p_{new}:

$$p_{new} \mapsto p_{new}\left(1 - \frac{\alpha}{N_{\bar{c}} + s}\right) \tag{2.38}$$

with $\alpha > 0$ decay rate and $s \gg 1$ constant that softens the reduction of p_{new} for small $N_{\bar{c}}$ (we use $s = N_c^{max}$).

It is possible to recover equations 2.35 and 2.36 from this model. For example at the very beginning, when $N \ll N_c^{max}$ (so $N_c \ll N_c^{max}$ and $N_{\bar{c}} \ll N_c^{max}$), we can assume that it is much more probable to draw core words than non-core words,

i.e. $1 - p_c^0 \ll 1$ because at the very beginning of vocabulary growth most of the new word types belongs to the set of core words. Therefore it follows from equations 2.38 and 2.37 that $p_{new} \approx$ const, hence $V \sim T$.

In this chapter we introduced two of the most important word frequency laws, explaining the reasons why researchers focused on the creation of mathematical models for them. In addition to this we analyzed some classical proposes that take into account different aspects: Ferrer i Cancho and Solé's model is based on the least effort principle, Simon's model relies on stochastic processes and Gerlach and Altmann's model is connected to the process of formation of texts. In the next chapter we will present and study some other models that use stochastic processes to interpret the creation of texts and more generally try to find a probability structure in language manifestations. In particular, we will talk about models for Zipf's and Heaps' laws that consider expanding sample probability space.

Chapter 3

Sample-space-varying models

This chapter focuses on modern stochastic models that reproduce power-laws in general and Zipf's and Heaps' laws in particular. As already mentioned in the previous chapter, several models have been proposed to understand the dynamics that lie under the emergence of power-laws in nature. We already introduced a classic stochastic model, Simon's model, that reproduces Heaps' and Zipf's laws. In this chapter we present some more recent stochastic models that have the sample space changing (reducing or expanding) as main feature. Those models are based on the idea of modeling a process that changes his possible states with time, reducing it (SSR processes) or expanding it (generalization of Polya's urn model). Those two antithetical ideas were born from the observation of two different kinds of phenomena: the reducing space originates to recreate a system that has constraints that become tighter with time, while the expanding space models a phenomena whose possibilities of expanding grow with time.
First, we describe three models characterized by the reduction of their sample space that reproduce the power-laws with different exponents. Subsequently we give a formalization that unifies those three models and perform its theoretical study. Eventually we introduce the works by Tria et al. showing that they recall both Zipf's and Heaps' laws.

3.1 SSR processes

The SSR processes have been inspired by history-dependent systems characterized by the reduction of their sample space. In other words, the set of the possible outcomes of these phenomena changes over time, reducing as they age. One example of history-dependent system with sample space reduction is the creation of sentences: while the first word of a sentence can be chosen from the space of almost all the existing words, the choice of the second has grammar and contextual constraints that become stronger the more the text length grows. This is the main idea behind the sample space reducing (SSR) processes, a class of models that are able to represent the features of history-dependent processes and that leads to power-laws in the rank distribution of their outcomes.
The discussion of the SSR processes follows the lines of the works by Corominas-Murtra et al. [3, 4], reference also for the figures of this section.

3.1.1 Simple SSR processes

The simple sample space reducing (SSR) process could be illustrated by a set of N fair dice with different number of faces. The first has one face, the second has two faces and so on to the N-th that has N faces. To start the process we take the dice with N faces and throw it, getting a result K. Then we throw the dice with K-1 faces getting a new result. Once we reach the dice with one face we restart the process by throwing the N-faced dice again (see figure 3.1).
In an equivalent way, we can interpret the process considering a staircase with N steps: imagine a ball that randomly falls downstairs but never can climb to higher levels. The ball first hits any of the N steps with uniform probability, $P_N(i) = \frac{1}{N} \quad \forall i = 1 \dots N$. Then it can only fall down to a lower level with uniform probability and so on until it reaches the bottom step. Once landed on the first step, it restart jumping randomly to any of the steps (see figure 3.2). Therefore, if at time t the ball is in the i-th step $i \neq 1$, at time $t+1$ all lower levels $j < i$ can be reached with the same probability. It is forbidden to go upstairs, hence $P(j|i) = 0$ if $j > i$.

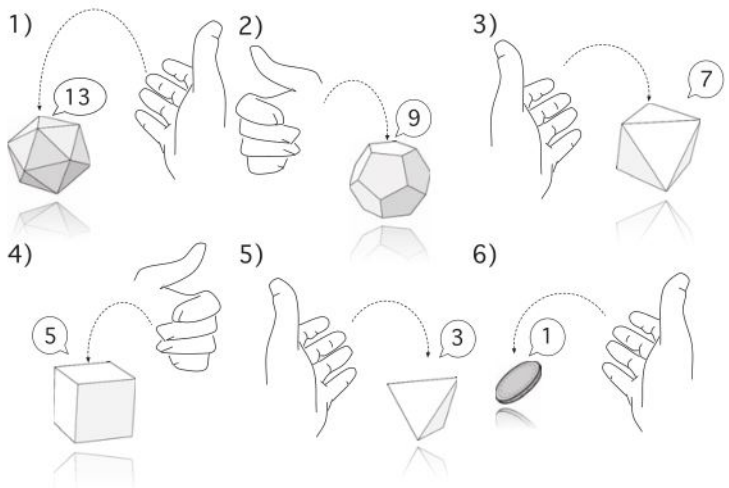


Figure 3.1: Representation of a SSR process with N=20. If the process is repeated many times, the distribution of face-values (rank-ordered) gives Zipf's law.

Now we are interested in the occupation probability, i.e. how often a given site i is occupied on average. The probability of jumping on the level i depends on the probability of being on an higher level. In other words

$$P(i) = \sum_{j=1}^{N} P(i|j)P(j)$$

We want to prove that $P(i) = c\frac{1}{i} \quad \forall i = 1 \dots N$. From the previous equation we have

$$\begin{aligned} P(i) &= \sum_{j=i+1}^{N} P(i|j)P(j) + P(i|1)P(1) = \\ &= c\frac{1}{N}\frac{1}{N-1} + c\frac{1}{N-1}\frac{1}{N-2} + \ldots + c\frac{1}{i+1}\frac{1}{i} + c\frac{1}{N} = \\ &= c\left(\frac{1}{N-1} - \frac{1}{N} + \frac{1}{N-2} - \frac{1}{N-1} + \ldots + \frac{1}{i} - \frac{1}{i+1} + \frac{1}{N}\right) = \frac{c}{i} \end{aligned} \tag{3.1}$$

Therefore $P(i) \propto \frac{1}{i}$, i.e. this process exhibits an exact Zipf's law, the power-law with exponent -1, in the occupation probabilities.
It is possible to generalize the process in order to recall a power-law with a wider

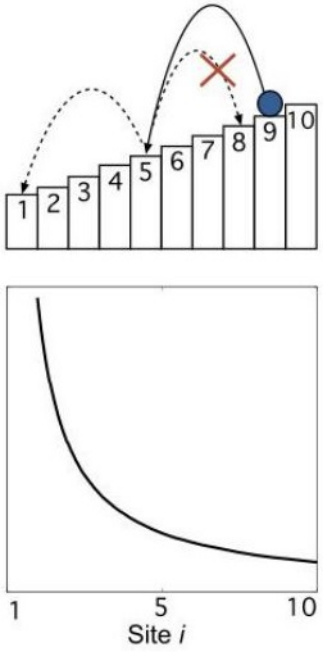


Figure 3.2: Representation of the simple SSR process in the interpretation of a ball that bounces downward on a staircase. Once it reaches the lower lever, it restarts jumping to any step. The process tends to a distribution that is a power law $P(i) = \frac{c}{i}$ where c is the normalization constant.

range of exponents. This generalization is called noisy SSR process and it is obtained perturbing the simple SSR process by noise.

3.1.2 Noisy SSR processes

This model is inspired by the behaviour of many real systems, whose sample space from time to time can expand or reduce during the process. Considering the simple SSR process, we can generalize it allowing upward moves from time to time. Therefore the previous process ϕ is perturbed by noise.
In particular, we consider a superposition of the simple SSR process ϕ and an unconstrained random walk ϕ_R with mixing ratio λ. We define the noisy SSR process $\phi^{(\lambda)}$ as follows:

$$\phi^{(\lambda)} = \lambda\phi + (1-\lambda)\phi_R \tag{3.2}$$

In the analogy of the staircase this process can be seen as a ball that bounces on any step of the staircase with uniform probability, while $\phi^{(\lambda)}$ can be interpreted in terms of a ball that jumps downward on a staircase. Every time the ball hits a step it can move downstairs with probability λ and with probability $(1-\lambda)$ it can jump to any position, therefore upward moves are allowed (see figure 3.3 for a graphic representation). Every time the ball hits the last step it jumps randomly to any step.
Note that being a mixing ratio, $0 \leq \lambda \leq 1$. Moreover, if $\lambda = 0$ the process corresponds to the unconstrained random walk ϕ_R while if $\lambda = 1$ it is the simple SSR process ϕ.
As a consequence of 3.2 we have that the probability of jumping from level i to level j is

$$P(j|i) = \begin{cases} \frac{\lambda}{i-1} + \frac{1-\lambda}{N} & \text{if} \quad j < i \\ \frac{1-\lambda}{N} & \text{if} \quad j \geq i > 1 \\ \frac{1}{N} & \text{if} \quad j \geq i = 1 \end{cases} \tag{3.3}$$

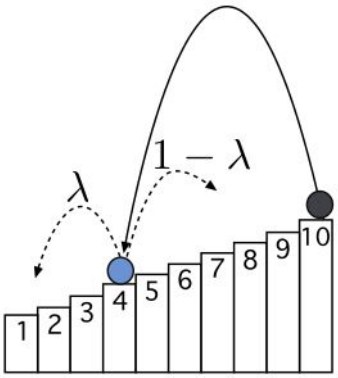


Figure 3.3: Representation of the SSR process with noise. With probability λ the ball can only jump downwards and with probability $(1-\lambda)$ it can jump to any step with uniform probability, therefore upwards moves are allowed.

In the same way as before we can look for the stationary distribution that gives the probability of finding a ball on the i-th step at equilibrium.

$$P(i) = \sum_{j=1}^{N} P(i|j)P(j) = \frac{1-\lambda}{N} + \sum_{j=i+1}^{N} \frac{\lambda}{j-1} P(j) + \frac{1}{N} P(1) \tag{3.4}$$

Therefore the recursive relation

$$P(i+1) - P(i) = -\frac{\lambda}{i} P(i+1) \tag{3.5}$$

holds, from which one obtains

$$\frac{P(i)}{P(1)} = \prod_{j=1}^{i-1} \left(1 + \frac{\lambda}{j}\right)^{-1} = exp\left[-\sum_{j=1}^{i-1} \log\left(1 + \frac{\lambda}{j}\right)\right] \sim exp\left(-\sum_{j=1}^{i-1} \frac{\lambda}{j}\right) \sim \\ \sim exp(-\lambda \log(i)) = i^{-\lambda} \tag{3.6}$$

Since $P(1)$ is given by the normalization condition $\sum_i P(i) = 1$, we find

$$P(i) \propto i^{-\lambda} \tag{3.7}$$

3.1.3 SSR cascades

A third model has been introduced in order to explain power laws with a different range of power-law exponents.
Recalling the model of the ball that jumps on a staircase, to define the SSR cascades process we set a value λ and suppose that at every time every ball is split in λ new balls. In other words, at time $t = 0$, we have λ balls jumping to any state. At time $t = 1$ each of these λ balls divide into λ new balls which all jump to any state below the original state. Whenever a ball hits the bottom step, it is eliminated from the system (see figure 3.4). In this way we are superimposing a multiplicative process that is characterized by the parameter λ. If $\lambda < 1$ our process is the SSR noisy process described in the previous section, and if $\lambda = 1$ no new elements are created, hence the process is the simple SSR process.
Considering one ball at the time we can write the probability of jumping from

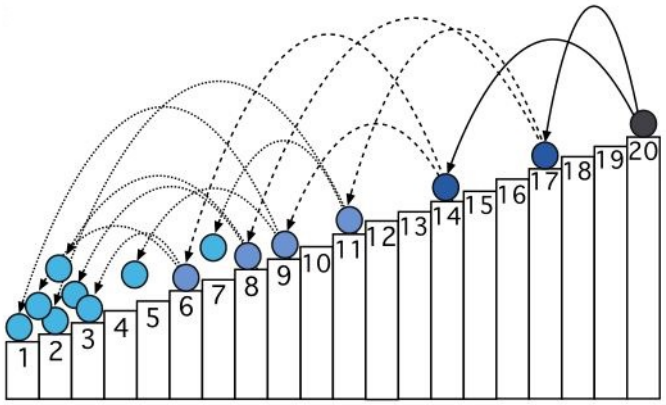


Figure 3.4: Representation of the SSR cascades process with $\lambda > 1$. Whenever a ball hits a state it creates λ balls which continue their random jumps.

step i to j, $P(j|i)$.

$$P(j|i) = \begin{cases} \frac{1}{i-1} & \text{if } j < i \\ 0 & \text{if } j \geq i \end{cases} \tag{3.8}$$

In addition to this, if an element is on the j-th step at time t, there are λ trials to reach any state $i < j$ at $t+1$. Now we take into account the expected number

of jumps from j to i. Since the jumps from j to i of each ball is independent, the expected number of jumps from j to i, $n(j \to i)$ can be approximated as

$$n(j \to i) = \lambda P(i|j) \tag{3.9}$$

We consider the expected number of elements that will hit state i in a given SSR cascade and denote them by $n_i \quad \forall i$. Up to a factor, the sequence $n_1, n_2 \dots n_N$ is identical to the histogram of visits. From equations 3.9 and 3.8 we get:

$$n_i = \sum_{j>i} n(j \to i) n_j = \lambda \sum_{j>i} \frac{n_j}{j-1} \tag{3.10}$$

Therefore, the recursive relation

$$n_i = \left(1 + \frac{\lambda}{i}\right) n_{i+1} \tag{3.11}$$

holds, from which one can find

$$n_i = \prod_{1<j\leq i} \left(1 + \frac{\lambda}{1-j}\right)^{-1} n_1 \tag{3.12}$$

Using the same procedure of equation 3.6, we obtain that

$$\prod_{1<j\leq i} \left(1 + \frac{\lambda}{1-j}\right)^{-1} \sim i^{-\lambda}$$

Therefore we have

$$n_i \sim i^{-\lambda} \tag{3.13}$$

that is, the multiplication factor λ becomes the scaling exponent.
Now we approach the problem of the theoretical formalization of the SSR processes. We will rewrite the first two processes in terms of Markov chains and show that this method does not apply to the SSR cascades process. Eventually we give a different interpretation that is able to recall all the processes and their results.

3.1.4 A unifying model for the SSR process

We first start with a simple observation: the noisy SSR process can be recast in terms of Markov chains.

Definition 3.1. Let $\phi^{(\lambda)} = \{X_t\}_{t\in\mathbb{N}}$ $X_t : \Omega \longmapsto \chi = [1, 2 \dots N]$ be a homogeneous Markov chain, with $P(X_1 = i) = \frac{1}{N} \forall i$ and

$$P(j|i) = \begin{cases} \frac{\lambda}{i-1} + \frac{1-\lambda}{N} & \text{if} \quad j < i \\ \frac{1-\lambda}{N} & \text{if} \quad j \geq i > 1 \\ \frac{1}{N} & \text{if} \quad j \geq i = 1 \end{cases}$$

where

Definition 3.2. Let $\phi = \{X_t\}_{t\in\mathbb{N}}$ $X_t : \Omega \longmapsto \chi = [1, 2 \dots N]$ be a homogeneous Markov chain, with $P(X_1 = i) = \frac{1}{N} \quad \forall i$ and

$$P_{ij} = P(j|i) = \begin{cases} \frac{1}{i-1} & \text{if} \quad j < i \\ 0 & \text{if} \quad j \geq i > 1 \\ \frac{1}{N} & \text{if} \quad j \geq i = 1 \end{cases}$$

and

Definition 3.3. Let $\phi_R = \{X_t\}_{t\in\mathbb{N}}$ $X_t : \Omega \longmapsto \chi = [1, 2 \dots N]$ be a iid process defined by the probability vector $P(X_t = i) = \frac{1}{N} \quad \forall i = 1 \dots N \quad \forall t \in \mathbb{N}$.

We can see $X_t = i$ as the event that at time t the ball is in the level i, therefore considering the interpretation of the staircase this Markov chain is equivalent to the noisy SSR process.
Note that if $\lambda = 1$ $\phi^{(\lambda)} = \phi$ and it is the simple SSR process.
Now we want to compute the vector $\vec{\pi}$ that makes the process stationary. First we prove that the transition probability matrix is regular.

Lemma 3.1.1. *The transition probability matrix P associated with the process $\phi^{(\lambda)}$ is regular, $\forall 0 \leq \lambda \leq 1$.*

Proof. If $0 \leq \lambda < 1$ upward moves are allowed, hence it is always possible to jump from a step of the staircase to any other.If $\lambda = 1$, since every level is linked to the lower step and the lower step is connected to all the levels it is always possible to connect two steps with a path of length at most 2. □

As a consequence, we can use the Perron-Frobenius theorem to find the probability distribution to which the process converges. The vector $\vec{\pi}$ is a probability vector s.t. $\vec{\pi}P = \vec{\pi}$. Therefore we have

$$\vec{\pi}_i = \sum_{j=1}^{N} P(i|j)\vec{\pi}_j \tag{3.14}$$

Using the definition 3.1, the previous becomes

$$\vec{\pi}_i = \frac{1-\lambda}{N} + \sum_{j=i+1}^{N} \frac{\lambda}{j-1}\vec{\pi}_j + \frac{1}{N}\vec{\pi}_1 \tag{3.15}$$

We can compute the recursive relation

$$\vec{\pi}_{i+1} - \vec{\pi}_i = -\frac{\lambda}{i}\vec{\pi}_{i+1} \tag{3.16}$$

This equation is equal to 3.5, therefore we have

$$\vec{\pi}_i \propto i^{-\lambda} \tag{3.17}$$

As a consequence, in the case of the noisy SSR process it is possible to rewrite it in terms of Markov chain and recall the principal results. On the other hand, the same reasoning does not hold for the cascades process. In particular, since the number of balls is not conserved the transition probability matrix P is not a stochastic matrix. To solve this problem, we give a slightly different formalization of the whole class of the SSR models that allows to recast the three models in a single unifying picture.

We introduce a vector $\vec{n}_t$ that represents the number of balls on every step at time t. Every time the balls at level $i > 1$ are redistributed in the lower levels using the same rules defined in the previous models with the only difference that their number (in the continuous limit) is rescaled by a factor λ. Moreover, the balls that hit the lower step are redistributed uniformly on all steps rescaled by a factor c. The proper choice of constant c ensure the conservation of the number of balls. In particular, if $\lambda = 1$ the number of balls is preserved at each time steps, if $\lambda > 1$ it grows and if $0 < \lambda < 1$ it decreases. As a consequence, if $\lambda < 1$ the constant c is bigger than 1 and if $0 < \lambda < 1$ c is smaller than 1. We will show that the system admits a stationary a stationary solution $\vec{n}$ if and only if c and λ are conveniently related.

Definition 3.4. Let $\vec{n}_t = \vec{n}_{t-1}A$ where A is a N $\times$ N matrix $A = (A_{ij})_{i,j=1\dots N}$,

$$A_{ij} = \begin{cases} \frac{\lambda}{j-1} & \text{if } i < j; \\ 0 & \text{if } i \geq j \neq 1; \\ \frac{c}{N} & \text{if } i \geq j = 1. \end{cases} \tag{3.18}$$

where c is a constant and $\lambda \in [0, +\infty[$.

Now we note that this model reproduces the results of the SSR processes presented before.

Theorem 3.1.2. *The matrix A has a left eigenvector with eigenvalue 1 if and only if*

$$c = \prod_{j=1}^{N-1} \left(1 + \frac{\lambda}{j}\right)^{-1} N$$

Proof. Let $\vec{n}$ be the left eigenvector, $\vec{n} = \vec{n}A$. Therefore considering the definition 3.4, the previous equation becomes:

$$\vec{n}_i = \sum_{j=i+1}^{N} \frac{\lambda}{j-1} \vec{n}_j + \frac{c}{N} \vec{n}_1$$

Hence the following recursive relation

$$\vec{n}_i = \left(1 + \frac{\lambda}{i}\right) \vec{n}_{i+1} \tag{3.19}$$

holds, from which one obtains

$$\vec{n}_i = \prod_{j=1}^{i-1} \left(1 + \frac{\lambda}{j}\right)^{-1} \vec{n}_1 \tag{3.20}$$

In the same way, we have

$$\vec{n}_i = \prod_{j=i+1}^{N-1} \left(1 + \frac{\lambda}{j}\right) \vec{n}_N = \prod_{j=i+1}^{N-1} \left(1 + \frac{\lambda}{j}\right)^{-1} \frac{c}{N} \vec{n}_1 \tag{3.21}$$

Where we used $\vec{n}_N = \frac{c}{N}\vec{n}_1$ because $A_{N1} = \frac{c}{N}$. From equations 3.20 and 3.21

$$c = \prod_{j=1}^{N-1} \left(1 + \frac{\lambda}{j}\right)^{-1} N \tag{3.22}$$

follows. □

Theorem 3.1.3. *Let A be the matrix of definition 3.4 and c as in theorem 3.1.2. Let the $\vec{n}$ be the left eigenvector of the matrix A. We have*

$$\vec{n}_i \propto i^{-\lambda}$$

Proof. Recalling the recursive relation 3.20 we obtain

$$\frac{\vec{n}_i}{\vec{n}_1} = \prod_{j=1}^{i-1} \left(1 + \frac{\lambda}{j}\right)^{-1} = exp\left[-\sum_{j=1}^{i-1} \log\left(1 + \frac{\lambda}{j}\right)\right] \sim exp\left(-\sum_{j=1}^{i-1} \frac{\lambda}{j}\right) \sim \\ \sim exp(-\lambda \log(i)) = i^{-\lambda} \tag{3.23}$$

Therefore we have

$$\vec{n}_i \propto i^{-\lambda} \tag{3.24}$$

□

That is, the same results of the SSR processes are recalled.
Now we want to apply the Perron-Frobenius theorem to the matrix A. First we need to prove that the matrix is non-negative, irreducible, aperiodic and its maximum eigenvalue is 1. From the definition 3.4 follows that A is non-negative. Now we prove that also the other properties hold.

Proposition 3.1.4. *Let A be the matrix of definition 3.4 and c as in theorem 3.1.2. The maximum eigenvalue is $\lambda_1 = 1$.*

Proof. From theorem 3.1.3 follows that the eigenvector relative to the eigenvalue $\lambda_1 = 1$ is $\vec{n}$ and $\vec{n} \geq 0$. Therefore λ_1 is the maximum eigenvalue. □

Proposition 3.1.5. *Let A be the matrix of definition 3.4 with c as in theorem 3.1.2. A is irreducible.*

Proof. We consider the adjacency matrix $M = (M_{ij})_{i,j=1\ldots N}$ associated with A and the graph associated with M. Every node of the graph is connected to the node 1 and the node 1 is linked to every node. As a consequence, a path of length 2 that connects two nodes exists always. Therefore $A^2_{ij} \neq 0 \quad \forall i, j = 1 \ldots N$. □

Proposition 3.1.6. *Let A be the matrix of definition 3.4 with c as in theorem 3.1.2. A is aperiodic.*

Proof. From definition 3.4 follows that $A_{11} \neq 0$ and $A_{ij} = 0 \quad \forall i, j \neq 1$. Moreover, from the proposition 3.1.5 follows that $A^2_{ij} \neq 0 \quad \forall i, j$. Since $\mathrm{GCD}\{2, 1\} = 1$ the period of A is 1. □

Therefore we can apply the Perron-Frobenius theorem for irreducible non-negative matrices and find:

Theorem 3.1.7. *Let A be the matrix of definition 3.4 and c as in theorem 3.1.2. Therefore*

$$\lim_{k\to\infty} \vec{m}A^k = c\vec{n} \quad \forall \vec{m} \geq 0$$

Therefore every initial distribution of balls converges to the stationary distribution $\vec{n}$. From the theorem 3.1.3 follows that this distribution is a power-law with exponent λ, $\lambda \in [0, +\infty[$.
Note that if $\lambda > 1$ this model is equivalent to the cascades model, if $0 < \lambda < 1$ it reproduces the results of the noisy model and if $\lambda = 1$ it is the simple SSR model. In this last case the matrix A is equal to the conditional probability matrix P for the simple SSR process given in definition 3.2 if and only if $c = 1$. In fact from theorem 3.1.2, if $\lambda = 1$ we have

$$\frac{c}{N} = \prod_{j=1}^{N-1} \left(1 + \frac{\lambda}{j}\right)^{-1} N = \frac{1}{2} \cdot \frac{2}{3} \cdots \frac{N-2}{N-1} \cdot \frac{N-1}{N} = \frac{1}{N}$$

Therefore this process is a generalization of the SSR models and it is able to recall the same results on the probability distribution of those processes.

3.2 Urn models

A new model that generates Zipf's law is based on the observation of the dynamics of correlated novelties [20]. The correlated novelties model, a generalization of Polya's urns model, was inspired by the process of exploring a space (physical, biological or conceptual) that enlarges whenever a novelty occurs. This model predicts statistical laws for the rate at which novelties happen (Heaps' law) and for the probability distribution on the space explored (Zipf's law).
The correlated novelties model mimics the process by which one novelty sets the stage for another, in the sense that once a novelty is discovered, the space of our possibilities (the adjacent possible) grows. In other words, the adjacent possible consist of all those things that are one step away from what actually exists, and hence can arise from incremental modifications and recombinations of existing material. Whenever something new is created in this way, part of the formerly adjacent possible becomes actualized, and is therefore substituted by a fresh adjacent possible. In this sense, every time a novelty occurs, the adjacent possible

expands.
We define novelties as everything that is new to the subject. They include innovations (something that is created for the first time, hence new for everybody) and discoveries for the subject. In this sense innovations are novelties for everyone. For example, a novelty could be a concept new to the subject. This discovery may induce the subject to search for further information and therefore could lead to the expansion of the personal adjacent possible.

3.2.1 Polya's urn

Since the correlated novelties model is based on a generalization of the Polya's urns, we first introduce the latter. In the classical (and simplest) version of this

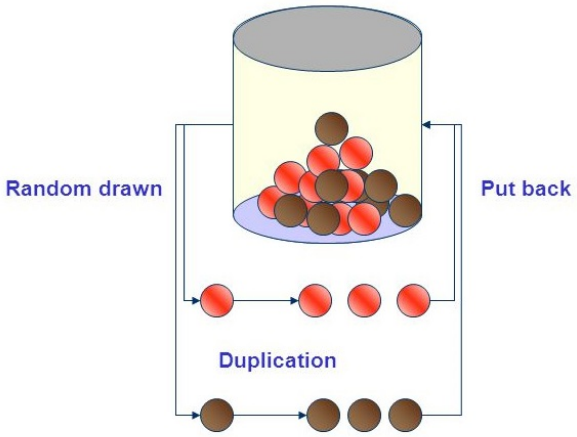


Figure 3.5: Representation of the classical Polya's urn model. Whenever a red ball is withdrawn, the ball is placed back in the urn and a certain number of new balls of the same color are added (in the example, 2 new balls). The same happens if the brown ball is withdrawn.

model, balls of two different colors are placed in an urn. A ball is withdrawn at random and placed back in the urn along with a certain number of new balls of the same color, thereby increasing that color's likelihood of being drawn again in later rounds (see figure 3.5 for a visual representation). In particular, suppose to

have a urn with balls of two different color, say black and red. At time 0 in the urn there are 2 balls, one red and one black. Every time a black ball is drawn from the urn, it is put back in the urn along with a second black ball. In the same way, if a white ball is drawn, it is put back in the urn together with a new white ball. Obviously that the number of balls in the urn at time t is $N_t = t+2$. Note that the number of black balls in the urn at time t depends only on the composition of the urn at time $t-1$. In fact, the probability of drawing a black ball from the urn at time t depends only on the number of balls of that color in the urn at the previous time. Therefore we can define a Markov chain $\mathcal{B}$ that indicates the number of black balls in the urn.

Definition 3.5. Let $\mathcal{B}$ be a Markov chain, $\mathcal{B} = \{B_t\}_{t\in\mathbb{N}}, B_t : \Omega \longmapsto \chi = \mathbb{N}\setminus\{0\}$, with $P(B_1 = 1) = 1$ and

$$P(B_t = k|B_{t-1} = h) = \begin{cases} \frac{h}{t+1} & \text{if } k = h+1; \\ \frac{t-h+1}{t+1} & \text{if } h = k; \\ 0 & \text{if } k \neq h \end{cases} \qquad \forall t \in \mathbb{N}_{>1}$$

where the random variable B_t indicates the number of black balls in the urn at time t. Since the conditional probability depends on the time t the process is not homogeneous.
If we define in the same way the Markov chain $\{R_t\}_{t\in\mathbb{N}}$ associated with the number of red balls in the urn at time t, $P(B_t = k) = P(R_t = t+2-k)$ because at time t there are $t+2$ balls in the urn. Therefore the Markov process B_t is enough to represent the number of both the black and the red balls.
Now we consider a more general version of the Polya's urn model. Suppose to have in the urn N_0 balls at time 0, α red and β black. Thus, $N_0 = \alpha + \beta$. Every time a red ball is drawn from the urn, it is replaced in the urn together with a red balls and b black ones. If the drawn ball is black, it is replaced into the urn

together with c red and d black balls. To represent the model, we can define

$$R = \begin{pmatrix} a & c \\ b & d \end{pmatrix}, \quad U_0 = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

where U_0 is called the initial composition vector and R the replacement matrix. Now suppose that the first ball drawn is red, hence the composition of the urn at time 1 is

$$U_1 = R \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

On the contrary, if the extracted ball is black the composition of the urn at time 1 is

$$U_1 = R \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Therefore the stochastic process $\{U_n\}_{n\in\mathbb{N}}$ describes the composition of the urn. A version of this model can be used to represent the dynamic of novelties and to recall Zipf's and Heaps' laws in this context. We think of the urn as the space of possibilities and the sequence of ball that are withdrawn represents the history that is actually realized. The Polya's urn model can be generalized to allow for novelties to occur and to facilitate the appearance of further novelties. This way we can model phenomena that are characterized by the expansion of their possibility space.

3.2.2 Urn model with triggering

We consider an urn $\mathcal{U}$ containing N_0 distinct elements, represented by balls of different colors. These elements represent human experiences and products of human creativity. The sequence $\mathcal{S}$ of elements generated through successive extractions from the urn represents a series of inventions or experiences. To model the expansion of the adjacent possible when something new occurs the content of the urn itself is assumed to enlarge whenever a new element is withdrawn.

In more detail, at each time step t an element s_t is selected at random from $\mathcal{U}$ and recorded in the sequence. Then the element is put back in the urn along with ρ additional copies of itself. The parameter ρ represent the reinforcement process, i.e. the more likely use of an element in a given context. If the element s_t appears to be novel, $\nu + 1$ brand new distinct elements are put in the urn. These new elements represent the set of new possibilities triggered by the novelty s_t. Hence $\nu + 1$ is the size of the new adjacent possible made available once we have a novel experience.
In other words, at each time step the following procedure is implemented:

1. an item is randomly extracted from $\mathcal{U}$ with uniform probability and added to $\mathcal{S}$;
2. the extracted element is put back into $\mathcal{U}$ together with ρ copies of it;
3. if the extracted element has never been used before in $\mathcal{S}$ (it is a new element), then $\nu + 1$ brand new distinct elements are added to $\mathcal{U}$.

Note that the number of elements of the sequence $\mathcal{S}$, $|\mathcal{S}|$, equals to the number of times t we repeated the above procedure.
We can define the process using the replacement matrix and the initial composition vector. We divide the balls in the urn in two classes: the balls that have already appeared in the sequence $\mathcal{S}$, i.e. the balls that have already been extracted, and the balls that do not appear in $\mathcal{S}$, i.e. the balls that have not been extracted yet. At the beginning all the N_0 balls are new. We have:

$$U_0 = \begin{pmatrix} N_0 \\ 0 \end{pmatrix}, \quad R = \begin{pmatrix} \nu & 0 \\ 1+\rho & \rho \end{pmatrix}.$$

Now we show that this model yields to Zipf's law for the frequency distribution of distinct elements and Heaps' law for the growth of the number of unique elements as a function of the total number of elements. Let D be the number of distinct elements that appear in $\mathcal{S}$, then the total number of elements in $\mathcal{U}$ after t steps is

$$|\mathcal{U}|_t = N_0 + (\nu + 1)D + \rho t \tag{3.25}$$

We also consider a second model in which the reinforcement does not act when an element is chosen for the first time. Hence, point 2. will be changed into:

2a. the extracted element is put back in $\mathcal{U}$ together with ρ copies of it only if it is not new to the sequence.

Therefore the replacement matrix becomes:

$$R = \begin{pmatrix} \nu & 0 \\ 1 & \rho \end{pmatrix}$$

while the initial composition vector U_0 remains the same. In this version of the model, equation 3.25 becomes:

$$|\mathcal{U}|_t = N_0 + (\nu + 1)D + \rho(t - D) = N_0 + (\nu + 1 - \rho)D + \rho t \tag{3.26}$$

We call $U_D(t)$ the number of elements in the urn that at time t have not yet appeared in $\mathcal{S}$, and $U(t) = |\mathcal{U}|_t$ the total number of elements in the urn at time t. Considering the first version of the model and recalling equation 3.25, it is obvious that $U_D(t) = N_0 + \nu D$ where the term νD comes from the fact that each time a new element is introduced in the sequence $U_D(t)$ is increased by ν elements (since $\nu + 1$ brand new elements are added to $\mathcal{U}$, while the chosen element is no longer new). Therefore the time dependence of the number D of different elements in the sequence $\mathcal{S}$ obeys the following differential equation:

$$\frac{dD}{dt} = \frac{U_D(t)}{U(t)} = \frac{N_0 + \nu D}{N_0 + (\nu + 1)D + \rho t} \tag{3.27}$$

In the same way, for the second version of the model:

$$\frac{dD}{dt} = \frac{U_D(t)}{U(t)} = \frac{N_0 + \nu D}{N_0 + (\nu + 1 - \rho)D + \rho t} \tag{3.28}$$

If we put $\alpha \equiv \nu + 1$ for the first model and $\alpha \equiv \nu + 1 - \rho$ for the second we can analyze both versions simultaneously.
Since we are interested in the behaviour of the equation at large times $t \gg N_0$,

we approximate the equations 3.27 and 3.28 by

$$\frac{dD}{dt} = \frac{\nu D}{\alpha D + \rho t} \tag{3.29}$$

And the asymptotic behaviour of $D(t)$ for large t is:

1. If $\nu < \rho$ $D \sim (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}$, i.e. $D(N) \sim N^{\frac{\nu}{\rho}}$;

2. if $\nu > \rho$ $D \sim \frac{\nu-\rho}{a} t$, i.e. $D(N) \sim N$;

3. if $\rho = \nu$ $D \log D \sim \frac{\nu}{\alpha} t \rightarrow D \sim \frac{\nu}{a} \frac{t}{\log t}$, i.e. $D(N) \sim \frac{N}{\log N}$.

This result proves that it is possible to predict Heaps' law from both the models. Moreover, the balance between reinforcement of old elements and triggering of new elements affect this prediction. Now we show that the same holds for Zipf's law.
We call n_i the number of occurrences of an element i in the sequence $\mathcal{S}$. Therefore

$$\frac{dn_i}{dt} = \frac{n_i \rho + 1}{N_0 + \alpha D + \rho t} \tag{3.30}$$

Two cases can be distinguished:

1. If $\nu \leq \rho$, when $lim_{t \rightarrow +\infty} \frac{D}{t} = 0$. Considering only the leading term for $t \rightarrow +\infty$,

   $$\frac{dn_i}{dt} \simeq \frac{n_i}{t}$$

   Let t_i denote the time at which the element i occurred for the first time in the sequence, then the solution $n_i(t)$ starting from the initial condition $n_i(t_i) = 1$ is given by

   $$n_i = \frac{t}{t_i}$$

   Considering the cumulative distribution $P(n_i < n)$, from equation 1 we can write

   $$P(n_i < n) = P(t_i \geq \frac{t}{n}) = 1 - P(t_i < \frac{t}{n})$$

Therefore we can estimate

$$P\left(t_i < \frac{t}{n}\right) \simeq \frac{D\left(\frac{t}{n}\right)}{D(t)} = n^{-\frac{\nu}{\rho}} \tag{3.31}$$

2. If $\nu > \rho$ when $D \simeq \frac{\nu-\rho}{a}t$. Considering $t \gg N_0$,

$$\frac{dn_i}{dt} \simeq \frac{\rho n_i}{\left(\rho + a\frac{\nu-\rho}{a}\right)t} = \frac{pn_i}{\nu t} \tag{3.32}$$

which yields the solution

$$n_i = \left(\frac{t}{t_i}\right)^{\frac{\rho}{\nu}} \tag{3.33}$$

Proceeding as the previous case, we find $P(n_i < n) = P(t_i \geq tn^{-\frac{\nu}{\rho}}) = 1 - P(t_i < tn^{-\frac{\nu}{\rho}})$ and thus

$$P\left(t_i < tn^{-\frac{\nu}{\rho}}\right) \simeq \frac{D\left(tn^{-\frac{\nu}{\rho}}\right)}{D(t)} = n^{-\frac{\nu}{\rho}} \tag{3.34}$$

Obtaining the same result of the previous case.

The probability density function of the occurrences of the elements in the sequence is therefore

$$P(n) = \frac{\partial P(n_i < n)}{\partial n} \sim n^{-\left(1+\frac{\nu}{\rho}\right)} \tag{3.35}$$

which correspond to a frequency-rank distribution

$$f(R) \sim R^{-\frac{\rho}{\nu}}$$

A model with semantics

Now we show that with an easy modification, the models just explained could be correlated to text formation, therefor they could explain the correlation between texts and Zipf's and Heaps' laws. We simply endow each element with a label, representing its semantic group, and we allow for the emergence of dynamical correlations between semantically related elements. We consider the same

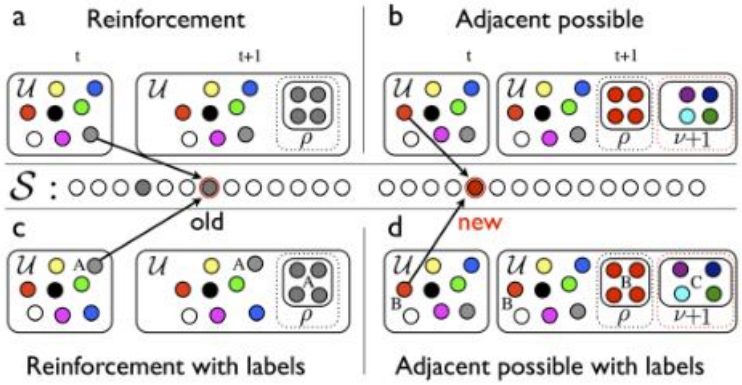


Figure 3.6: Representation of the simple urn model with triggering (a) and (b), and of the urn model with semantic triggering (c), (d). The first half represents the case in which an element (the gray ball) that had been previously drawn is drawn again and the second half the case in which the ball drawn is new.

urn $\mathcal{U}$ with N_0 elements, but we divide those elements into $\frac{N_0}{\nu+1}$ groups. The element in the same group share a common label. To construct the sequence $\mathcal{S}$ the first element is randomly chosen from the urn. Then at each step t (see figure 3.6):

1. we give a weight 1 to:

   (a) each element in $\mathcal{U}$ with the same label (say A) as s_{t-1};

   (b) the element that triggered the entry into the urn of the elements with label A;

   (c) the elements triggered by s_{t-1};

   A weight $\eta \leq 1$ is assigned to all the other elements in $\mathcal{U}$.

2. we choose an element s_t from $\mathcal{U}$ with a probability proportional to its weight and write it in the sequence;

3. we put the element s_t back in $\mathcal{U}$ along with ρ additional copies of it;

4. if and only if the chosen element s_t is new we put $\nu+1$ brand new distinct elements in $\mathcal{U}$, all with a common brand new label. These $\nu+1$ new

elements are given a weight $\eta = 1$ at the next time step $t+1$ and each time the same mother element s_t is picked.

Note that with $\eta = 1$ this model reduces to the simple urn model with triggering introduced earlier.
This extended model can again reproduce both Heaps' and Zipf's laws. We start from the estimate of the Heap's law exponent as function of the parameters ν, ρ and η. Supposing that the last element added to $\mathcal{S}$ has label s, therefore the probability of drawing a new ball is equal to the probability of drawing a new ball with label s plus the probability of drawing a new ball with a different label. Formally,

$$\begin{aligned} &P(new) = P(new, label = s) + P(new, label \neq s) = \\ &= P(new|label = s)P(label = s) + P(new|label \neq s)P(label \neq s) \end{aligned} \tag{3.36}$$

therefore, if we call $N^s(t)$ the number of elements with label s, $N_D^s(t)$ the number of new (never used in the sequence $\mathcal{S}$) elements with label s, $N^{\bar{s}}(t)$ the number of elements with label different from s and $N_D^{\bar{s}}(t)$ the number of new elements with label different from s that are present in $\mathcal{U}$ at time t, we have:

$$\frac{dD(t)}{dt} = \frac{N^s(t)}{N^s(t) + \eta N^{\bar{s}}(t)} \frac{N_D^s(t)}{N^s(t)} + \frac{\eta N^{\bar{s}}(t)}{N^s(t) + \eta N^{\bar{s}}(t)} \frac{N_D^{\bar{s}}(t)}{N^{\bar{s}}(t)} = \frac{N_D^s(t) + \eta N_D^{\bar{s}}(t)}{N^s(t) + \eta N^{\bar{s}}(t)} \tag{3.37}$$

The following relations hold:

$$\nu D(t) = N_D^s(t) + \eta N_D^{\bar{s}}(t) \tag{3.38}$$

and, calling $U(t)$ the number of total elements in the urn,

$$U(t) = N^s(t) + N^{\bar{s}}(t) \tag{3.39}$$

Note that if $\eta = 1$ one recovers equation 3.27.
On the contrary, if we do not know the label of the last added element, we can

write the general equation for $D(t)$:

$$\frac{dD(t)}{dt} = \sum_k P(k)\frac{\nu D_k(t)}{U_k(t)} = \sum_k P(k)\frac{N_D^k(t) + \eta N_D^{\bar{k}}(t)}{N^k(t) + \eta N^{\bar{k}}(t)} = \\ = \sum_k P(k)\frac{N_D^k(t) + \eta(\nu D(t) - N_D^k(t))}{N^k(t) + \eta(U(t) - N^k(t))} \tag{3.40}$$

where the sum is over all the labels k present at time t in $\mathcal{U}$, $P(k)$ is the probability that the last added element to the sequence $\mathcal{S}$ at time t had the label k. Now we have to estimate $N^k(t)$ and $N_D^k(t)$. Note that $N_D^k(t) \leq \nu + 1$, therefore this term can be neglected for large t with respect to $D(t)$. The problem of calculating $N^k(t)$ is complex, and to avoid it we calculate the probability $P(n)$ that $N^k(t) \equiv n$. We can rewrite the equation 3.40 substituting the sum over k with the sum over the labels with the same number of occurrences n in the urn. Therefore, asymptotically the following equation holds:

$$\frac{dD(t)}{dt} = \sum_n P(n)\frac{\eta\nu D(t)}{n(1-\eta) + \eta U(t)} \tag{3.41}$$

Now we consider two cases:

1. We keep in the sum only the terms $n \simeq U(t)$. This approximation is sufficiently good when the frequency-rank distribution for the elements in $\mathcal{S}$ is sufficiently steep, corresponding to a high Zipf's exponent. Solving the previous equation within this approximation, we obtain the result for Heaps' exponent $\beta = min\left(\frac{\eta\nu}{\rho}, 1\right)$.

2. When the probability $P(n)$ is large only for $n \ll U(t)$, we can neglect the term $n(1-\eta)$ with respect to $\eta U(t)$. Solving the equation within this approximation we obtain $\beta \simeq min\left(\frac{\nu}{\rho}, 1\right)$.

Summarizing, we have obtained lower and upper bounds for β, $min\left(\frac{\eta\nu}{\rho}, 1\right) \leq \beta \leq min\left(\frac{\nu}{\rho}, 1\right)$. Thus, the hypothesized mechanism of a relentlessly expanding of the adjacent possible is consistent with the dynamics of correlated novelties.

Chapter 4

Network of words

In the previous chapters we introduced and studied some models that recreate the features of Zipf's law and try to explain it. Most of them are built starting from assumptions on the dynamics of the phenomena they want to explain. A different approach consists in studying the structure of written texts in order to deepen the understanding of natural language and give an explanation of the appearing of Zipf's law in this context.
From this perspective, we focused on the study of the English language creating a network of words on the basis of the structure of some texts, then we studied its features trying to link them to Zipf's law and in particular to its changing slope. The main idea was to convey the structure of texts in the features of a network in order to study it with respect to Zipf's slope change. Since we are interested in getting information about the structure of the network, we applied a topological study of its features.
Similar networks of words have already been used to study the feature of texts. In particular, they have been used in the context of authorship attribution ([2],[1]), while others underlined some of the properties of such graphs relating their absence to language disorders ([8]). Our aim is to connect the feature of the network to the change in slope of Zipf's law. More in particular, we think that the network would have the following structure: a central part constituted by the most frequent words, i.e. the words that populate the first slope of Zipf's law and a

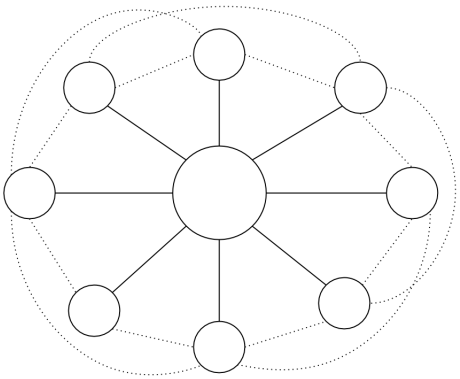
peripheral area inhabited by rare words, the ones that correspond to the second slope. The most frequent words are connecting words (such as *and*, *but*, *the* ...) and words that are related to common topics, while the rare words deal with the specific topics. In our idea, the outer part of the graph is constituted by areas where words with the same topic are aggregated. Every topic area is connected to the central part and some of them can be connected one to another with rare links. Therefore we could depict it with a daisy structure (see Fig. 4.1(a)).

4.1 The dataset

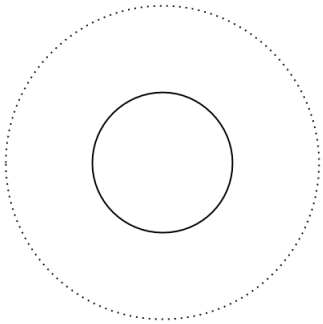
In order to study Zipf's law and its slope in relation with language, we used a dataset of texts written in English collected from the Gutenberg Project ebook collection (see table 4.1). We studied 13 different texts chosen on the basis of the length. The length of the texts is important because it is necessary to be able to detect the changing slope of Zipf's law, which is not appreciable when the corpus is larger than 10^4 words. To have a general idea of language, we chose a various dataset: the texts deal with different subjects and include both prose and poetry.

We decided to apply a stemmatizer to the corpus of texts in order to aggregate words and to make it possible to see Zipf's changing slope at lower ranks. The stemming process is a process that reduces inflected (or derived) words to their word stem, base or root form. In simple words, it deletes the *-s* from plurals of words and third person of verbs, the ending *-ly* from adjectives, the ending *-ing* from verbs and so on. For example, the stemmer aggregates the words *tree* and *trees*, the verbs *see* and *sees*, and the adjective *slow* and the adverb *slowly*. Considering the first lines of the book *Pelle the conqueror* by M. A. Nexø, the unstemmed (plain) text is: *'It was dawn on the first of May, 1877. From the sea the mist came sweeping in, in a gray trail that lay heavily on the water'*, while the stemmed counterpart is: *'It wa dawn on the first of Mai, 1877. From the sea the mist came sweep in, in a grai trail that lai heavili on the water'*.

This process is useful because it reduces the number of words in the sample with-



(a) Hypothesized configuration of the network of words co-occurrences: the daisy structure. It is composed by a central part populated by frequent words and leaves for the rare words, aggregated on the basis of the topic they deal with. Every pair of links can communicate with rare links (dashed in the picture).



(b) Structure of the network after a shuffle of the texts. The inner and outer parts are still recognizable because of the density of their links but the most rare words are not aggregated on the basis of their topic.

Figure 4.1: Graphical representation of the network on the sample (a) and on the shuffled sample (b). The distinction in inner and outer part holds also for the shuffled graph but not the aggregation of the rare words on the basis of their topic.

Table 4.1: The corpus sorted by number of total words of the texts. It is composed by 13 non-copyrighted ebooks available at the Gutenberg Project ebook collection. All the books are in English and three of them are translations of non-English books. The total number of words is $\sim 9 \times 10^7$, while the number of different words is $\sim 7.6 \times 10^4$. We applied a stemming process to the texts in order to reduce the total number of different words, reaching a value of $\sim 5 \times 10^5$ different words.

Author	**Work**	**Total number of words**	**Number of distinct words**	
			stemmed	**unstemmed**
G. Parker	Complete Works of Gilbert Parker	2,235,926	21,113	33,845
W. D. Howells	The Entire PG Edition of W. D. Howells	1,488,669	17,668	29,714
	The world English Bible	837,943	8,825	12,826
J. Bunyan	The works of John Bunyan volume 2	761,335	9,571	15,356
J. Bunyan	The works of John Bunyan volume 3	617,467	9,756	15,773
J. G. Whittier	The Complete Works of Whittier	590,268	16,224	25,792
L. Tolstoj (translated)	War and peace	572,550	10,944	17,554
M. A. Nexø (translated)	Pelle the conqueror	437,809	9,182	14,789
C. A. Fyffe	History of Modern Europe 1792-1878	434,904	10,721	16,100
T. Moore	The Complete Poems of Sir Thomas Moore	380,104	15,021	22,162
P. Shelley	The complete poetical works of Percy Bysshe Shelley	376,675	13,981	20,977
W. M. Thackeray	The Newcomes	372,986	11,775	17,921
L. Mühlbach (translated)	Joseph II and his court	370,230	9,619	15,094
Total		**9,476,866**	**49,987**	**76,270**

out losing meaning or information on the structure of the texts. In Tab. 4.1 it is possible to compare the number of different words of every text, considering both the stemmed and unstemmed versions. On average, the stemmed texts have 40% less different words than the original ones. Obviously after the stemming process the total number of words remains the same.
Another way of aggregating words is to use a lemmatizer. This process is more refined than the stemmatizer because it aggregates words considering their real roots. The correct root of a word is tied to the context, for example the word *saw* could have two meanings: it could be the past tense of the verb to see or the noun for a working tool and its meaning can be inferred only from the topic of the section of text it is used in. Since it is not a mechanical removal of the endings of words basing on specified rules, the lemmatizer is more efficient than the stemmatizier in aggregating words according to their meaning. For example, considering the first two sentences of *Pelle the conqueror*, a lemmatizer would have recognized the word *was* as the past tense of the verb to be, while the stemmatizer only deleted the last 's'. On the other hand, being less elaborate a stemmatizer is faster. Because of all the advantages they lead to, stemmatizers and lemmatizers are frequently used in text analysis ([2],[9]). Since ours is a quantitative study, we do not need an accurate lemmatization of the texts, therefore we decided to use a stemmatizer.
A stemmatizer does not only reduce the number of words without losing information on the structure of the text. Another advantage lies in the fact that on a stemmed sample the changing slope of Zipf's law is visible with a smaller number of words. In fact, since the words are aggregated on the basis of their stems the corpus still conserves all its features but counts less different words, which are the ones used to build the frequency rank of the sample.

4.1.1 The Zifp's law on the sample

Since we aim to study the characteristics of the texts in relation with the slope of Zipf's law, we need to know if the sample presents a changing slope in

its frequency-rank distribution. It is also important to know where the change occurs in terms of ranking. Fig. 4.2 shows the frequency-rank plot on both the stemmed and unstemmed corpus. The change in slope is visible in both cases, but it appears earlier (rank-wise) on the stemmed corpus. In fact, the change happens around 2×10^3 and the angle between the two power-laws is sharper than the angle on the unstemmed texts.

From now on we will only consider the stemmed corpus.

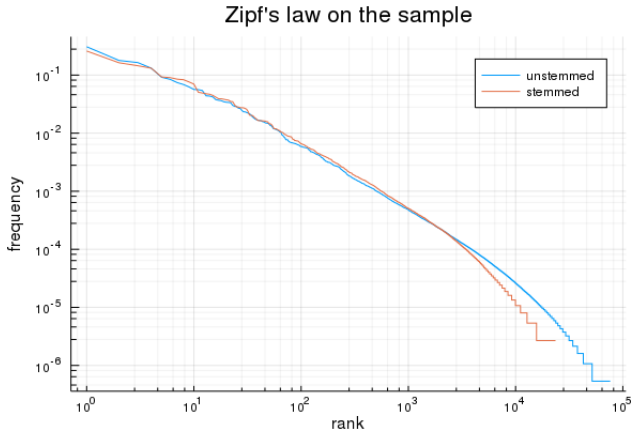


Figure 4.2: Frequency-rank (Zipf's law) log-log plot on the stemmed and unstemmed corpus. Considering the stemmed sample, the slope changes earlier and the angle is sharper for the stemmed texts than for the unstemmed ones.

Ascertained that the changing slope is visible in our sample, we can start analyzing the relation between the corpus and Zipf's double slope. For studying the features of written texts we decided to create a structure that could be representative of it and that could be studied in a rigorous way. In order to make this possible, we used the words of the sentences and their order to define a network. Note that Zipf's law does not depend on the order of the words in the texts but only on their frequency. As a consequence, Fig. 4.2 would be observed also for the shuffled sample. On the other hand, the network we are going to build is based on the order of words, therefore a shuffle of the text would change completely the topology of the graph. Our idea is that with a shuffling the daisy structure would be lost, while the distinction between a central part and an outer part would be conserved. In other words, the outer part would not be divided in

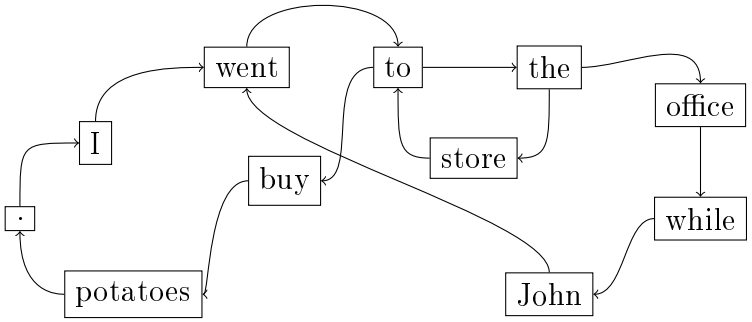
different areas on the basis of the topic but it would still be recognizably different from the central part (see Fig.4.1(b) for a graphical representation).
As we said, after a shuffle of the text Zipf's law will not change but the topology of the network will be completely different. However, we decided to study the network because we think it is related to Zipf's law on a deeper level. In fact, the only thing that distinguishes the real text from its shuffles is the dynamic of the building of sentences, that is the order according to which sentences are created. We think that Zipf's law is related to the human limitations: if humans were omniscient, there is reason to think that they would not distinguish a set of common words and a set of rare ones but use all of them indistinctly. Under this assumption, the changing slope of Zipf's law would probably not appear. Therefore the evidence of a power-law distribution in the frequency rank of texts leads us to think that the creation of texts, and more in general the human language, has some constraints that are related to our mind. In fact, since everyone is experienced in some field of the human knowledge, every book and human speech deal only with few of the possible topics. As a consequence, every meaningful text populates only some petals of the daisy structure, while every of them presents the central component. Therefore, considering different texts as we did in our analysis, the structure of the network will present several petals that are strongly connected to the central part of the graph and weakly connected one with the other.
To sum up, we study the topology of the graph because we think it is related to Zipf's law on a deeper level: Zipf's law does not depend on the order of the words, but it is related to the human capacities and mind. Also the network's topological structure depends on the mechanisms of sentence construction and it is reasonable to think that it mirrors all the constraints imposed by the human nature to the language. Therefore studying the network we will gather information also on the Zipf's law.

4.2 Construction of the network

The network had to conserve the main features of the texts considered, therefore we decided to construct it according to the sentences structure.
Supposing to have a written text, we first deleted the punctuation and symbols such as commas, brackets, and so on. We decided not to delete numbers, full stops, question and exclamation marks. We left the numbers because they have the same role as words in sentences. We did not remove the exclamation marks, question marks and full stops because having a stopping function they are important to the structure of the network. Since they have the same role, we considered question and interrogative marks the same as full stops.
After cleaning the sample, we defined a network where every vertex corresponds to a different word of the corpus and the links connect one word with the following word in the sentence. Since the full stops are at the beginning and ending of every sentence, every first and last word of a sentence is connected to the full stop (we artificially added a full stop at the beginning of every text). Therefore we can consider the full stop the central vertex of the network and we call it ρ.
More formally,

Definition 4.1. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network, where $\mathcal{V}$ is the set of all the words that appear at least once in the sample and $\mathcal{E} = \{(i,j), i,j \in \mathcal{V}, i \text{ precedes } j \text{ in the text}\}$.

Note that from this definition follows that the graph we are defining is directed. In fact, since usually in sentences with meaning the order is important, we could have $(i,j) \in \mathcal{E}$ but $(j,i) \notin \mathcal{E}$.
Notation-wise, we call $i \mapsto j$ the link (i,j). We will write $i_1 \to i_n = i_1 \mapsto i_2 + i_2 \mapsto i_3 + \ldots + i_{n-1} \mapsto i_n$ for the concatenation of the links $(i_1,i_2),(i_2,i_3)\ldots(i_{n-1},i_n)$.
We will call $i_1 \to i_n$ the path that connects i_1 with i_n.
For example, considering the simple sentence 'I went to the office while John went to the store to buy potatoes' we can construct the network $\mathcal{G}$:



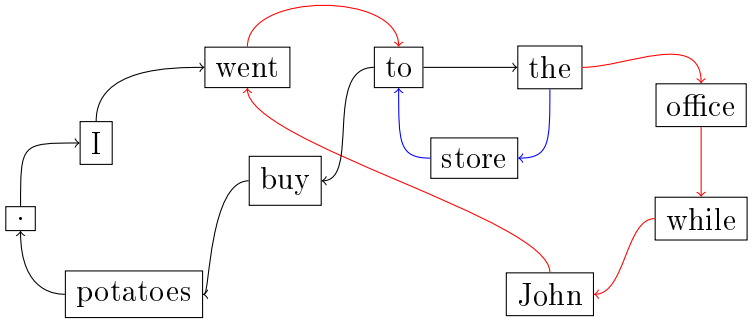
Where every vertex is labeled with the word it is associated with.
Note that some words occurred more than once but this evidence is not recorded in the graph. In our first approach to the study of the network we will only take into account the different words that are used in the text but not their number of occurrences.

4.3 Analysis of the network

Having defined the network, the next step is to study it. We started analyzing some simple features, as the number of nodes of the network. We wanted to understand the structure and the shape of the graph. More in detail, we wanted to check if the topological structure of the network could be related to Zipf's law. In order to do this, we defined some distances and analyzed the graph in relation to them.

4.3.1 Topological distance

Initially, we considered the plain topological distance, i.e. the distance between the nodes u and v is the minimum number of links that connects them. For instance, considering the previous example, the distance between the nodes *the* and *to* is 2. There is more than one path that connects the two words, for example the blue and the red ones, but the blue is the shortest and it is long 2.



Note that this definition of distance takes into account the direction of the links. In fact, if we were not considering it, the shortest path would be the link $to \mapsto the$ that is long 1.
In more formal words,

Definition 4.2. Let $i_1, i_n \in \mathcal{V}$, $i_1 \to i_n$ the oriented path that connects the vertices i_1 with i_n. Let $i_1 \to i_n = i_1 \mapsto i_2 + i_2 \mapsto i_3 + \ldots + i_{n-1} \mapsto i_n$. We define the length of the path as

$$\ell(i_1 \to i_n) = |\{i_1, i_2 \ldots i_n\}|$$

Remark 19. The graph is constituted by one strongly connected component (SCC).

Proof. We have to prove that if $u, v \in \mathcal{V}$, a path $u \to v$ in $\mathcal{G}$ exists always. Due to the structure of the network $\mathcal{G}$, a path that connects u with the full stop ρ, $u \to \rho$ exists $\forall u \in \mathcal{V}$. For the same reason, a path $\rho \to v$ exists $\forall v \in \mathcal{V}$. Since $u \to v = u \to \rho + \rho \to v$ we have the thesis. □

Definition 4.3 (Topological distance)**.** Let $u, v \in \mathcal{V}$

$$d(u, v) = \min_{u \to v} \ell(u \to v)$$

Note that mathematically speaking this is not a distance. In fact, it is always larger or equal to 0 and $d(u, v) = 0$ if and only if $u = v$. Moreover it satisfies the

triangle inequality but, since the graph is directed, the symmetry property does not hold.

Remark 20. $d(u, v) \in \mathbb{N} \quad \forall u, v \in \mathcal{V}$.

Proof. As a consequence of definition 4.2, $\ell(u \to v) \in \mathbb{N} \quad \forall u, v \in \mathcal{V}$, therefore also $d(u, v) \in \mathbb{N} \quad \forall u, v \in \mathcal{V}$. □

As advanced, the stopping function of the full stop is a peculiar role in the structure of the sentences. All the other words of the sample can occupy different positions in sentences while the full stop is always found at the beginning and at the end of them. Hence, the full stop has a remarkable peculiarity in respect to the positioning of words in the sentence. Since the network had to represent all the features of the sample, we took into account this evidence constructing the graph in such a way that the node ρ associated with the full stop can be considered the center of the network. For these reasons, we are interested in calculating the distance of every vertex from ρ. Moreover, because of the triangle inequality the distance of every word from the full stop can lead to an upper bound for the distance between every pair of words. Therefore we compute $d(\rho, u) \quad \forall u \in \mathcal{V}, u \neq \rho$.

The distance defined above takes into account only the minimum length of the paths, therefore it leads to an aggregating phenomenon. For example, a word that once appeared at the beginning of a sentence has distance 1, no matter of the position of its other occurrences. This feature affects the maximum distance and the population of the distances according to the rank of the vertices. In fact, from Fig. 4.3 it is possible to see that due to the definition of the network the maximum distance is quite low. While the maximum length of a sentence in the sample is about 1500 words, it is always possible to connect one word to the full stop with a path shorter or equal to 8.

Moreover, due to the structure of the network there are many links that enter and exit from a frequent word. This is because a frequent word is usually used in different sentences, hence it is followed by different words. This evidence com-

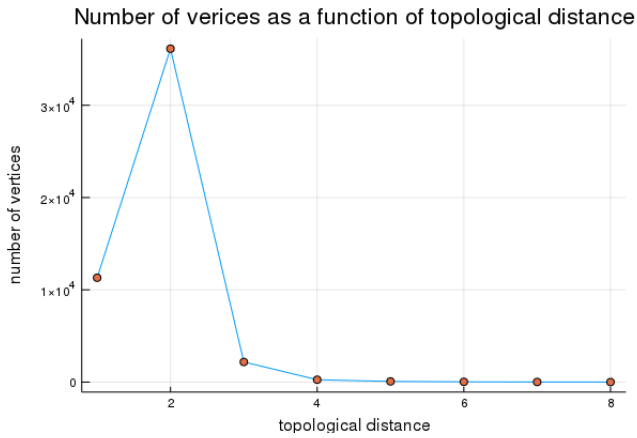


Figure 4.3: Number of nodes as a function of the topological distance. Due to the structure of the network and the definition of this distance, the maximum distance is much smaller than the maximum length of sentences and most of the words have a distance lower or equal to 2.

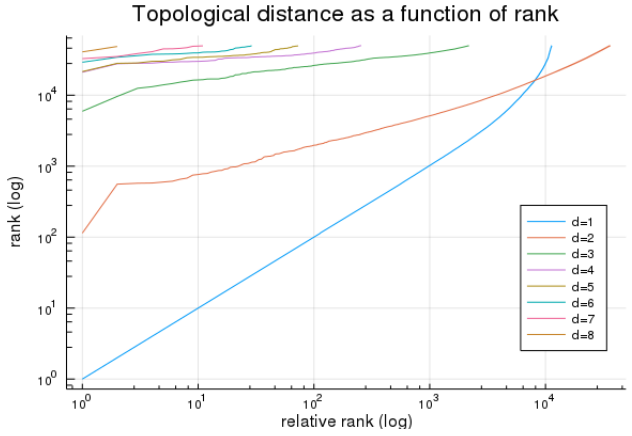


Figure 4.4: Log-log plot of the rank of the words divided on the basis of their topological distance. The more the distance grows, the more the words are rare. The y-axis is the absolute frequency rank of the words while the x-axis is the relative rank, that is the frequency rank of the words with respect to the distance they belong to.

bined with the aggregation feature brings to the fact that a frequent word has more chances to be at a low distance. In fact some sentence could start with it or, if it is never used at the beginning of a sentence, at least one of its links connects it to a word with low distance. This eventuality has an high likelihood for frequent words because they are connected to many words. Hence, at low distances we expect to find most words and the most frequent.
Moreover, frequent words trigger a multiplying effect. In fact, since a frequent word is usually at low distance, a word connected to it most likely has a low distance. Having many links, the frequent words let many other words have a low distance. For instance, it is highly probable to find in the sample a sentence that starts with the word *the*, therefore the vertex associated with that word has distance 1. As a consequence, all the words that follow *the* in every sentence of the sample have distance at most 2.
In Fig. 4.3 it is possible to see that the higher number of vertices is at distance 2, therefore the multiplying effect is strong only at the beginning, while the aggregating component takes over from distance 3 on.
As shown in Fig. 4.4, the most frequent words have distance less than to 3 while the rare words populate the other distances. In fact, distance 3 is populated by words that have rank higher than 3×10^4 while distance 1 and 2 are rank-wise uniformly populated, i.e. they host frequent words as well as rare words.
As we advanced, this definition of distance does not take into account the occurrences of every link, hence the frequency of the word is not fully represented.
To make the analysis more accurate, we gave a second definition of distance that also considers the number of occurrences of links.

4.3.2 Weighted distance

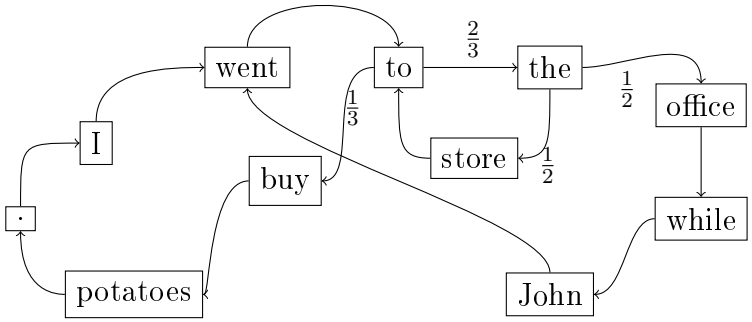
This second definition deals with probabilities. We consider the weight of every link as the estimated probability of crossing that particular link. In other words, if a link connects the word a with the word b, its weight is the probability of choosing the word b starting from a. More formally,

Definition 4.4. We define the weight of the link $i \mapsto j$ as:

$$W_{i,j} = \frac{w_{i,j}}{\sum_{\{k|(i,k)\in\mathcal{E}\}} w_{i,k}}$$

where $w_{i,k}$ is the number of occurrences of the link $i \mapsto k$.

Therefore the weight of the link $u \mapsto v$ corresponds to the measured transition probability of going from u to v.
Considering the previous example, the weighted graph would be:



Where we only indicated the links with weight $\neq 1$.
Now define a new distance on the weighted graph:

Definition 4.5. Let $i_1, i_n \in \mathcal{E}$, $i_1 \to i_n$ the oriented weighted path that connects the vertices i_1 with i_n. Let $i_1 \to i_n = i_1 \mapsto i_2 + i_2 \mapsto i_3 + \ldots + i_{n-1} \mapsto i_n$. We define

$$\ell_W(i_1 \to i_n) = \sum_{j=1}^{n-1} W_{i_j, i_{j+1}}$$

that is, the length of a path is the sum of the weight of its intermediate links.

Definition 4.6 (Weighted distance)**.** Let $u, v \in \mathcal{E}$

$$d_W(u, v) = \min_{u \to v} \ell_W(u \to v)$$

As we did with the topological distance, we consider $d_W(\rho, u) \quad \forall u \in \mathcal{V}, u \neq \rho$ and therefore we calculate the distance of every word from the center of the network.

Note that $d(u,v) \geq 0 \forall u, v \in \mathcal{V}$ and $d(u,v) = 0 \Leftrightarrow u = v$, but the symmetry and

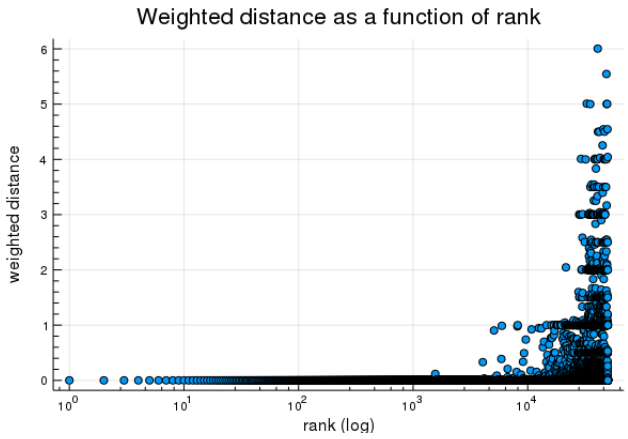


Figure 4.5: Weighted distance as a function of the rank of the words (log-lin scale). Words that populate the first part of Zipf's law have weighted distance ~ 0, while words that populated the second part have on average larger distance.

triangle inequality do not hold for this distance, therefore this is not a mathematical distance. Moreover, since every link has weight ≤ 1, the simple topological distance is an upper bound for the weighted distance. In fact, as shown in Fig. 4.5, the maximum of the weighted distance is ~ 6.

Remark 21. $d(u,v) \in \mathbb{Q}^+ \quad \forall u, v \in \mathcal{V}$.

Proof. As a consequence of definition 4.5, $\ell_W(u \to v) \in \mathbb{Q}^+ \quad \forall u, v \in \mathcal{V}$. Since $d(u,v)$ is a sum of rational numbers, $d(u,v) \in \mathbb{Q}^+ \quad \forall u, v \in \mathcal{V}$. □

Including the notion of frequency in the definition of distance helps us to find the correlation between the structure of the network and Zipf's double slope. In fact, Fig. 4.5 shows that the distance takes values larger than 0 only for words whose rank is larger than 2×10^3. That is, the words that populate the second part of the frequency-rank plot have on average distance larger than 0, while the distance of all the other words is ~ 0. Therefore the weighted distance could be seen as a way to distinguish rare and frequent words without counting their occurrences.

Since we defined the weight of a link as the estimated transition probability of going from the first vertex to the second, we can apply the definition of entropy to every word of the graph.

Remark 22. The entropy associated to the vertex $v \in \mathcal{V}$ is

$$E(v) = -\sum_{\{u|(v,u)\in\mathcal{E}\}} W_{v,u} \log(W_{v,u})$$

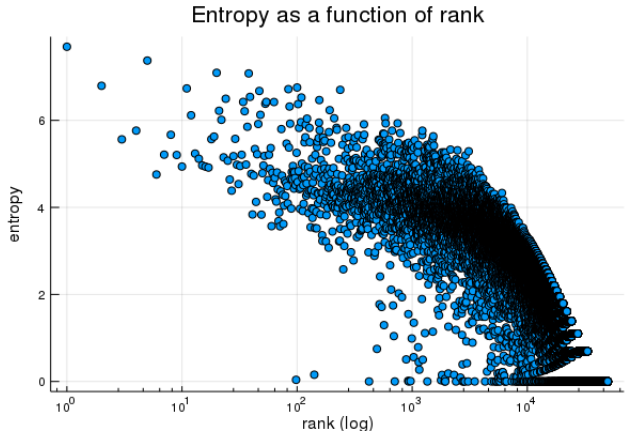
In Fig. 4.6(a) it is possible to see that the entropy just defined is higher for low-ranked words. That is because, as already discussed, usually frequent words have an higher number of exiting links. On the contrary, if a word is rare it will have few outneightbors and the entropy associated to it will be low. For example, if one word has only one outneighbour, the link will have probability 1 and the entropy associated to that vertex will be 0.
Since, as already discussed, the value of the entropy increases with the number of outneightbours of the vertex considered, we consider the normalized entropy, that is the entropy divided by the number of exiting links.

Definition 4.7 (Normalized entropy)**.** The normalized entropy associated with the vertex $v \in \mathcal{V}$ is

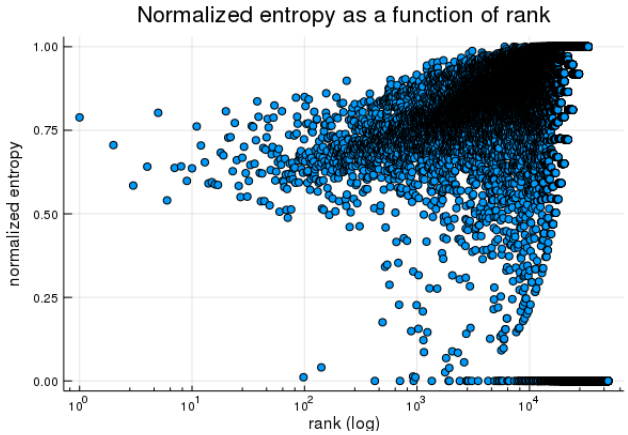
$$E(v) = -\frac{\sum_{\{u|(v,u)\in\mathcal{E}\}} W_{v,u} \log(W_{v,u})}{\log k_v}$$

where k_v is the number of outneighbours of the vertex v.

In Fig. 4.6(b) it is possible to see that the normalized entropy grows with the rank. Its behaviour is opposite to the entropy one, which means that the number of exiting links was a prevalent factor in the computation of the entropy. More in detail, the positive correlation between normalized entropy and rank seems to be stronger on words with rank larger than 10^3. In fact, words with rank larger than 10^3 have a wider range of entropy levels, while more frequent words seem to be more stable according to the entropy range. Note that some words have entropy 0 because they have only one exiting link.
To sum up, since the weighted distance takes into account the occurrences of the



(a) Entropy as a function of the rank (log-lin scale). Words with higher rank have a lower entropy, therefore they are negatively correlated.



(b) Normalized entropy as a function of the rank (log-lin scale). The entropy raises at the growing of the rank and this positive correlation seems to be stronger on words that have rank larger than 10^3.

Figure 4.6: Entropy and normalized entropy as functions of the rank (log-lin scale). The difference in behaviour is due to the fact that entropy takes into account the number of exiting links. If we normalize it by dividing by the logarithm of the number of exiting links, we find a value of entropy that does not depend on the number of the links.

links it seems to be more useful than the plain topological distance to study the relation between the structure of texts, i.e. the feature of the network, and the double slope of Zipf's law. Obviously this is because both the weighted distance and Zipf's law deal with the frequency of the words. Once we took into account the frequency of the words by means of their occurrences we have been able to detect the features of the network that are related to the rank in frequency of those words.
On the other hand, the aggregating factor seems to hide some of the properties of the texts. In fact, as already discussed, if once a word appears at the beginning of a sentence it has low distance, no matter the other occurrences. To help solving this problem, we defined a distance that takes into account both the occurrences of the words and their position in the sentences.

4.3.3 Mean distance

We defined a third distance as the average position occupied by a word in the sentence. This way we take into account the repetitions of the word and the position that the word has in the sentence with respect to the starting full stop.

Definition 4.8. Let $v \in \mathcal{V}$ a vertex associated with a word that occurs n times in the sample. Every occurrence is linked with a number d_i of the words that stand in between v and the preceding full stop. We define the distance

$$d_M(v) = \frac{\sum_{i=1}^{n} d_i}{n}$$

Note that this definition takes into account both the number of occurrences of the word and their distance from the center of the network. On the contrary, for the other two definitions of distance this was not important: they only took into account the minimum of the distance, that corresponds to only one occurrence.
To analyze the relation between the mean distance and Zipf's law we plotted the distance as a function of the rank (Fig. 4.7). The Fig. shows that high distances are related to high positions in the frequency rank and vice versa, but it does not seem to be related to the changing in slope of Zipf's law. In fact, the distance

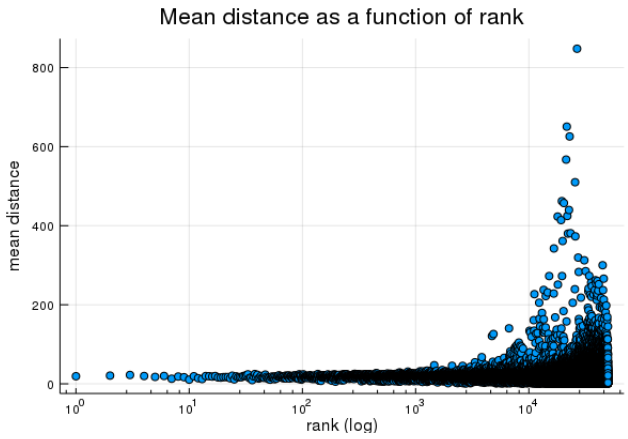


Figure 4.7: Mean distance as a function of the rank (log-lin scale). The growing of the distance with the rank is probably due to the unreliability of the mean of the distance on rare words.

seems to grow with the same rate before and after the point where Zipf's slope changes. On the other hand, for words with rank larger than 10^4 the distance seems to grow faster. However, this phenomenon is probably due to the low number of occurrences of the words. In fact, since those words occur only once or twice in the whole text, they do not offer a reliable field for analysis. This evidence is noticeable in Fig. 4.8, where we plotted the variance of the distance as a function of the rank. If the word is frequent, its variance is low but if the rank of a word is larger than 10^3 then the variance starts to be high. This means that the more the word is rare, the more the positions it assumes in the sentences move away from the mean value.

In conclusion, we defined three different distances and used them to analyze the network with the purpose of trying to find a relation between the structure of the graph and its topology. Taking into account more features such as the number of occurrences of words and links, the position of words in the sentences and so on, the definitions of distance become more accurate for bringing out the features of the network. However, for all the three distances the more the words get rare the more the distance grows. We could interpret this result in terms of structure of the network and use a different approach to analyze it.

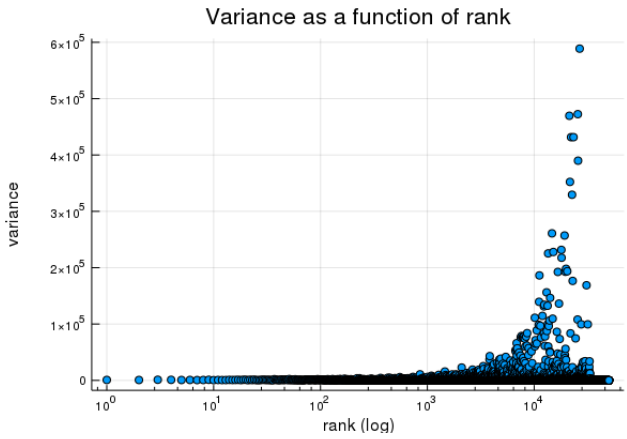


Figure 4.8: Variance of the mean distance as a function of the rank (log-lin scale). The variance grows with the rank, which means that the mean of the distance is not reliable for high ranks.

4.3.4 Centrality measures

One evidence of the previous analysis is that, if the distance used takes into account some simple features of the words, it is observed to grow the more the words become rare. Since we took into account the distance of the words from the central vertex ρ, that evidence means that the network has a shape such that the frequent words occupy its central part while the rare ones are in the outer part. One interpretation of this phenomenon could be that while the most frequent words would refer to general topics, the most rare would be specialized words that refer to particular and niche subjects. The frequent words would be characterized by many entering and exiting links that connect them to other frequent words and to rare words. On the contrary, since the rare words are specialized they would be linked to the frequent words but only to some other rare words, the ones that refer to the same particular topic. As a consequence, the network would be characterized by a big central component populated by the frequent words and some remote components inhabited by rare words. The central component would be characterized by an elevated number of links that connect frequent words with other words, frequent or rare. On the other hand, since rare words would be related only to frequent words or to words that refer

to the same topic, the remote part would be divided into different components on the basis of the subject. We want to find out whether the network has this structure and the best way of doing it is to use centrality measures.
Centrality measures are useful to study the configuration of the network because they associate every vertex of a graph with a number that represents its importance in the network. The definition of importance for a vertex could be related to many features, but the one that is necessary to our survey is the relation between the vertex and the links of the network.

Betweenness centrality

One of those measures is the betweenness centrality. Its definition uses the notion of shortest paths and it ranks the vertices on the basis of their position: the more shortest paths the vertex is in the middle of, the higher the centrality is.

Definition 4.9. Let $v \in \mathcal{V}$, the betweenness centrality associated with v is:

$$BC(v) = \sum_{\substack{s,t \in \mathcal{E} \\ s,t \neq v}} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $\sigma_{s,t}$ is the number of shortest paths between s and t and $\sigma_{s,t}(v)$ is the number of those shortest paths that pass through v.

In other words, the betweenness centrality of the vertex v takes into account the fraction of shortest paths that connect two vertices of the graph and pass through v.

We applied this measure to our network and found that the betweenness centrality decreases with the growth of the rank (see Fig. 4.9). Therefore frequent words have a central position in the network, while rare words are at the border. However, this decrease does not seem to be related with the change in slope. In fact, the rate of decrease of the betweenness centrality does not have any relevant change after the threshold of 10^3 words. The reason of this behaviour could lie in the fact that the definition of the betweenness centrality does not

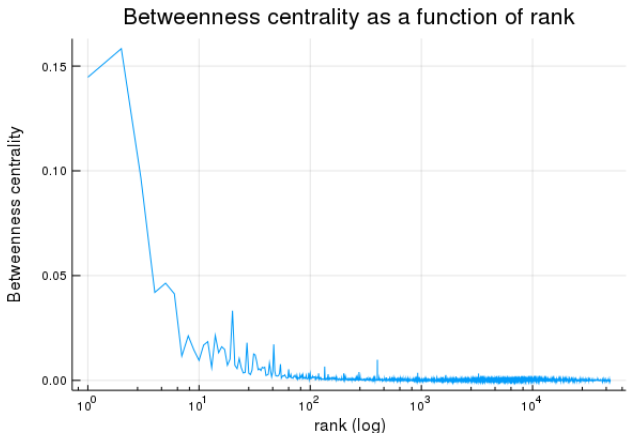


Figure 4.9: Betweenness centrality as a function of the rank (log-lin scale). It decreases with the growing of the rank, that is the most frequent words have a central position in the network.

take into account the weight of the links, therefore the occurrences of the words are irrelevant to this method.

Remoteness

One other way of defining a centrality measure on the network is the remoteness. The remoteness measures the difficulty of getting from one vertex to one other in the graph. We define the remoteness of a vertex v as the sum of the length of the shortest paths that connect v with every other vertex of the graph. Therefore the more a vertex is well connected to all the others, the lower the remoteness is.
Note that this measure requires the notion of length of the paths. We can use both the length definitions we gave above in order to find the one that is the most suitable for our analysis.
Using the first definition of length we have:

Definition 4.10 (Remoteness)**.** Let $v \in \mathcal{V}$. The remoteness of the vertex v is defined as

$$R(v) = \sum_{\substack{u \in \mathcal{V} \\ u \neq v}} \min_{v \to u} \ell(v \to u)$$

As a consequence of the definition of length used, the remoteness defined above does not take into account the occurrences of the words. As we expected, the remoteness and the rank are positively correlated (see Fig. 4.10): the more a word is rare the higher its remoteness is. This results makes sense in our inter-

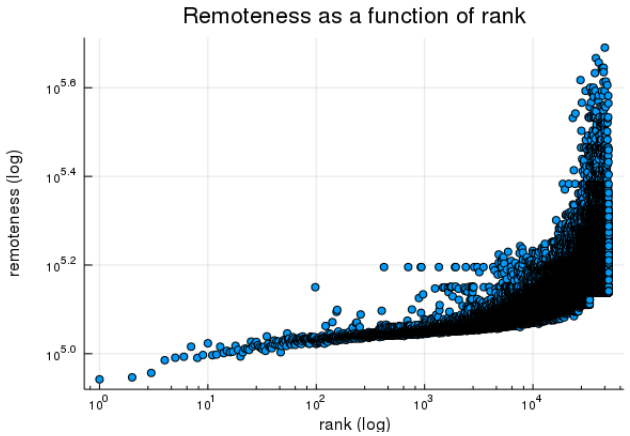


Figure 4.10: Remoteness as a function of the rank (log-log scale).The remoteness of frequent words is lower than the one of rare words, that means that the more a word is rare the more difficult it is to reach it.

pretation, because it confirms that rare words are the worst connected, therefore they are on the border of the network while frequent words are in the center. However, in Fig. 4.10 it is possible to see that the more the rank grows the wider the remoteness range is, i.e words with the same rank could have rather different remoteness values. Moreover, since the growth rate of the remoteness values does not change its behaviour passing the threshold of 2×10^3 words, it does not seem to be related with the changing slope of Zipf's law. Therefore we found that the remoteness could be related to the rank of the words, but now we need a more accurate definition of it in order to find a relation with the double slope.

For finding a definition of remoteness that is more suitable for our purposes we should consider also the number of occurrences of the words. As we did for the distance, we can adjust the definition of remoteness taking into account the weight of the links. Therefore, recalling the definition 4.4 we can define:

Definition 4.11 (Weighted remoteness)**.** Let $v \in \mathcal{V}$. The remoteness of the

vertex v is defined as

$$R(v) = \sum_{\substack{t \in \mathcal{E} \\ t \neq v}} \min_{v \to t} \ell_W(v \to t)$$

Using this definition for analyzing the network, we can plot the remoteness as a function of the rank of the words (Fig. 4.11). As we found for the previous definition of remoteness, also the weigthed remoteness grows with the rank of the words, therefore the rank and this version of the remoteness are related. Since Zipf's law is based on the frequency rank of the words, this evidence is vital for a correlation of the remoteness with Zipf's law.

The first difference with the result found with the previous definition is that

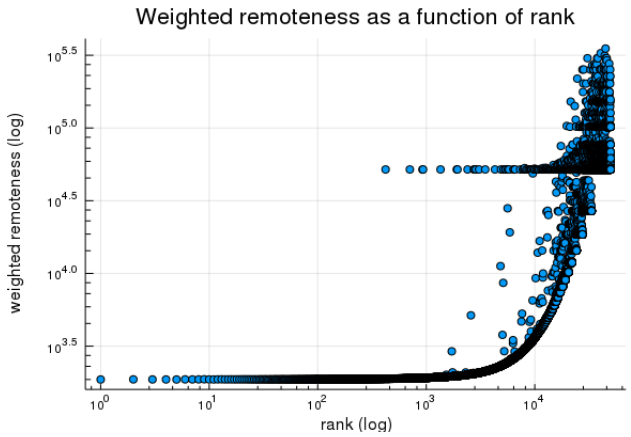


Figure 4.11: Weighted remoteness as a function of the rank (log-log scale). The weighted remoteness divide the words into two components of the graph, one more connected and one less connected. Since this phenomenon starts to be evident from a rank value of 2×10^3 it seems to be related with Zipf's law. The first words in the second component are all connected to the same node. For this reason they happen to have more or less the same value of weighted remoteness.

here the noise in the remoteness measure appears later. This evidence is a consequence of the decision of considering also the occurrences of the words. In fact, taking into account all the occurrences of the words we have a larger statistical basis and the measure is reliable also for words with high rank. Therefore this second definition is more accurate and reliable than the first.

Moreover, the weighted remoteness highlights an important feature of the net-

work. In fact, in Fig. 4.11 it is possible to see that from the rank value of 2×10^2 the words seem to be split in two components. In the first component there are words that still have a low remoteness value, while in the second there are words that have the same rank but higher value in remoteness. The phenomenon seems to become relevant once the rank passes the 2×10^3 threshold, therefore it seems to be related in some way to the change of Zipf's slope. In fact, only less than 10 words with rank smaller than 2×10^3 inhabit the second component, while all the other words have rank larger than 2×10^3. Eventually, also the words from the first component reach the high remoteness values of the second, but it happens for high values of rank, hence when the words are so rare that the remoteness measure is not reliable any more. Therefore this result shows that the central part of the network, the more connected one, is inhabited by frequent and less frequent words. On the other hand, some of those less frequent words start populating another component of the graph, less connected and central than the first. The important remark is that the threshold between frequent and less frequent words is set at the rank value of $\sim 2 \times 10^3$, that is where Zipf's law changes its slope. Note that, due to the structure of the network those two components can not be separated.
We found that taking into account the density of the links in the network is a suitable way for defining different components of the graph. Using this idea, another way of detecting the structure of the network could be based on the definition of some components with respect to the features of the links that characterize them.

4.3.5 Components of the network

Since the components of the network seem to be related to the density of its links, we define three sets of core words, in-words and out-words on the basis of the links of the network. Every set of words corresponds to a component of the graph. We took inspiration from the tie structure of the web [5] for dividing the graph into components but our definitions and analysis differ from the ones by Donato et al.

The core words constitute the central component, the one with the higher density of links. The density of the links is related to the length of the paths from one word to the other, in fact the more a component is dense with links the lower the distance between the words is. As we already advanced, in a directed graph the main obstacle for having a low distance is the direction of the links. In fact, if the link $u \mapsto v \in \mathcal{E}$ there is not guarantee that also $v \mapsto u \in \mathcal{E}$. As a consequence, considering the definition of length of a link in a graph without weights, even if $d(u, v) = 1$, the length of the path $v \to u$ is potentially longer than 1. On the contrary, not considering the direction of the links the two distances are the same: $d(u, v) = d(v, u) = 1$. The same applies to the graph with weights. Therefore in a directed graph a way of avoiding taking into account the directions of the links is to have couples of links $u \mapsto v$ and $v \mapsto u$. Hence we define the core words as those words that have at least one two-ways link:

Definition 4.12. Let $u \in \mathcal{V}$, it is a core word if $\exists v \in \mathcal{V}$ s.t. $u \mapsto v$ and $v \mapsto u \in \mathcal{E}$. We define CORE$= \{v, v$ core word$\}$.

Now we define the in-words as words that connect the outer part of the network with the core words and the out-words as the words that connect the core words with the outer part.

Definition 4.13. Let $u \in \mathcal{V}$, it is a in-word if $\exists v \in \mathcal{V}$, v core-word s.t. $(u, v) \in \mathcal{E}$ and $\nexists v \in \mathcal{V}$, v core-word s.t. $(v, u) \in \mathcal{E}$. We call IN$=\{v, v$ in-word$\}$

In other words, an in-word has at least one core word as outneighbour but no core words as inneighbours.

Definition 4.14. Let $u \in \mathcal{V}$, it is an out-word if $\exists v \in \mathcal{V}$, v core-word s.t. $(v, u) \in \mathcal{E}$ and $\nexists v \in \mathcal{V}$, v core-word s.t. $(u, v) \in \mathcal{E}$. We define OUT$= \{v, v$ out-word$\}$.

Therefore an out-word is a word that has at least one core word as inneighbour but no core words as outneighbours.
As a consequence of the previous definitions, the in-words and out-words have the role of connecting the central part of the graph with the outer part.

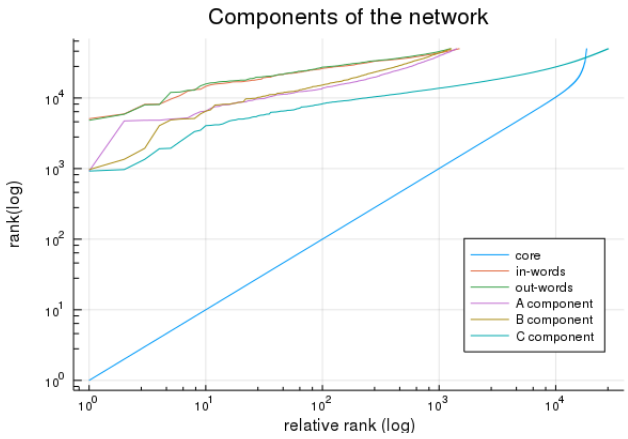


Figure 4.12: Rank of the words divided on the basis of the components of the network (log-log scale). The IN and OUT components are populated by words with rank higher than 3×10^3, while the A, B and C components start to be populated from a rank value of 10^3.

To study the three units of the network we plotted the rank of the words divided on the basis of the component they belong to (Fig. 4.12).

In that figure it is possible to see that the core component is larger than the other two and it is populated by words with various range in frequency, from frequent words to rare ones. On the contrary, the other two components have the same shape rank-wise and start to be populated from rank 3×10^4 on. Note that the three components gather less than 50% of the words of the sample, therefore they are only a minority.
We can interpret the in-words as the words that serve as connection between the the outer part of the graph and the core words and the out-words as the one that connect the core with the outer part. Therefore we could be interested in knowing how the rank of the words in the outer part is distributed. We define three new components

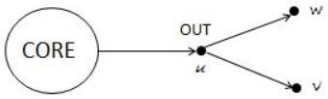
Definition 4.15. Let $v \in \mathcal{V}$, $v \notin \text{IN} \cup \text{CORE}$, $v \in \mathcal{A}$ if $\exists u$ in-word s.t. $v \mapsto u \in \mathcal{E}$.

Therefore $\mathcal{A}$ is the component of all the words that use an in-word as a link to the core.

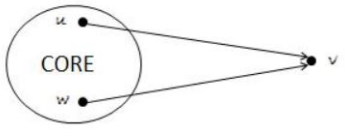
Definition 4.16. Let $v \in \mathcal{V}$, $v \notin \text{OUT} \cup \text{CORE}$, $v \in \mathcal{B}$ if $\exists u$ out-word s.t.

$u \mapsto v \in \mathcal{E}$.

As a consequence, $\mathcal{B}$ is the set of all the words to which the core words are linked with links that pass through the out-words (see Fig. 4.13(a) for a graphical representation).



(a) Graphical representation of the definition of $v, w \in \mathcal{B}$. Inverting the arrows, the OUT component becomes the IN component and the definition of $\mathcal{A}$ is recalled.



(b) Representation of the definition of $v \in \mathcal{C}$. Since $u \neq w$, the shortest possible loop in $\mathcal{C}$ consists at least of three links.

Figure 4.13: Graphical representation of the definition of the three components $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. The shortest possible loop in those components consists at least of three links.

Definition 4.17. Let $v \in \mathcal{V}$, $v \notin$CORE, $v \in \mathcal{C}$ if $\exists u, w$ core words s.t. $u \mapsto v, v \mapsto w \in \mathcal{E}$.

Note that $v \neq w$ because otherwise v would be a core word.
From the definition we have that $\mathcal{C}$ is the set of the words that do not use in-words or out-words as links to the core.
From the definitions, loops of length 2 are allowed only in the CORE. Therefore we distinguished a dense part of the graph where 2-links loops are allowed (the CORE) and one outside part where the shortest loops consist of at least 3 links. The outside part is divided into other components with respect to the relation of their vertices with the CORE (see Fig. 4.15). Note that this classification is linked to the density of links: the more the links are dense, the more 2-links loops are probable.
Note that not all the intersections between the components are empty. More in particular, we observe that $A \cap OUT \neq \emptyset$, $B \cap IN \neq \emptyset$, $A \cap B \neq \emptyset$, $C \cap A \neq \emptyset$ and $C \cap B \neq \emptyset$ (see Fig.4.14).

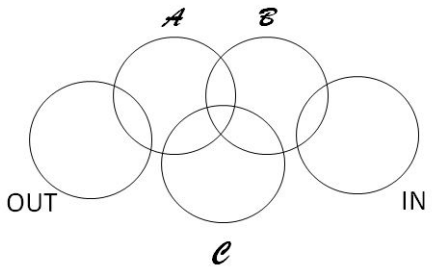


Figure 4.14: Scheme of the intersections between the components.

Fig. 4.14 is a graphical representation of the structure of the graph considering the components described above. The components $\mathcal{A}$ and $\mathcal{B}$ use respectively the in-words as a link entering in the core and the out-words as a link exiting from the core. The third component, $\mathcal{C}$, is constituted by words that do not belong to the core and do not use neither the in-words nor the out-words to be connected to the center of the network.

In Fig. 4.12 it is possible to see that the components $\mathcal{A}$ and $\mathcal{B}$ have the same

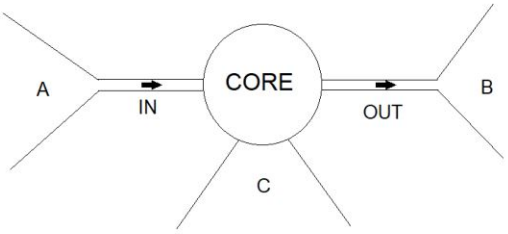


Figure 4.15: The division of the network in the CORE, IN, OUT, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ components. The in and out components serve as links from the outer part to the core.

composition in terms of rank and number of words, while the component $\mathcal{C}$ has a larger cardinality than the previous. Note that except for the first two or three words, the $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ components start to be inhabited by words with rank values from 2×10^3 on. This means that the composition of those components could be related to the changing slope of Zipf's law. In fact, since the change in

the slope is evident but smooth, it outlines a range of ranks that are intermediate in the change. Those intermediate ranks could be the first few words of the $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ components. The IN and OUT components are inhabited by words with rank larger than 3×10^3, therefore they seem to be related to the second slope of Zipf's law. A different definition of the components based on a more strict and precise concept of density of the links of a graph could help outlining better the components and could lead to more precise results in terms of ranks of the words of the components.

To understand if the previous definition is significant with respect to Zipf's law,

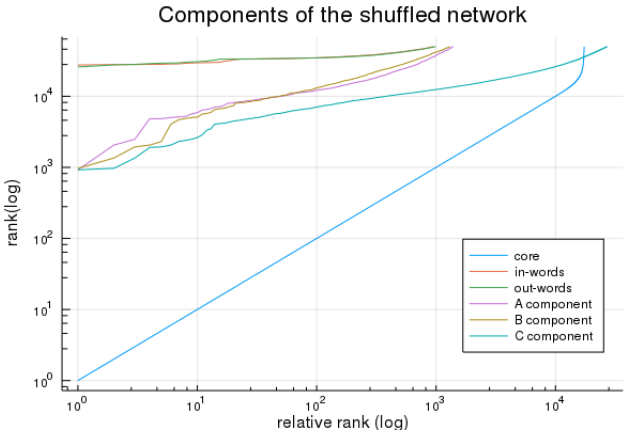


Figure 4.16: Rank of the words divided on the basis of the components of the network of the shuffled sample (log-log scale). The core, A, B and C components don't change significantly, while the IN and OUT components are populated by words with higher rank.

we consider the network built on the shuffled sample. As we advanced, the shuffling does not affect Zipf's law but it changes the structure of the network. As a consequence, the graph on the shuffled sample would be constituted by the same inner and outer parts, but the links between them would be different. More in detail, since the rare words occupy random positions in the tests, the outer part would not show different aggregation of words on the basis of the topic they deal with. In Fig. 4.16 it is possible to see that the IN and OUT component on the shuffled text are different from the one of the normal sample while all the other components do not vary. The IN and OUT words are populated by words that

are more rare, while all the other keep the same composition. This means that the IN and OUT words are significant to detect the structure of the network with respect to the order of the words and their meaning, while the other components are only related to the frequency of the words. That is, we found a structure that does not change with the shuffling of the texts, i.e. it is related to Zipf's law, and a structure that depends also on the order of the words and the topic they deal with, i.e. it is a deeper structure.

In conclusion, the definition of those components seem to be related to the frequency rank of the words of the sample. The core components contains words with a wide range of ranks, while the other components are more specifically populated. This means that basing only on the idea of density of the links, we found a central component of the network, the core component, and some outer ones, the in, out, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. This evidence seems to confirm the assumption that the frequency of words is reflected in the features of the network we built and more specifically in the structure of its links.

Conclusions

Zipf's law manifests itself in many different areas. For example, we find it in the distribution of earthquake magnitudes, in the distribution of the number of inhabitants of cities and in the energy distribution of cosmic rays, to cite a few. In its original form, Zipf's law consists of a pure power-law relation between the frequency of occurrence of certain events and their rank, whereas the rank is an integer number ranking the frequencies from the highest (rank 1) to the lowest. Generally and historically, that power-law relation has an exponent of -1.
Aim of this thesis is to deepen our understanding of power-laws in general and Zipf's law in particular. In order to do this, we first studied some synthetic models able to generate power-laws. Second, we investigated the relation between the frequency of words and their rank in written English texts. In this case this relation cannot be approximated by a conventional Zipf's law with constant exponent, but at least two regimes can be observed. One regime embraces the first few thousands of most frequent words with an exponent close to the canonical -1, while the other regime found in the case of the least frequent words is slightly less than -2.

In our theoretical approach we reformulated in a mathematical way some models that reproduce power-laws in terms of stochastic processes. Because of the relevance of Zipf's law, several scholars have studied it and have proposed some interpretations for its insurgence. In those interpretations, the observation of power-laws in the FR distribution has been related to many causes, e.g., to the structure of language itself, to the dynamics of language formation and the principle of least effort, to the human brain and its limits. In some cases model-

ing models have been proposed.
We considered some of those models, framed them within the theory of stochastic processes and studied their properties. The mathematical formalization we performed allowed us to apply theoretical tools to the models in order to detect their features. In particular, we considered two different classes of models based on the variation of their sample space, where the range of their possibilities grows (generalized urns models) or decreases (SSR models) with time. We leveraged the theory of stochastic processes. and Markov chains, which allowed us to apply the Perron-Frobenius theorem to the models. Finally,as the main result of our theoretical mathematical approach, we introduced a brand new population dynamics model that gives a unified picture of the SSR processes, being able to reproduce their behaviour on the full range of exponents. Such new model is of outstanding importance because it unifies three different processes, defined and treated differently in literature, in a unique elegant framework.
We followed this first theoretical analysis with a quantitative study, where we analyzed a corpus of texts and related some of their features to the varying local exponent of the word FR. We constructed a network of word adjacency where word A is connected with a direct link to word B if word A precedes word B. In order to have a minimal explanation of the existence of the double slope in the FR distribution of words, we conjectured that the network would have a topological structure resembling a 'daisy', i.e. a structure with a central part and some external petals. In fact, a daisy structure of the graph would imply that the change in slope of the FR is related to some features of the human knowledge. We conjecture that the knowledge of everyone is subdivided into a component of base notions shared by the whole population and a sectoral component depending on the individual that deals with more specialized topics. Since the words that refer to the base knowledge are generally in common with any context they are evenly used in any dissertation. As a consequence, we expect to find them in the inner component of the daisy. On the other hand, since the more specific words have the opposite behaviour we expect to find them in the outer part, aggregated in petals on the basis of the topic they deal with. Being more general,

we think that the words that populate the inner part of the daisy structure are used more frequently in the language than the specific words. This hypothesis leads us to conjecture that the inner part of the structure of the network will be populated by the words of the first slope of the FR and the specific words of the petals would be the ones that characterize the second slope.
Note that a reshuffling of the texts is irrelevant for the FR while the structure of the network, being based on the order of words, will be affected by it. The topological structure of the network is influenced by both the order and the frequency of words, so that in such a network one has two layers: a first weak one related to the order of words and a second strong one related only to their frequency. A shuffle of the texts will affect the first layer but not the second.
Since the considered graph is of high order and size so that its thorough structural analysis is computationally infeasible, we devised a simple and novel method to indirectly understand whether a daisy-like structure is actually present. This method relies on the study of reciprocal links and is based on the assumption that a good indicator of the density of links is the length of loops. In other words, the more a region of the network has dense links, the shorter the loops are.
Initially, we tried to check whether the assumption of a correlation between the topological structure of the graph and the FR of the word is founded. We defined some distances and investigated the structure of the network with respect to those measures. At first, we just used the plain topological distance, then we defined a new one that additionally considered the number of occurrences of the links. Finally, we defined a third measure that also took into account the positioning of the words in the sentences. As a result we found out that the more detailed the distance was the more it seemed to be correlated with the FR of the words. Hence this first evidence seems to confirm the assumption that the topology of the network is related to the FR of the words of the texts, despite its invariance to reshuffling.
Subsequently, we applied some centrality measures to the network. This kind of measures gives an indication of the importance of the nodes in the graph. Since we were interested in the topology of the network, we performed only the cen-

trality measures that were in some ways related to the density of the links. More in detail, for our study we used three different measures each based on a different distance. Accordingly with the topological study, we found that the more refined the centrality measure was the more it seemed to be correlated to the change in slope of the FR. Worth noting, the centrality measure that took into account the number of occurrences of the links seemed to outline two different components of the network. In other words, also these results seem to be consistent with the conjecture that the structure of the network is related to the frequency of the words. More in particular, it seems that the density of the links naturally outlines a structure of the network divided in components.

Inspired by this result, we decided to define some components of the network on the basis of the density of their links. As already mentioned, this division is based on a completely new way of quantifying the density of links and leads to some interesting results. In fact, studying the composition of the components, we found out that they are differently populated frequency-wise. More in detail, we observed that some components are populated only by words that belong to the second slope of the FR while others are inhabited by words more evenly distributed. This evidence outlines that the composition of some of the components is related to the FR of the words, hence the structure of the graph could be associated with the FR. This means that the daisy structure hypothesis could be founded.

To check the hypothesis of a strong structure of the network based only on the frequency of words and not on their order we performed the same analysis on the shuffled network, where the word adjacencies are determined after shuffling all the words in the texts. In particular, we found that the IN and OUT components, defined as the components that have the function of connecting the dense part of the network with the outer part, changed their statistical composition while the other did not. This evidence outlines that the composition of some of the components is related to the FR of the words, while the IN and OUT are also related to the order of the words. Therefore the IN and OUT components are probably related to the weak structure of the network, while the others may

be related to the strong one.
Concluding, our thesis dealt with the FR in the context of linguistics. We addressed this problem because its study could reveal some hidden properties of the human mind and deepen our understanding of the structure of language. We performed the analysis of this problem with both a theoretical and an experimental approach. With regard to the theoretical study, we formalized some models that reproduce power-laws and showed that stochastic processes and Markov chains are powerful tools for the analysis of those models. On the quantitative experimental side, we built a network of word adjacencies and found some evidence that led us to think that its topological features are related to the FR of the words. In addition to this, we outlined with a novel method some components of the network that seem to confirm its hypothesized daisy structure. This is an important result because it means that the double slope of the FR is related to the structure and the limits of the human mind. It is a completely new approach to the study of the change in slope of the FR and it could be deepened in future. For example it would be interesting to analyze in the same way a larger corpus of texts and to build a model that represents the network and its properties. In the context of quantitative linguistics this study could lead to a development in the field of text mining and analysis: from the daisy structure of the network it is possible to infer the topics that are addressed in the text, paving the way for more modern techniques of text mining.

Bibliography

[1] Akimushkin C., Amancio D.R. and Oliveira O.N.Jr, On the role of words in the network structure of texts: application to authorship attribution, *Physica A: Statistical Mechanics and its Applications*, **495**,(2018), pp 49-58.

[2] Amancio D.R., Altmann E.G., Oliveira O.N.Jr and Fountura Costa L., Comparing intermittency and network measurements of words and their dependence on authorship, *New journal of physics* **13** (2011) 123024.

[3] Corominas-Murtra B., Hanel R. and Thurner S. (2017), Sample space reducing cascading processes produce the full spectrum of scaling exponents, *Sci Rep.*;7(1):11223.

[4] Corominas-Murtra B., Hanel R. and Thurner S. (2015), Understanding scaling through history-dependent processes with collapsing sample space, *PNAS*, 5348-5353.

[5] Donato D., Leonardi S., Millozzi S. and Tsaparas P. (2008), Mining the inner structure of the Web graph,*Journal of Physics A: Mathematical and Theoretical*, **41**, 22.

[6] Euler L. (1736), Solutio problematis ad geometriam situs pertinentis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae.*

[7] Ferrer i Cancho R. and Solé R.V. (2003), Least effort and the origins of scaling in human language, *PNAS* **100**, 3, pp.788-791.

[8] Ferrer i Cancho R. and Solé R.V. (2001), The small world of human language, *Proc. R. Soc. Land.*, **268**, 1482, pp. 2261-2265.

[9] Font-Clos F., Boleda G. and Corral A. (2013), A scaling law beyond Zipf's law and its relation to Heaps' law, *New Journal of Physics*, **15**, 093033.

[10] Gerlach M. and Altmann E.G. (2013), *Stochastic model for the vocabulary growth in natural languages*, Physical Review X, 021006.

[11] Heaps H.S. (1978), *Information Retrieval: Computational and Theoretical Aspects*, Orlando, Academic Press.

[12] Mandelbrot B. (1935), An informational theory of the statistical structure of language, *Communication theory*, London, W. Jackson, pp. 486-502.

[13] Manning C.D. and Schütze H., Foundations of Statistical Natural Language Processing, MIT Press (1999).

[14] Mariotti F. (1880), *Dante e la statistica delle lingue*, Firenze, G. Barbera Editore.

[15] Markov A.A. (1913), An example of statistical investigation of the text *Eugene Onegin* concerning the connection of samples in chains, *Science in Context*, **19**(4),pp.591-600.

[16] Meyer C.D. (2001), *Matrix analysis and applied linear algebra*, SIAM.

[17] Newman M.E.J. (2005), Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, **46**, 5, pp.323-351.

[18] Simon H.A., On a class of skew contribution functions, *Biometrika* **42**, 3-4, pp. 425-440.

[19] Sowa J.F. (1991), *Principles of semantic networks: explorations in the representation of knowledge*, Morgan Kaufmann, California.

[20] Tria F., Loreto V., Servedio V.D.P. and Strogatz S.H. (2014), The dynamics of correlated novelties, *Scientific Reports*, 4:5890.

[21] Tria F., Loreto V. and Servedio V.D.P. (2018), Zipf's, Heaps' and Taylor's laws are determined by the expansion into the adjacent possible, *Entropy*, **2018**, 20, 752.

[22] Yule G.U. (1925), A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis. *Philos. Trans. R. Soc. London B*, **213**, 21-87.

[23] Zanette D.H. (2012), *Statistical patterns in written language.*

[24] Zanette D.H. and Montemurro M.A. (2005), Dynamics of text generation with realistic Zipf's distribution, *J. Quant. Linguistics*, **12**, 1, pp. 29-40.

[25] Zipf G. K. (1949), *Human behaviour and the principle of least effort. An introduction to human ecology*, Cambridge, Addison-Wesley.

[26] Zipf G. K. (1936), *The psycho-biology of language. An introduction to dynamic philology*, London, Routledge.